Yeminy *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:16
DOI 10.1186/s13636-016-0094-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Single microphone speech separation by diffusion-based HMM estimation

Yochay R. Yeminy*, Yosi Keller and Sharon Gannot

Abstract

We present a novel non-iterative and rigorously motivated approach for estimating hidden Markov models (HMMs) and factorial hidden Markov models (FHMMs) of high-dimensional signals. Our approach utilizes the asymptotic properties of a spectral, graph-based approach for dimensionality reduction and manifold learning, namely the diffusion framework. We exemplify our approach by applying it to the problem of single microphone speech separation, where the log-spectra of two unmixed speakers are modeled as HMMs, while their mixture is modeled as an FHMM. We derive two diffusion-based FHMM estimation schemes. One of which is experimentally shown to provide separation results that compare with contemporary speech separation approaches based on HMM. The second scheme allows a reduced computational burden.

Keywords: Single microphone speech separation, Manifold learning, Diffusion maps, Factorial hidden Markov models

1 Introduction

Single-channel speech separation (SCSS) is one of the most challenging tasks in speech processing, where the aim is to unmix two or more concurrently speaking subjects, whose audio mixture is acquired by a single microphone. The goal is therefore to decompose the single input signal into multiple output channels, each dominated by a single speaker. The core obstacle in such tasks is the lack of spatial information, and the common statistical characteristics of the mixed signals.

Single-channel speech separation was studied by several schools of thought, where computational auditory scene analysis (CASA) proved to be among the most effective. CASA-based methods are motivated by the ability of the human auditory system to separate acoustic events, even when using a single ear (although binaural hearing is advantageous). CASA techniques imitate the human auditory filtering known as cochlear filtering, where time-frequency bins of the speech mixture are clustered using psychoacoustic cues such as the pitch period, temporal continuity, onsets and offsets, etc. The clustering associates each time-frequency bin with a particular source. The time-frequency bins associated with the desired source are retained, while those associated with the interference are attenuated. Such approaches were studied by Weintraub [1], Parsons [2], and Brown and Cooke [3]. Contemporary CASA schemes utilize oscillatory correlations [4], and common amplitude modulation [5], but do not utilize prior information regarding the source signals and their number. All they require is that each time-frequency bin is dominated by a single speaker. A probabilistic interpretation of CASA was proposed by Wang and Brown [6], and applied by Vincent and Plumbley [7], who proposed a Bayesian formulation of the separation based on a harmonic model of the sources.

The association of each time-frequency bin with a particular speaker is usually referred to as *binary* or *hard* masking. In the ideal case, where the time-frequency bin association of each source is perfectly known, the mask is usually referred to as the ideal binary mask (IdBM), and it was shown by Li and Wang [8] to be optimal in terms of source to noise ratio (SNR). Yilmaz and Rickard [9] showed that (ideal) binary masking enables the separation of up to ten sources from a single mixture.

Alternatively, a *soft* mask can be used, where each time-frequency bin is assumed to be associated with multiple signals (with different weights), and their relative spectral content in each time-frequency bin has to be estimated.

Blind source separation (BSS)-based approaches are commonly implemented via independent component

*Correspondence: yochay.ye@gmail.com
Faculty of Engineering, Bar-Ilan University, Building 1103, Ramat-Gan, Israel

analysis (ICA), and are widely used in multi-microphone speech separation. In the SCSS context they are formulated as an undetermined BSS problem [10, 11]. Zibulevsky and Pearlmutter [12] used the Fourier coefficients to represent speech signals and utilize their sparseness for separation.

Van der Kouwe et al. [13] conducted an experimental study that compared CASA [4] to multi-microphone BSS approaches (joint approximate diagonalization of eigenmatrices (JADE) [14] and second order blind identification (SOBI) [15] algorithms), and showed the advantage of the latter approaches that utilize spatial information. However, in many important application, such a spatial information is unavailable.

Recent separation schemes applied non-negative matrix factorization (NMF), where the magnitude of the Fourier transformed frames is factorized as the product of two non-negative matrices. The first comprising the basis functions, and the second encoding the weights of the corresponding basis functions. The matrix of basis functions is speaker-adaptive, learned in a training phase. In the separation stage, the power spectral density (PSD) of the mixture is modeled by a linear combination of the basis functions of both speakers. The corresponding weight matrices are estimated, and utilized to estimate the underlying sources. Virtanen [16] proposed an NMF approach that encourages PSD continuity and sparseness. Smaragdis [17] proposed a convolutive form of the NMF to model the time dependencies of the PSD. A semi-supervised real-time NMF algorithm was proposed by Joder et al. [18]. Only one source is learned from training data whereas the other source is estimated based on the recent past of real-time data.

Benaroya et al. [19] express each source as a weighted sum of temporal Gaussian stationary processes, with positive, slowly time-varying weights. The PSD is approximated by the weighted sum of the variances of the Gaussian processes, yielding a non-negative representation, and the sources are recovered utilizing the Wiener filter. Blouet and Cohen [20] extended this factorization [19] to separate speech from speech-music mixtures, where the weighted sum of processes approximates the short-time Fourier transform (STFT) of each source as complex-Gaussian stationary processes. A sinusoidal modeling of the time-domain was proposed by Mowlaee et al. [21], where a codebook is trained for each source and utilized in the separation stage. This work was extended in [22], and includes a preceding stage of detecting double-talk or single-talk frames, as well as a speaker identification system.

Machine learning approaches were applied to speech separation by Bach and Jordan [23]. They proposed to treat the separation problem in the time-frequency domain as a segmentation problem and to apply the relevant segmentation tools to the audio features extracted from the speech spectrogram.

Deep learning techniques are gaining popularity following their success in single-talker automatic speech recognition tasks. Essentially, the networks are trained based on parallel sets of mixtures and their constituent target sources. They are optimized to predict the source of the target class, usually for each time-frequency bin. For example, in [24], the speakers are estimated by jointly optimizing a soft time-frequency mask layer with deep recurrent neural networks. However, these works often assume speaker-dependent models with few target speakers that are known during training. In addition, they usually work on limited vocabulary and grammar. Yu et al. [25] proposed a speaker-independent method for multi-talker speech separation by using permutation invariant training. It first determines the best output-target assignment and then optimizes the separation regression error given the assignment. Another speaker-independent technique is proposed in [26], where contrastive embedding vectors are assigned to each time-frequency region of the spectrogram. It results in implicit prediction of the segmentation labels of the target spectrogram from the input mixtures. Separation is obtained by optimizing K-means with respect to the unknown assignment. In [27], the authors propose to use an ensemble of deep neural networks and demonstrate the superiority of this structure over speech separation algorithms based on a single network.

A plethora of approaches utilize statistical models of speech signals. In [28], the PSD of each speech frame is computed by iterating between randomly drawing frequency bins from a mixture of multinomial distributions, and scaling the histogram of the drawings. Given the probabilistic models, the minimum mean square error (MMSE) estimate of the desired source is derived. Essentially, this method is virtually indistinguishable from methods applying NMF.

Gaussian mixture models (GMMs) and HMMs are extensively utilized in speech separation tasks. Kristjansson et al. [29] modeled the log-spectrum of each speaker by an GMM. They approximate the joint probability density function (p.d.f.) of the log-spectra of the speakers given the log-spectrum of the mixture, by a normal distribution. Using this approximation, the posterior distributions of the log-spectrum of each speaker are computed, and the MMSE estimator is derived. GMM-based representations can be modified to improve the temporal modeling. Such an approach was proposed by Benaroya et al. [30], where the variances in the GMM are scaled by time-varying factors, to incorporate source dynamics.

It is common to apply the GMM and HMM framework using the *log-max* approximation, that was first proposed by Nádas et al. [31], in the context of speech recognition, and was denoted as MIXMAX. Burshtein and Gannot [32] reformulated MIXMAX, by assuming that the log-spectrum of the clean speech segments can be modeled by GMMs, while the noise log-spectrum is normally distributed. In particular, the log-spectrum of the noisy speech is approximated by the maximum of the log-spectrum of the speech and noise. This result was extended by Yeminy et al. [33], who proposed a generalized formulation, incorporating the correlation between adjacent frequency bins. Reddy and Raj [34] applied the MIXMAX model to the speech separation task, where both input signals were modeled by GMMs, and derived two estimators. The first optimizes the MMSE, and the second uses a soft mask, computed using the posterior probability of the observed log-spectrum to match the log-spectrum of the desired speech. Radfar and Dansereau [35] used a similar model as in [34] with an additive error model, assuming that the error is zero-mean and normally distributed. The log-spectra of the speakers are recovered by optimizing the MMSE.

Roweis [36] modeled the log-spectra of the two speakers as the output of HMM processes. Using the log-max approximation, the log-spectrum of the mixture signal is approximated by the maximum between the corresponding outputs of the two underlying HMMs, and is modeled by an FHMM. The compound state of the log-spectral vector of the mixed signal consists of two states, corresponding to the speakers. A variant of the Viterbi algorithm, denoted as *factorial Viterbi* algorithm, is used to reconstruct the log-spectra of the speakers, while a binary mask is applied to recover the source. This approach was extended by Radfar et al. [37], to the case where the mixture of the two sources is a weighted sum of the noise-free speech signals. The gain factors are recovered by an iterative formulation of the FHMM, resulting in improved experimental results. However, in [37], the complexity of the algorithm is quadratic in the number of states, and the speaker-to-speaker ratio range must be an input of the algorithm. Hu and Wang [38] proposed an iterative separation scheme that estimates the gain factor without any prior knowledge. An HMM is trained for each source, where the utterances of the sources are scaled to have a known equal energy. The mixing model is the log-max approximation and the compound states are inferred using factorial Viterbi algorithm. Each loop of the algorithm consists of several steps. A separation phase, where the sources are estimated. Then, the speaker-to-speaker ratio is assessed using the estimated sources. Eventually, the pre-trained models of the speakers and the mixture are adapted to the estimated gain factors. It was reported in [38] that best results were obtained when the separation was based on the MAP estimation.

Hershey et al. [39] also proposed to utilize both FHMM and log-max approximation. The FHMM encodes the grammar dynamics of the sources, when the structure and vocabulary of each unmixed speech are known in advance. At each grammar state, the dynamics of the acoustics of each source is encoded by an GMM which is based on the log-max approximation. The grammar dynamics takes into account temporal long-term components of the speech, with respect to the acoustic dynamics, yielding improved experimental results, that in some scenarios even outperform human listeners. Weiss and Ellis [40] proposed a similar model, where the speech characteristics are unknown a priori, but adapted iteratively. Ming et al. [41] proposed a data-driven technique that is also based on long-term temporal dynamics, intended for the general scenario in which the vocabulary and grammar are unknown. A combined FHMM and NMF approach was presented by Mysore et al. [42], and denoted non-negative hidden Markov model (N-HMM). For each source, several small spectral dictionaries are learned. Their evolvement in time is also learned via HMM. The composite signal of the two sources is separated by applying a soft mask, generated by an estimation-maximization (EM) procedure that estimates the contribution of each source at each time-frequency bin. Good separation performance is reported for this N-HMM technique. A related work is [43], where a new model called factorial scaled hidden Markov model (FS-HMM) combines Gaussian scaled mixture model and NMF. The FS-HMM is utilized in [43] for speech separation and polyphonic audio representation.

The high dimensionality of the log-spectral vectors and the large number of states of the factorial model, imposes high computational burden for the FHMM inference. Roweis proposed to mitigate the high computational burden by detecting pairs of states with the highest observation likelihood, and limiting the factorial Viterbi calculations to their corresponding paths. In [38], the beam search [44] is used to speed up the inference process. A band quantization approach that reduces the number of HMM states, was proposed by Rennie et al. [45] to reduce the computational complexity.

An efficient belief propagation technique for the inference, rather than the exact factorial Viterbi algorithm was proposed by Hershey et al. [39] for the temporal inference. They also presented two methods that together efficiently compute the acoustic likelihood estimation of the observed mixed signal, which is required for the temporal inference. The first is called band-quantization, and it suggests to approximate some of the acoustic GMM states that differ only in a few features. Each Gaussian is approximated using a shared set of a smaller number of Gaussians in each frequency bin. The second technique is joint-state pruning, which utilizes the fact that several states pairs have significantly larger probabilities than the rest of joint-states. This feature stems from the sparseness of the model. Those states can be used to explain the observations, rather than using all of the possible pairs of states. Rennie et al. [46] applied a similar model able to separate up to four speakers using loopy belief propagation and variational inference that reduce the inference complexity. The max-sum algorithm, which is also a belief propagation technique, is also employed by [47] to track multiple pitch trajectories described by an FHMM. Reyes-Gomez et al. [48] proposed to group several frequency bins into frequency bands, such that each frequency band is modeled by an HMM.

Dimensionality reduction schemes were also applied to speech separation. Michalevsky et al. [49] applied the diffusion framework [50] to speaker identification, where a feature vector consisting of the mel frequency cepstral coefficients (MFCC) and their first temporal derivative is used to parameterize the manifold of each speaker, by embedding them into a lower dimensional space. Samples are classified by a k-nearest neighbors (k-NN) classifier applied to the embedding of the corresponding feature vector.

In this work we propose the following contributions:

> **First**, we derive a novel non-iterative speech separation approach based on the diffusion framework [50], to compute HMM and FHMM models. A comprehensive set of experimental results exemplify the applicability of the proposed method. It is shown that the proposed scheme provides separation results that compare with contemporary speech separation approaches.
>
> **Second**, by analyzing the asymptotics of Markov random walks, we show that the proposed scheme allows to directly estimate the states of the HMMs and FHMMs, without having to assume any underlying observation model, nor to apply EM-based iterative training. Hence, the estimation of the Markov states and transitions is decoupled from the estimation of the emission p.d.f.s, and their corresponding parametric model. Thus, we propose two FHMM-based approaches that estimate the underlying HMM in the diffusion domain. The first, provides a direct extension of the FHMM, where the underlying HMMs are computed in the diffusion domain, and the Gaussian observation models utilize the log-max approach applied in the original domain. We denote this approach hybrid FHMM (HFHMM), as it utilizes both the temporal and diffusion domains. In the second approach, denoted dual FHMM (DFHMM), we estimate the emission p.d.f.s in the diffusion domain, without assuming an explicit emission p.d.f. model. The underlying HMMs are computed similarly to the HFHMM approach.
>
> **Last**, we propose to utilize the diffusion embedding as a nonlinear projection of the input mixture onto the manifolds spanning each of the speakers. Thus, we aim to utilize the diffusion embedding as a manifold-adaptive projection operator, where the states of each speaker are detected by an FHMM in its manifold. The HFHMM is experimentally shown to compare with previous results [36], and is shown to outperform DFHMM. The latter requires low computational cost, and can be applied alongside other approaches [39] in the low-dimensional space to further reduce the computational complexity.

The remainder of this paper is organized as follows. Section 2 formulates the monaural speech separation problem of two equi-power and pre-trained sources. The diffusion framework, the Nyström extension [51] and the HMM inference using the diffusion maps are surveyed in Section 3. The proposed diffusion-domain speech separation schemes are presented in Section 4, where we detail the separation and training procedures, and propose two alternative mask functions. The proposed approaches are experimentally verified in Section 5, and their performance is compared with contemporary schemes. The computational complexity of the HFHMM and the DFHMM schemes is discussed in Section 6, while conclusions and future directions are discussed in Section 7.

2 Problem formulation

The speaker separation problem is often formulated as an FHMM problem. Let $a[n]$ and $b[n]$ be the speech signals of the first and second speakers, respectively, where n is the discrete time index. We assume that the speech signals are of equal power

$$\sum_n a^2[n] = \sum_n b^2[n] = 1, \tag{1}$$

and that $a[n], b[n]$ are zero mean and statistically independent. For the unequal power model, the reader is referred to [37, 38]. The observed signal is a mixture of the two speakers

$$z[n] = a[n] + b[n], \tag{2}$$

and the objective of the separation scheme is to compute the estimates, $\hat{a}[n]$, $\hat{b}[n]$ of $a[n]$ and $b[n]$, respectively, given $z[n]$. Let $A(\ell, k)$ be the STFT of $a[n]$, where ℓ is the temporal frame index and $0 \le k \le K-1$ is the frequency index. Denote $\mathbf{a}_\ell$ as the $K/2+1$ dimensional vector, whose kth element is

$$\mathbf{a}_\ell^k = \log|A(\ell, k)|; \qquad k = 0, 1, \ldots, K/2.$$

Hence, $\mathbf{a}_\ell$ is the log-spectrum of $a[n]$, and $\mathbf{b}_\ell$ is the respective log-spectrum of $b[n]$.

The core assumption of our approach is that $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are the observed outputs of two separate HMMs, one per speaker, denoted HMM_a and HMM_b, respectively. Let S^a be the number of states in HMM_a, and s_ℓ^a be the state of HMM_a at time frame ℓ. The probabilistic attributes of HMM_a are given by the initial probabilities $\Pr\left(s_1^a = s^{i,a}\right) = \pi_i^a$, $i = 1, \ldots, S^a$, where $s^{i,a}$ is the ith state of the first speaker. The transition probabilities are given by $\Pr\left(s_\ell^a = s^{j,a} | s_{\ell-1}^a = s^{i,a}\right) = p_{ij}^a$; the emission p.d.f. is $p\left(\mathbf{a}_\ell | s_\ell^a = s^{i,a}\right) = \mathcal{N}\left(\mathbf{a}_\ell; \boldsymbol{\mu}_i^a, \mathbf{Q}_i^a\right)$, i.e., normally distributed with mean vector $\boldsymbol{\mu}_i^a$, and a covariance matrix $\mathbf{Q}_i^a$.

HMM_b is defined mutatis mutandis such that there are S^b states with $\Pr\left(s_1^b = s^{i,b}\right) = \pi_i^b$; $\Pr\left(s_\ell^b = s^{j,b} | s_{\ell-1}^b = s^{i,b}\right) = p_{ij}^b$; and $p\left(\mathbf{b}_\ell | s_\ell^b = s^{i,b}\right) = \mathcal{N}\left(\mathbf{b}_\ell; \boldsymbol{\mu}_i^b, \mathbf{Q}_i^b\right)$, where we assume $S^a = S^b$ in sake of simplicity.

The mixture process can be modeled by the FHMM [36] that comprises two underlying HMMs evolving independently over time, each corresponding to a single speaker. At each time instant ℓ, the observed sample $\mathbf{z}_\ell$ depends on the states of HMM_a and HMM_b, emitting the latent outputs $\{\mathbf{a}_\ell, \mathbf{b}_\ell\}$, respectively, such that $\mathbf{z}_\ell = \xi\left(\mathbf{a}_\ell, \mathbf{b}_\ell\right)$, where $\xi\left(\cdot, \cdot\right)$ is the mixing function.

Roweis [36] used the *log-max* mixing function approximation (see [31, 52])

$$\mathbf{z}_\ell \approx \max(\mathbf{a}_\ell, \mathbf{b}_\ell). \tag{3}$$

Nádas computed the probability of $\mathbf{z}_\ell$ analytically. However, in our scenario, as two speech signals are involved, the sparsity of the signals can be utilized

$$p(\mathbf{z}_\ell | s_\ell^a = i, s_\ell^b = j) = \mathcal{N}(\mathbf{z}_\ell; \boldsymbol{\mu}_{ij}, \mathbf{Q}) \tag{4}$$

where $\boldsymbol{\mu}_{ij} = \max(\boldsymbol{\mu}_i^a, \boldsymbol{\mu}_j^b)$ and $\mathbf{Q}$ is the covariance matrix of the observation, assuming that $\forall i, j\, \mathbf{Q}_i^a = \mathbf{Q}_j^b = \mathbf{Q}$.

3 The diffusion framework and the Nyström extension

The diffusion framework is the core computational tool used in our work. The fundamentals of the diffusion framework [50] are detailed in Section 3.1, and the Nyström extension is described in Section 3.2. The asymptotic properties of random walks that pave the way for a novel approach for HMM and FHMM estimation, are discussed in Section 3.3. A systematic approach for estimating the HMM parameters based on the diffusion framework is presented in Section 3.4.

3.1 Diffusion maps

The diffusion framework is an advanced approach for dimensionality reduction [53]. Let $\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$ be a set of L points, such that $\mathbf{x}_i \in \mathbb{R}^d$. Ω is viewed as the nodes of an undirected graph, where the weight $w(\mathbf{x}_i, \mathbf{x}_j)$ of the edge connecting $\mathbf{x}_i$ and $\mathbf{x}_j$ is the affinity between the two nodes. The kernel function $w(\cdot, \cdot)$ is symmetric, nonnegative, and is commonly based on an application-specific distance measure between the points Ω. For instance, in speech processing, it is common to compute the distance between MFCC features [49], or log-spectrum feature vectors. The radial basis function (RBF) kernel is often used:

$$w(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\varepsilon^2\right) = \exp\left(-d_{ij}^2/\varepsilon^2\right) \tag{5}$$

where $d_{ij} \triangleq \|\mathbf{x}_i - \mathbf{x}_j\|^2$. In addition, here $\varepsilon > 0$ such that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar $w(\mathbf{x}_i, \mathbf{x}_j) \sim 1$, and $w(\mathbf{x}_i, \mathbf{x}_j) \sim 0$ if they are dissimilar, implying that the corresponding affinity graph nodes are disconnected. ε is a scale factor quantifying the scale of the similarity, as $\mathbf{x}_i$ and $\mathbf{x}_j$ have nonzero affinity for $d_{ij} < 3\varepsilon$, approximately.

Let $\mathbf{W}$ be the affinity matrix such that $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$, and a corresponding Markov matrix is computed by

$$\mathbf{P} = \mathbf{F}^{-1}\mathbf{W} \tag{6}$$

where $\mathbf{F}$ is a matrix such that $f_{ii} = \sum_j w_{ij}$ and the non-diagonal entries are zeros. p_{ij} can be viewed as the transition probability from $\mathbf{x}_i$ to $\mathbf{x}_j$ in a single time step. Taking the tth power of $\mathbf{P}$ is equivalent to running the Markov chain forward t time steps. This Markov chain has a unique stationary distribution ϕ_0 such that $\phi_0^T P = \phi_0^T$ [50]. In the pre-asymptotic regime, the transition probability from $\mathbf{x}_i$ to $\mathbf{x}_j$ can be expressed using the biorthogonal spectral decomposition:

$$p_t(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l \geq 0} \lambda_l^t \psi_l(\mathbf{x}_i) \phi_l(\mathbf{x}_j), \tag{7}$$

where $1 = \lambda_0 \geq |\lambda_1| \geq |\lambda_2| \geq \ldots$ are the (right) eigenvalues of $\mathbf{P}$, and $\{\psi_l\}$, $\{\phi_l\}$ are the corresponding right and left eigenvectors, respectively. Due to the spectrum decay, the term $p_t(\mathbf{x}_i, \mathbf{x}_j)$ in (7) can be well approximated by summing only a few elements.

The induced Markov chain is utilized to define the *diffusion distance*

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\mathbf{y} \in \Omega} \frac{(p_t(\mathbf{x}_i, \mathbf{y}) - p_t(\mathbf{x}_j, \mathbf{y}))^2}{\phi_0(\mathbf{y})}. \tag{8}$$

This metric evaluates the connectivity of the pair of nodes $\mathbf{x}_i, \mathbf{x}_j$ through the entire graph, by the weighted distance between the conditional probabilities $p_t(\mathbf{x}_i, \cdot)$ and $p_t(\mathbf{x}_j, \cdot)$ induced by the random walk. The diffusion distance can be computed by the eigenvalues and right-eigenvectors of $\mathbf{P}$ [50]

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l \geq 1} \lambda_l^{2t} \left(\psi_l(\mathbf{x}_i) - \psi_l(\mathbf{x}_j)\right)^2. \tag{9}$$

Due to the spectral decay, the diffusion distance $D_t(\mathbf{x}_i, \mathbf{x}_j)$ can be approximated by a relatively low number of eigenvectors in (9). Hence, the right eigenvectors $\{\psi_l\}$ can be used as *a new set of coordinates* for the set Ω, such that the Euclidean distance between these diffusion coordinates $\{\psi_l\}$ approximates the diffusion distance in (9). Let $m(t)$ be the number of terms retained, then the *diffusion map* (embedding) is given by

$$\Psi_t : \mathbf{x}_i \longmapsto \left(\lambda_1^t \psi_1(\mathbf{x}_i), \lambda_2^t \psi_2(\mathbf{x}_i), \ldots, \lambda_{m(t)}^t \psi_{m(t)}(\mathbf{x}_i)\right). \tag{10}$$

for $\mathbf{x}_i \in \Omega$. Note that $\psi_l(\mathbf{x}_i)$ for $1 \le l \le m(t)$ is the ith coordinate of ψ_l, the lth eigenvector of $\mathbf{P}$.

3.2 Nyström extension

The Nyström extension [51, 54] is a numerically efficient approach to extend the embedding vectors $\{\psi_l\}$ defined on a set $\Omega = \{\mathbf{x}_i\}_{i=1}^L$ to a sample point $\tilde{\mathbf{x}} \notin \Omega$. Namely, we aim to compute $\Psi_t(\tilde{\mathbf{x}})$. The crux of the Nyström extension is to compute $\Psi_t(\tilde{\mathbf{x}})$ without having to recompute the embedding of Ω by forming a $(L+1) \times (L+1)$ affinity matrix and its eigenvectors. The Nyström extension is given by

$$\overline{\psi}_l(\tilde{\mathbf{x}}) = \frac{1}{\lambda_l} \sum_{i=1}^{L} p(\tilde{\mathbf{x}}, \mathbf{x}_i) \psi_l(\mathbf{x}_i) \tag{11}$$

for each eigenvector ψ_l, and

$$p(\tilde{\mathbf{x}}, \mathbf{x}_i) = \frac{w(\tilde{\mathbf{x}}, \mathbf{x}_i)}{\sum_{j=1}^{L} w(\tilde{\mathbf{x}}, \mathbf{x}_j)}. \tag{12}$$

The weighting by $1/\lambda_l$ implies that due to the decaying spectrum of the Markov matrix, the Nyström extension can be applied to a limited number of eigenvectors to ensure numerical stability.

Since our models will be trained on a massive amount of data, the Nyström extension can help reduce the dimensions of $\mathbf{P}$ and, thus to reduce the computational burden and to mitigate the storage requirements.

3.3 The asymptotics of random walks and their convergence properties

The diffusion maps scheme utilizes numerically induced random walks to analyze data sets, by forming the diffusion kernel and the corresponding *discrete* eigenfunction $\{\psi_l\}$ computed with respect to the discrete domain (set) $\Omega = \{\mathbf{x}_i\}_{i=1}^L$. This approach aims to study the intrinsic *continuous* manifold $\widehat{\Omega}$ of the data, via its sampled finite manifestation Ω.

Nadler et al. [55] showed that as the number of data points $L \to \infty$ the random walk on the discrete graph manifested by the Markov matrix $\mathbf{P}$, as in (6), converges to a random walk on the continuous space, manifested by the Fokker-Planck operator. The convergence is given by the convergence of the eigenfunctions of the discrete graph to those of the underlying continuous Fokker-Planck operator $\{\widehat{\psi_l}\}$

$$\lim_{N \to \infty} \psi_l = \widehat{\psi_l}. \tag{13}$$

In the continuous domain, systems with potential wells are characterized by their eigenfunctions, where the stable states of the underlying Markov process, corresponding to the potential wells are points of high density in the metric eigenfunctions space $\{\widehat{\psi_l}\}$. Nadler et al. [56] extended these classical results asymptotically to the discrete domain, showing that as the discrete eigenfunction $\{\psi_l\}$ approximate the continuous ones, the points of high density in the discrete domain estimate those of the corresponding continuous ones.

This implies that a diffusion embedding computed as in Section 3.1 with respect to a finite and discrete set of points Ω can be used to approximate the states of a *latent* Markov walk $\widehat{\Omega}$ given its discrete manifestation Ω. In particular, given that the intrinsic representation of the data and corresponding Markov system is assumed to be low dimensional, implies that it can be represented by a few leading eigenvectors. In speech analysis, the low dimensionality of the system stem from the multiple constraints induced on a human speech process by the physical attributes (the anatomical structure of the mouth, tongue etc.), as well as social conventions.

Lafon and Lee [57] studied the quantization of the corresponding Markov chain and graph, aiming to derive a computational approach for recovering the stable meta-states and showed that the optimal quantization with respect to the diffusion distance as in (9) is given by the centroids computed by the K-means quantization scheme with an Euclidean distance metric, in the diffusion domain. Their result stems from the equivalence between the optimal L_2 distances (optimized by the K-means scheme) and the diffusion distances. Hence, given a set of L points, their quantization in the diffusion space allows to optimally approximate the *meta-states* of the *latent* Markov system [57], in terms of diffusion distance distortion.

3.4 Learning an HMM with diffusion maps

We propose to apply the diffusion framework to estimate an HMM model, by computing a reduced dimensionality representation of the training set and estimating the meta-states. This emphasizes the gist of our HMM estimation approach, as the Markov states and transition probabilities can be estimated non-parametrically in the diffusion domain, based on the asymptotics of Markov random walks. Our approach decouples the estimation of the Markov states and the transition probabilities from the estimation of the emission p.d.f.s, and avoids the iterative training of the classical EM-based approach.

Let the set of samples $\Omega = \{\mathbf{x}_\ell\}_{\ell=1}^{L} \in \mathbb{R}^d$ be a sequence of HMM emissions. The HMM has S states with transition probabilities $\{p_{ij}\}_{i,j=1}^{S}$, initial probabilities $\{\pi_i\}_{i=1}^{S}$, mean vectors $\{\boldsymbol{\mu}_i\}_{i=1}^{S}$, and covariance matrices $\{\mathbf{Q}_i\}_{i=1}^{S}$.

The *diffusion embedding* of $\{\mathbf{x}_\ell\}_{\ell=1}^{L}$ is denoted by $\{\bar{\mathbf{x}}_\ell\}_{\ell=1}^{L} \in \mathbb{R}^D$, where $D \ll d$. In order to identify the meta-stable states of the random process manifested by $\{\mathbf{x}_\ell\}_{\ell=1}^{L}$, we quantize the embedding coordinates $\{\bar{\mathbf{x}}_\ell\}_{\ell=1}^{L}$ into S *meta-states*, denoted as $\{s^i\}_{i=1}^{S}$, using the K-means algorithm. The L_2 distance in the embedding domain corresponds to the diffusion distance, and allows to coarsen the corresponding Markov chain.

The computed meta-states are utilized to estimate the parameters of the HMM. The transition probabilities p_{ij} are estimated by running the training samples through the meta-states $\left\{s^i\right\}_{i=1}^{S}$, and accumulating the transitions in a $S \times S$ transition histogram. For that, each sample $\{\bar{\mathbf{x}}_\ell\}_{\ell=1}^{L}$ is associated with its closest meta-state in $\{s^i\}_{i=1}^{S}$, in terms of the L_2 norm. $\boldsymbol{\mu}_i$ is the average of the high-dimensional vectors $\{\mathbf{x}_\ell\}_{\ell=1}^{L}$ belonging to a state s^i. The initial probabilities and the covariance matrices are estimated similarly.

An example demonstrating the learning procedure and its results are given in Appendix.

4 Speech separation by diffusion maps

In this section, we introduce two novel speech separation schemes based on the diffusion framework. In both, we derive data-driven speech models for recovering latent state-space models, where S_a and S_b, are the first and second speakers, respectively, and the FHMM models are trained with respect to them.

4.1 Hybrid-FHMM

We propose to train an HMM model per speaker using the diffusion framework following Section 3.4. Given the speaker's estimated meta-states and the corresponding probabilities found in the training step, the observation (emission) p.d.f.s are computed in the log-spectral domain, using the log-max formulation in (4). With these emission p.d.f.s, the factorial Viterbi algorithm is carried out in the log-spectral domain to infer the states sequences of the speakers, as suggested in [36]. Finally, a masking mechanism is applied to the mixed signal based on the decoded states. We denote this method HFHMM, as it utilizes both the (original) log-spectral domain as well as the (embedded) diffusion domain. The training of the model is carried out in the low-dimensional space, and the inference in the high-dimensional log-spectral domain. By testing the HFHMM (and comparing it to [36]), it will be easy to demonstrate that the new training procedure, at the very least, does not come with performance penalty.

4.1.1 Training phase

Let $u[n]$ be the discrete temporal samples forming the training set of S_a. We start by computing the log-spectral frames $\Omega_a = \{\mathbf{u}_\ell\}_{\ell=1}^{M} \in \mathbb{R}^d$, where each frame $\mathbf{u}_\ell$ comprises $d = K/2 + 1$ frequency bins. The *diffusion embedding* of $\{\mathbf{u}_\ell\}_{\ell=1}^{M}$ is denoted by $\{\bar{\mathbf{u}}_\ell\}_{\ell=1}^{M} \in \mathbb{R}^D$, where $D \ll d$. Throughout this paper, over-bar designate term in the embedded space. In order to identify the meta-stable states of the random process manifested by $\{\mathbf{u}_\ell\}_{\ell=1}^{M}$, we quantize the embedding coordinates $\{\bar{\mathbf{u}}_\ell\}_{\ell=1}^{M}$ into $S^a = S^b = S$ *meta-states*, denoted $\{s^{i,a}\}_{i=1}^{S}$, using the K-means algorithm. Although one can set a different value of states to each speaker, the same value was used for both, in sake of simplicity. S is on the order of tens of states due to the limited number of training data points. The transition probabilities p_{ij}^a and the log-spectral mean vectors $\{\boldsymbol{\mu}_i^a\}_{i=1}^{S}$ are estimated following section 3.4, and a similar procedure is applied to $v[n]$, the temporal samples of S_b.

The training phase is depicted in Fig. 1, where the HMM of the speakers are trained separately, and their coupling is formulated by the observation probability within the FHMM framework.

4.1.2 Test phase (latent state estimation)

The decoding phase of the proposed HFHMM scheme identifies with that of [36]. Let $z[n]= a[n]+b[n]$ be a test mixture signal, where $a[n]$ and $b[n]$ are the utterances from S_a and S_b, respectively, and $\{\mathbf{z}_\ell\}_{\ell=1}^{N} \in \mathbb{R}^d$ its corresponding log-spectral vectors. The observation p.d.f. is modeled in the log-spectral domain utilizing the log-max approximation, similarly to (4). The underlying states of the speakers are estimated using the factorial Viterbi algorithm (see Algorithm 1) given the models inferred in the training phase. The separation is implemented by adaptively masking the speakers (see Section 4.3).

4.2 Dual-FHMM

In the second proposed approach, we derive a novel formulation for estimating the emission (observation) p.d.f.s of the speech mixture directly in the diffusion domain, as opposed to the HFHMM scheme. The meta-states of the underlying Markov processes (modeling the speakers) are estimated by K-means based training in the diffusion domain, as in the previous section.

We propose to synthesize an artificial mixture signal consisting of randomly combined training segments of both speakers. Two FHMM are trained in two different diffusion domains, one FHMM per speaker. Since each diffusion domain is based on the segments of one particular speaker, it is best adapted to that speaker. Denote these models FHMM$_a$ and FHMM$_b$, respectively. In the test phase, we process the input data with both

Algorithm 1 Factorial Viterbi algorithm for the HFHMM scheme

```
1. Preprocessing:
  For $i = 1{:}S^a, j = 1{:}S^b$
      $\tilde{\pi}_i^a = \log \pi_i^a;\quad \tilde{\pi}_j^b = \log \pi_j^b$
      $\tilde{p}_{ij}^a = \log p_{ij}^a;\quad \tilde{p}_{ij}^b = \log p_{ij}^b$
      $\tilde{p}\left(\mathbf{z}_\ell^a | s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b}\right) =$
          $:= \log p\left(\mathbf{z}_\ell^a | s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b}\right).$
2. Forward:
  For $\ell = 2$ to $N$
      For $i = 1{:}S^a, j = 1{:}S^b$
          $v_1(i,j) = \tilde{\pi}_i^a + \tilde{\pi}_j^b +$
              $+\tilde{p}\left(\mathbf{z}_1 | s_1^a = s^{i,a}, s_\ell^b = s^{j,b}\right)$
          $v_\ell(i,j) =$
          $= \max_{\substack{1 \le r \le S^a \\ 1 \le q \le S^b}} \left\{ v_{\ell-1}(r,q) + \tilde{p}_{ri}^a + \tilde{p}_{qj}^b \right\} +$
              $+\tilde{p}\left(\mathbf{z}_\ell | s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b}\right)$
          $\delta_\ell(i,j) =$
          $= \arg\max_{\substack{1 \le r \le S^a \\ 1 \le q \le S^b}} \left\{ v_{\ell-1}(r,q) + \tilde{p}_{ri}^a + \tilde{p}_{qj}^b \right\}.$
3. Set $\left(\hat{s}_N^a, \hat{s}_N^b\right) = \arg\max_{\substack{1 \le i \le S^a \\ 1 \le j \le S^b}} \delta_N(i,j).$
4. Backward:
  For $\ell = N-1$ to $1$
      $\left(\hat{s}_\ell^a, \hat{s}_\ell^b\right) = \delta_{\ell+1}\left(\hat{s}_{\ell+1}^a, \hat{s}_{\ell+1}^b\right).$
```

models, where at each time segment, FHMM$_a$ is used to infer the latent state of S_a, and FHMM$_b$ infers the state of S_b.

4.2.1 Training phase

We detail the computation of FHMM$_a$, the FHMM defined in the diffusion embedding domain of S_a, and the procedure is applied to FHMM$_b$ mutatis mutandis. In general, in order to derive an FHMM, two quantities need to be estimated. First, the states and the corresponding transition probabilities for each speaker (Markov process), and second, the observation p.d.f. associating an input measurement with an underlying Markov states. The meta-states of S_a, i.e. $\left\{s^{i,a}\right\}_{i=1}^S$, the meta-states of S_b, $\left\{s^{i,b}\right\}_{i=1}^S$, and the corresponding $S \times S$ transition matrices are computed similarly to Section 4.1.1.

In order to find the *observation p.d.f.* of a mixture signal, we first embed $\{\mathbf{u}_\ell\}_{\ell=1}^M$, the training set of S_a, to yield $\{\bar{\mathbf{u}}_\ell\}_{\ell=1}^M$ and the corresponding eigenvalues $\{\lambda_i^u\}_{i=1}^D$. Similarly, $\{\bar{\mathbf{v}}_\ell\}_{\ell=1}^M$ is the embedding of the training sequences of S_b into his diffusion domain.

The observation p.d.f. is estimated in the same diffusion domain, by synthesizing the mixture signal:

$$w[n] = u[n] + v[n]\,. \tag{14}$$

Since both $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ have S Markov states, $\mathbf{w}_\ell$ has possible S^2 states. We define the mixture's observation p.d.f. as

$$P\left(\bar{\mathbf{w}}_\ell^a | s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b}\right) = \mathcal{N}\left(\bar{\mathbf{w}}_\ell^a; \bar{\boldsymbol{\mu}}_{ij}^a, \bar{\mathbf{Q}}_{ij}^a\right), \tag{15}$$

where $\bar{\mathbf{w}}_\ell^a$ is the embedding of $\mathbf{w}_\ell$ in the diffusion domain of S_a using the Nyström extension. That is, $\mathbf{w}_\ell$ substitutes $\tilde{\mathbf{x}}$ in (11), and N, λ_l, and $\mathbf{x}_i$ are substituted by M, λ_i^u and $\mathbf{u}_\ell$, respectively. s_ℓ^a and s_ℓ^b are the corresponding states of S_a and S_b, at time ℓ. $\bar{\boldsymbol{\mu}}_{ij}^a$ and $\bar{\mathbf{Q}}_{ij}^a$ are the mean and the covariance of the mixture embedding $\bar{\mathbf{w}}_\ell^a$ related to the states $s_\ell^a = s^{i,a}$ and $s_\ell^b = s^{j,b}$. Namely, let

$$\gamma_{ij}^a \triangleq \left\{\bar{\mathbf{w}}_\ell^a | \bar{\mathbf{u}}_\ell \in s^{i,a} \text{ and } \bar{\mathbf{v}}_\ell \in s^{j,b}\right\}. \tag{16}$$

be the set of points in the mixture embedding, associated with the states $s^{i,a}$ and $s^{j,b}$, then the estimates of $\bar{\boldsymbol{\mu}}_{ij}^a$ and $\bar{\mathbf{Q}}_{ij}^a$ are

$$\hat{\bar{\boldsymbol{\mu}}}_{ij}^a = \frac{1}{|\gamma_{ij}^a|} \sum_{\bar{\mathbf{w}}_\ell^a \in \gamma_{ij}^a} \bar{\mathbf{w}}_\ell^a \tag{17}$$

and

$$\hat{\bar{\mathbf{Q}}}_{ij}^a = \frac{1}{|\gamma_{ij}|} \sum_{\bar{\mathbf{w}}_\ell^a \in \gamma_{ij}} \left(\bar{\mathbf{w}}_\ell^a - \hat{\bar{\boldsymbol{\mu}}}_{ij}^a\right)\left(\bar{\mathbf{w}}_\ell^a - \hat{\bar{\boldsymbol{\mu}}}_{ij}^a\right)^T. \tag{18}$$

In contrast to the HFHMM where all states share a single covariance matrix (in the high-dimensional domain), in the DFHMM we chose to define a distinct covariance matrix (in the low-dimensional space) for each state, so the algorithm is as general as possible.

The training procedure of FHMM$_a$ and FHMM$_b$ is summarized in Algorithm 2. The overall scheme is depicted in Fig. 2.

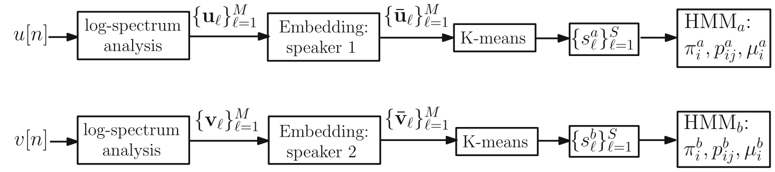

$u[n]$
log-spectrum analysis
$\{\mathbf{u}_\ell\}_{\ell=1}^M$
Embedding: speaker 1
$\{\bar{\mathbf{u}}_\ell\}_{\ell=1}^M$
K-means
$\{s_\ell^a\}_{\ell=1}^S$
HMM$_a$: $\pi_i^a, p_{ij}^a, \mu_i^a$
$v[n]$
log-spectrum analysis
$\{\mathbf{v}_\ell\}_{\ell=1}^M$
Embedding: speaker 2
$\{\bar{\mathbf{v}}_\ell\}_{\ell=1}^M$
K-means
$\{s_\ell^b\}_{\ell=1}^S$
HMM$_b$: $\pi_i^b, p_{ij}^b, \mu_i^b$

Fig. 1 Training in HFHMM scheme. The training phase of the hybrid log-spectral-diffusion approach, for speaker 1 (*top*) and speaker 2 (*bottom*)

Algorithm 2 Dual FHMM training

```
Compute the log-spectra $\{\mathbf{u}_\ell, \mathbf{v}_\ell\}_{\ell=1}^M$ of the training signals $u[n]$ and $v[n]$.
Compute $\{\lambda_\ell^u, \bar{\mathbf{u}}_\ell\}_{\ell=1}^M$ and $\{\lambda_\ell^v, \bar{\mathbf{v}}_\ell\}_{\ell=1}^M$, the embedding of $\{\mathbf{u}_\ell\}_{\ell=1}^M$ and $\{\mathbf{v}_\ell\}_{\ell=1}^M$, respectively.
Estimate the HMM models of $S_a$ and $S_b$ as in Fig. 1.
Define the synthetic mixture $w[n] = u[n] + v[n]$, and compute its log-spectrum $\{\mathbf{w}_\ell\}_{\ell=1}^M$.
Embed $\{\mathbf{w}_\ell\}_{\ell=1}^M$ onto the diffusion space of $Sa$ using the Nyström Extension and obtain $\left\{\bar{\mathbf{w}}_\ell^a\right\}_{\ell=1}^M$.
Embed $\{\mathbf{w}_\ell\}_{\ell=1}^M$ onto the diffusion space of $Sb$ using the Nyström Extension and obtain $\left\{\bar{\mathbf{w}}_\ell^b\right\}_{\ell=1}^M$.
Compute $\left\{\bar{\boldsymbol{\mu}}_{ij}^a, \bar{\boldsymbol{\mu}}_{ij}^b, \bar{\mathbf{Q}}_{ij}^a, \bar{\mathbf{Q}}_{ij}^b\right\}_{i,j=1}^S$, the Gaussian parameters of the observation p.d.f.s using (16)–(18).
```

4.2.2 Latent state estimation

In the test phase a mixed utterance $z[n] = a[n] + b[n]$ is measured, where $a[n]$ and $b[n]$ are the (unknown) separate speech signals. The latent states corresponding to $z[n]$ are estimated by embedding $\mathbf{z}_\ell$ on the two diffusion spaces, yielding $\{\bar{\mathbf{z}}_\ell^a\}_{\ell=1}^N$ and $\{\bar{\mathbf{z}}_\ell^b\}_{\ell=1}^N$, and applying the embedded domain FHMMs, namely FHMM$_a$ and FHMM$_b$, to $\{\bar{\mathbf{z}}_\ell^a\}_{\ell=1}^N$ and $\{\bar{\mathbf{z}}_\ell^b\}_{\ell=1}^N$, respectively. Each FHMM is used to infer the latent state of the speaker used for its own embedding. Hence, we use FHMM$_a$ only to recover the state sequence of S_a, while discarding the states sequence obtained for S_b. The states sequences are recovered using the factorial Viterbi algorithm with the parameters of FHMM$_a$. It is identical to Algorithm 1, with $\bar{\mathbf{z}}_\ell^a$ substituting $\mathbf{z}_\ell$. Similarly, FHMM$_b$ is used to estimate the latent states of S_b, i.e., using Algorithm 1 with $\bar{\mathbf{z}}_\ell^b$ instead of $\mathbf{z}_\ell$.

The gist of this approach, as can be deduced from Section 3.3, is that an embedding space, either FHMM$_a$ or FHMM$_b$, encodes the speech attributes of the respective speaker, and hence would best estimate the latent states of the corresponding speaker.

The procedure for estimating the latent state sequence is summarized in Fig. 3.

4.3 Masking

Masking is a common approach in speech separation given latent states, that is often implemented in the STFT domain, which provides a sparse representations of speech signals. The separated log-spectral vectors of the test signal are reconstructed by associating each frequency bin of the input signal $\mathbf{z}_\ell$, with either S_a or S_b. There are various ways to define the mask, and here we stick to [36]. Formally, given the estimates of latent states $s_\ell^a = s^{i,a}$ and $s_\ell^b = s^{j,b}$, Roweis proposed [36] to estimate the log-spectral domain vector of S_a by

$$\hat{\mathbf{a}}_\ell^k = \begin{cases} \mathbf{z}_\ell^k & \boldsymbol{\mu}_i^a(k) > \boldsymbol{\mu}_j^b(k) \\ \mathbf{m}_0(k) & \text{else} \end{cases}, \tag{19}$$

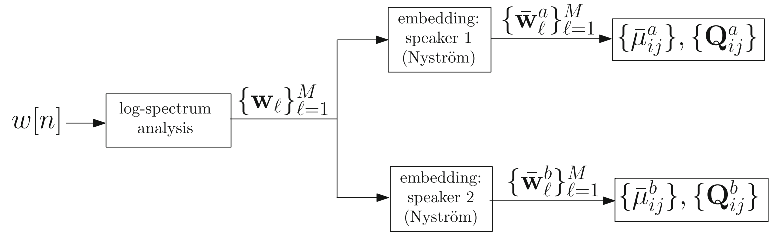
where $\mathbf{m}_0(k)$ is a tunable parameter that determines the attenuation of the frequency bins. Also recall that k is the frequency index. In [36], it is proposed to set $\mathbf{m}_0(k) = -\infty$, $\forall k$ resulting in a hard mask. As the use of a hard mask might result in noticeable distortions and artifacts in the output signals, we applied a soft estimator instead, by setting $\mathbf{m}_0(k) = \boldsymbol{\mu}_i^a(k)$

$$\hat{\mathbf{a}}_\ell^k = \begin{cases} \mathbf{z}_\ell^k & \boldsymbol{\mu}_i^a(k) > \boldsymbol{\mu}_i^b(k) \\ \boldsymbol{\mu}_i^a(k) & \text{else} \end{cases}. \tag{20}$$

In this estimator, the log-spectral content of the weaker source is not attenuated as in (19), but synthesized according to the estimated HMM. This masking was shown by Radfar and Dansereau [35] to correspond to the MMSE estimator given a zero model error. Recovering the log-spectrum of S_b is carried out mutatis mutandis.

5 Experimental results

The proposed HFHMM and DFHMM schemes were experimentally verified by studying common state-of-the-art speech separation tasks. The quality of the result is evaluated using both objective criteria and (informal) listening tests. The proposed schemes are compared to the separation scheme proposed by Roweis [36] (for both hard and soft masks), the iterative FHMM-based estimator by Hu and Wang [38], and to the MIXMAX estimator by Radfar and Dansereau [35].



$w[n]$
log-spectrum analysis
$\{\mathbf{w}_\ell\}_{\ell=1}^M$
embedding: speaker 1 (Nyström)
$\{\bar{\mathbf{w}}_\ell^a\}_{\ell=1}^M$
$\{\bar{\mu}_{ij}^a\}, \{\mathbf{Q}_{ij}^a\}$
embedding: speaker 2 (Nyström)
$\{\bar{\mathbf{w}}_\ell^b\}_{\ell=1}^M$
$\{\bar{\mu}_{ij}^b\}, \{\mathbf{Q}_{ij}^b\}$

Fig. 2 The Dual FHMM training scheme

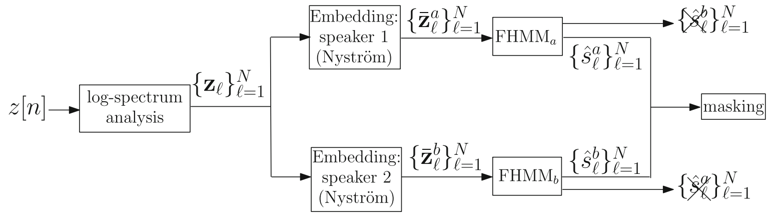

$z[n]$
log-spectrum analysis
$\{\mathbf{z}_\ell\}_{\ell=1}^N$
Embedding: speaker 1 (Nyström)
$\{\bar{\mathbf{z}}_\ell^a\}_{\ell=1}^N$
FHMM$_a$
$\{\hat{s}_\ell^b\}_{\ell=1}^N$
$\{\hat{s}_\ell^a\}_{\ell=1}^N$
masking
Embedding: speaker 2 (Nyström)
$\{\bar{\mathbf{z}}_\ell^b\}_{\ell=1}^N$
FHMM$_b$
$\{\hat{s}_\ell^b\}_{\ell=1}^N$
$\{\hat{s}_\ell^a\}_{\ell=1}^N$

Fig. 3 The DFHMM state inference scheme. The states are inferred by applying two FHMMs. Each FHMM infers the states of the speaker whose training set was used to compute its embedding

5.1 Experimental setup

A training set of 450 noiseless sentences per speaker drawn from the speech separation challenge [58] is used. Each sentence is 1–2 s long, and was down-sampled from 25 kHz to 8 kHz to shorten the running time of the code. The STFT was computed using 256 samples long frames, having an overlap of 128 samples between successive frames (50 % overlap). Consequently, each log-spectral STFT feature vector is 129 coefficients long, and a Hann window was used in both the analysis and synthesis stages. On average, the training set of each speaker consisted of 55,000 log-spectral vectors.

The application of the diffusion embedding to the training set is carried out in two steps. First, 30 random sentences per speaker are embedded by extracting the eigenvectors of the Markov matrix defined by the diffusion framework. In the second stage, the remainder of the 420 sentences are embedded by applying the Nyström extension. This embedding scheme was chosen due to complexity and memory considerations.

From the complexity aspect, the dimensionality of the Markov matrix determines the number of operations for the DFHMM in the test phase. The embedding of the mixed signal involves the Nyström extension, which is calculated via (11) and thus affected proportionally to the dimensionality of the Markov matrix. It can also be deduced from (12) that the samples creating the Markov matrix, used for calculating the weight functions measuring the graph connectivity, should be kept in the memory.

An RBF kernel is used to compute the spectral embedding, with kernel bandwidth ϵ. In general, a kernel bandwidth that is too large can result in HMM states which are almost identical, since all the data points are fully connected. An excessively small ϵ, might result in a model consisting of mostly disconnected graph nodes, with an increased number of states, that might be computationally intractable. A kernel bandwidth $\epsilon \sim 110$ was found to be a good compromise, as it retains 5 % of the edges connected. This is a common approach used in previous works on diffusion based embedding [59], where the embedding was shown to be robust to the kernel bandwidth.

Each FHMM uses $S = 70$ meta-states computed by applying K-means with Euclidean distance measure to the embedded vectors (see Section 3.4). The proposed schemes is evaluated using different combinations of speakers' gender: male-female, male-male and female-female, where each combination is tested using four pairs of speakers, each pair contributing 15 mixtures. Therefore, each gender combination is evaluated using 60 sentences. The pairs of speakers (numbers refer to [58]) are listed in Table 1. The individual signals are noiseless, and the source to interference ratio (SIR) of the mixed signals is set to 0 dB for all experiments (to comply with our model).

When generating the mixed signal $w[n]$ for the DFHMM scheme (refer to (14)), each of the signals $u[n]$ and $v[n]$ was created by concatenating utterances from the database in a random order. This implementation stems from the unique structure of the utterances. Each sentence is composed of six words that are ordered in the following manner: command, color, preposition, letter, number and an adverb. For example, a valid sentence is "bin blue at Z three please". Each component of the utterances has a final set of possible values. For instance, the command word can be only one of the following: "bin", "lay", "place", or "set". If $u[n]$ and $v[n]$ are summed without shuffling the utterances from the database, an undesired situation can occur in which the mixture of the signals depicts only states in which the speakers utter the same word.

Several variants of the proposed schemes were implemented to assess the influence of the various components on the performance. First, an ideal DFHMM (iDFHMM)

Table 1 Tested speakers. The pairs of speakers used for testing each algorithm, each pair contributing 15 sentences

Male+male	Male+female	Female+female
1 + 32	14 + 25	15 + 20
14 + 30	19 + 20	18 + 29
19 + 28	26 + 34	22 + 33
26 + 27	32 + 23	16 + 31

For the male-female case, the left number is associated with the male, and the right with the female with the accurate factorial states, instead of their estimated counterparts, is implemented by computing the embeddings and the meta-states of the separate (unmixed) speakers.

Second, the hard mask (19) and the soft mask (20) are compared, and the corresponding schemes are denoted as DFHMM-H (hard mask), DFHMM-S (soft mask), iDFHMM-H (idealized DFHMM with hard mask) etc.

Third, two training schemes are compared. The first, uses the entire training set to form the Markov matrix $\mathbf{P}$, while in the second, the matrix is based on 30 sentences only and the Nyström extension is used to embed the remaining sentences. These variants are denoted as HFHMM-H-E, HFHMM-S-E for the exact embedding, and HFHMM-H-N, HFHMM-S-N for the procedure that utilizes the Nyström extension. Only the Nyström extension based training is used in the DFHMM scheme, as detailed earlier.

The proposed approaches are compared with contemporary state-of-the-art schemes: (1) the work of Roweis, using 70 HMM states per speaker, that are inferred by the EM procedure. The HMMs are used to define an FHMM, as detailed in [36]; (2) the iterative algorithm by Hu and Wang [38], with the separation part implemented by inferring FHMM states for the mixed signal and then applying MAP estimation. As recommended in [38], the FHMM comprises 256 Gaussians per speaker, and a maximum number of 4 iterations is allowed. To reduce the computational complexity, the beam search uses the 16 most likely preceding state pairs.

The MIXMAX estimator by Radfar and Dansereau [35] uses 512 mixtures per GMM, that are trained on the same 450 sentences as the HFHMM and DFHMM schemes. Such a high dimensional GMM per speaker imposes a heavy computational burden. Therefore, we separated the mixed signals only the most probable states pair [35] and, as indicated by the authors, this procedure achieves comparable scores to that of full estimation. Finally, in order to reduce some of the artifacts and distortions of the hard mask, we set $\mathbf{m}_0(k) = -8, \forall k$ for the algorithm proposed by Roweis [36], and for the HFHMM and the DFHMM schemes (when the hard mask is applied). This value was chosen to reflect the average level of low power time-frequency bins. It achieved the best balance between speech intelligibility and separation performance.

5.2 Figures-of-merit

In order to quantify the performance of the proposed separation schemes, we utilized the SIR, source to distortion ratio (SDR) and source to artifacts ratio (SAR) criteria, proposed by Vincent et al. [60] and implemented as a Matlab toolbox [61]. The SIR measures the attenuation of the interference with respect to the desired speech signal, and the SAR evaluates the level of artifacts (e.g. musical noise) in the processed signal. The SDR is the desired signal level with respect to the total contribution of all the other distortion factors. The SDR and the SAR criteria are informative when a hard mask is applied. The outcome of the algorithm was also assessed by *informal* listening tests.

5.3 Results

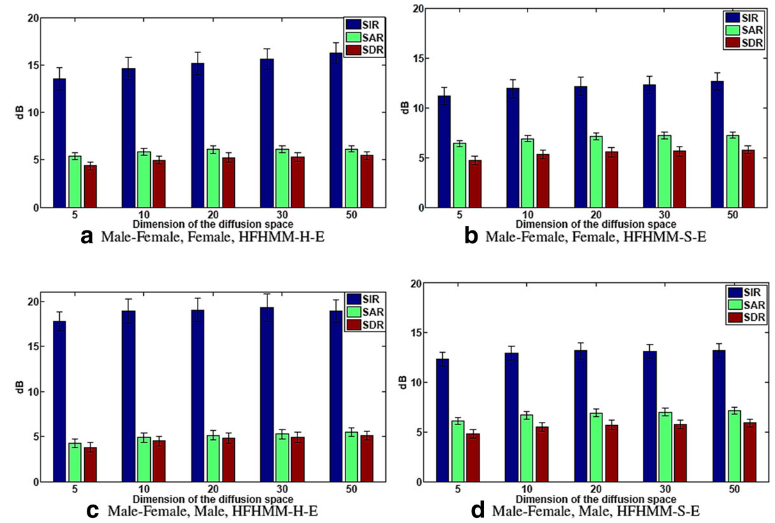
5.3.1 HFHMM

We start by evaluating the performance of the HFHMM scheme. The results are depicted in Fig. 4 for the exact diffusion training. The figure presents the results for the male-female case. It indicates that for the male speaker, 30 dimensions yield the best score, whereas for the female speaker $D = 50$ is better than $D = 30$ by 0.7 dB. For the same gender mixture a similar trend is observed, and thus not reported in the figure. The results for the training procedure that incorporates the Nyström extension are less satisfying. Consequently, they are not extensively presented due to space limitations. The quantitative results are reported in Table 2, with $D = 30$ for the male speaker and $D = 50$ for the female speaker across all mixtures. For the male-female mixtures, we report the results related to each gender separately, and for male-male and female-female pairs, we extracted both speakers, and averaged the results. It follows that the HFHMM-S approach outperforms the HFHMM-H formulation in both the SDR and SAR figures-of-merit for most mixtures, although a degradation in the SIR is encountered. The results indicate a performance gap between the HFHMM-H-N, HFHMM-S-N and the HFHMM-H-E, HFHMM-S-E, in favor of the latter. Consequently, we conclude that the Nyström extension leads to performance degradation.

5.3.2 DFHMM

Sonograms and time-frequency maps of the DFHMM-H and DFHMM-S schemes for the male-female mixture are depicted in Figs. 5 and 6. In Fig. 6, the white regions correspond to time-frequency bins associated with the female speaker, while the darker ones with the male speaker. It follows that the mask resembles the clustering of the mixed signal (but not perfectly so). Figure 7 depicts the performance metrics of the DFHMM-H estimator and its ideal counterpart iDFHMM-H as a function of the dimensionality for the male-female mixture. It follows that an embedding space of $D = 20$ suffices for all schemes and the iDFHMM-H outperforms the DFHMM-H by close to 20 % for the male speaker, and $\sim$ 10 % for the female speaker.

The results of applying a soft mask to the male-female mixture are similar to those of the HFHMM. As expected, the SAR and SDR measures indicate that the DFHMM-S yields lower distortion levels as compared with the DFHMM-H scheme. However, the SIR measure deteriorates. This can be attributed to the higher level of



a Male-Female, Female, HFHMM-H-E

b Male-Female, Female, HFHMM-S-E

c Male-Female, Male, HFHMM-H-E

d Male-Female, Male, HFHMM-S-E

Fig. 4 HFHMM, separation results for the male-female mixture. Speech separation results measured by SDR, SIR and SAR of the HFHMM scheme for the male-female mixtures. Confidence intervals are also indicated. Training stage was based on exact diffusion maps, without utilizing the Nyström extension

residual interference, being a consequence of the softer mask.

We also subjectively compared the DFHMM-H and the DFHMM-S. The male speaker was recovered by the DFHMM-H without noticeable artifacts. However, the separated female speech sometimes sounds disrupted when a hard mask is used, and applying a soft mask resolved this artifact. Similar trends were observed for the male-male combination.

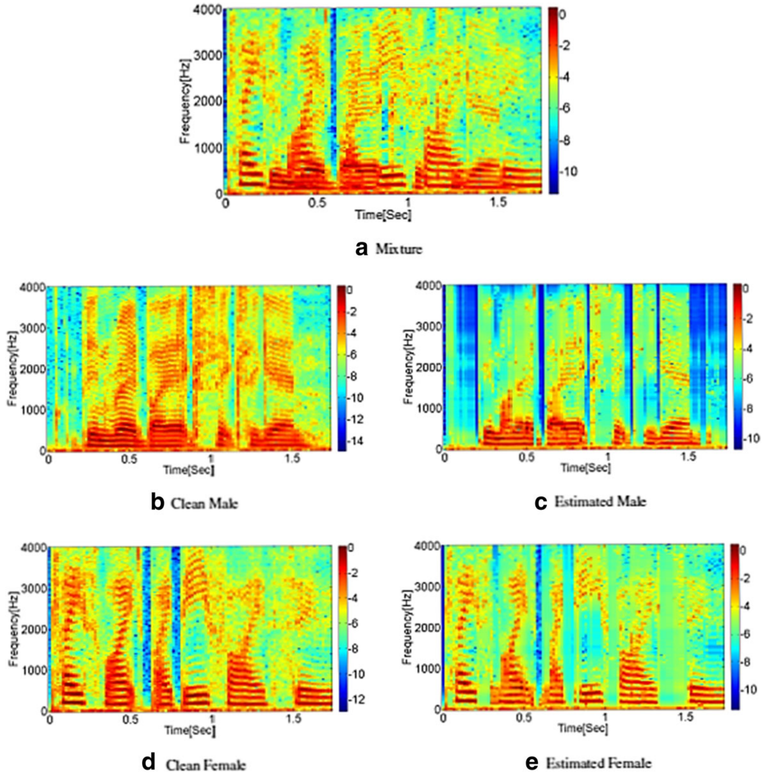
Application of the DFHMM-H resulted in audible artifacts, that are evident by the low SAR and SDR levels. As in the male-female mixture, applying the DFHMM-S, improves the results, and the ideal estimators, iDFHMM-H and iDFHMM-S, respectively, outperform their non-ideal counterparts. We attribute this to the possible overlap of the meta-state of speakers of the same gender. The female-female mixtures, exhibit similar results to the male-male case.

We further study the effect of the overlapping spectral components, by depicting the hard mask of the female-female mixtures, obtained by the DFHMM-H scheme, in Fig. 8. The white regions of the mask correspond to the time-frequency bins, where S_b is estimated to have higher spectral content. As most of the mask is white, this indicates that S_b is the dominant speaker in the segment. We fail to identify S_a accurately due to the overlapping

Table 2 Quantitative results

	Male+female						Male+male			Female+female			Average		
	Male			Female			-			-			-		
	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR
Hu and Wang [38]	16.2	7.3	6.5	15.3	8.0	7.0	**12.6**	6.0	4.6	**11.6**	5.5	4.0	14.0	6.7	5.5
Roweis-H	18.8	5.4	5.0	15.8	6.1	5.3	10.8	3.3	1.8	9.3	2.6	0.9	13.7	4.3	3.2
Roweis-S	13.0	7.2	5.8	12.6	7.2	5.7	7.7	5.5	2.6	6.5	4.8	1.7	9.9	6.2	3.9
HFHMM-H-E	**19.4**	5.4	5.1	**16.0**	6.2	5.4	11.1	3.3	1.8	9.9	3.0	1.3	**14.1**	4.5	3.4
HFHMM-S-E	13.2	7.2	5.9	12.5	7.2	5.7	7.7	5.5	2.6	6.9	5.2	2.1	10.1	6.3	4.1
HFHMM-H-N	16.9	4.7	4.1	14.4	5.3	4.4	10.6	3.1	1.6	6.4	1.9	-0.8	12.1	3.7	2.3
HFHMM-S-N	12.1	6.4	5.0	11.7	6.3	4.8	7.5	5.2	2.4	4.5	4.0	0.0	8.9	5.5	3.0
DFHMM-H	13.7	5.0	4.1	15.5	4.6	4.0	9.4	2.7	0.7	7.0	4.0	-0.4	11.4	3.9	2.1
DFHMM-S	10.8	6.4	4.7	12.7	5.6	4.5	6.7	4.8	1.5	5.6	4.2	0.4	8.9	5.2	2.3
MIXMAX	15.6	**8.6**	**7.6**	15.1	**8.6**	7.4	10.7	**6.8**	**4.8**	9.7	6.3	**4.1**	12.3	**7.6**	**6.0**

Separation performance of the different schemes. The leading result per category is marked in bold



a Mixture

b Clean Male

c Estimated Male

d Clean Female

e Estimated Female

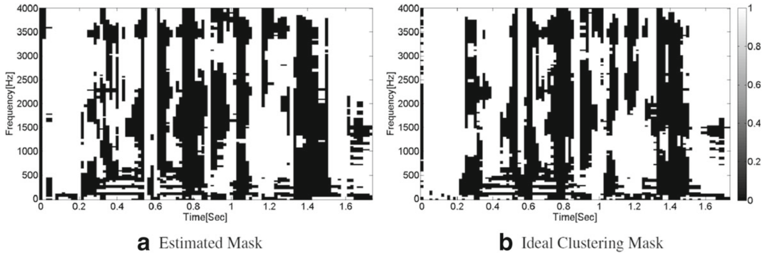
Fig. 5 DFHMM, illustrative sonograms for the male-female mixture. Male-female mixture. Sonograms of the clean male speech, the clean female speech, the mixture signal and the estimated sources are depicted. The estimated signals were constructed by the DFHMM-S scheme, for a more informative presentation

spectral content of the same gender speakers, especially in the lower frequency band.

5.3.3 *Comparison with competing algorithms*

In this section, we compare the performance of our proposed algorithms to several single microphone separation algorithms, namely the algorithms proposed by Roweis [36], Hu and Wang [38], specifically the iterative FHMM-based inference and MAP estimator variant[1], and the MIXMAX-based separation scheme [35][2]. For implementing the proposed algorithms we have set the following parameters: for the DFHMM scheme we used $D = 30$. For the HFHMM we used $D = 30$ for the male speaker and $D = 50$ for the female speaker. Note that increasing D in the HFHMM only influences its training phase.

The comparative study is summarized in Table 2. For the male speaker in male-female mixture, the HFHMM-H-E outperforms the other estimators with respect to the SIR measure. However, it obtains lower SAR and SDR than the MIXMAX algorithm. Hu and Wang, Roweis-S and HFHMM-S-E also obtain good SAR score, but worse than the MIXMAX. The HFHMM-H-E has a better SIR result also for the female speaker, with the DFHMM-H,



a Estimated Mask

b Ideal Clustering Mask

Fig. 6 DFHMM, illustrative masks for the male-female mixture. Male-female mixture. Sonograms of the estimated and ideal masks constructed by the DFHMM-H scheme, corresponding to Fig. 5. White regions are bins associated with the female speaker and the black bins with the male speaker

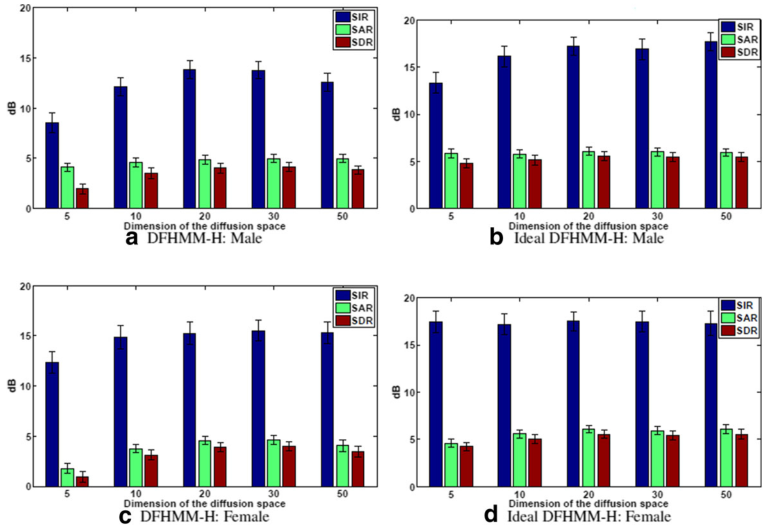


Fig. 7 DFHMM, results for the male-female scenario. **a**, **c** Male-female separation using the DFHMM-H. **b**, **d** using iDFHMM-H, respectively

Roweis-H scoring below. The SAR and SDR of the MIXMAX are again superior to the respective measures obtained by the competing algorithms. The algorithm of Hu and Wang also exhibits satisfactory SAR and SDR, although still being inferior to the score obtained by the MIXMAX.

For the male-male mixture, the best SIR result is obtained by Hu and Wang algorithm. The second best results are obtained by the HFHMM-H-E, and then by Roweis-H, HFHMM-H-N and the MIXMAX, respectively, with similar performance. The MIXMAX scores the best results in the SAR and SDR measures. The algorithm of Hu and Wang demonstrates better measures for the female-female mixtures in terms of SIR, as well. Again, also for these mixtures the MIXMAX gains the highest SAR score, and shares the best SDR score with Hu and Wang algorithm. However, the HFHMM-S-E and Roweis-S also obtain good SAR results.

By looking at the overall performance, described in the three right-hand-side columns of Table 2, it follows that the HFHMM-H-E and Hu and Wang iterative algorithm obtained the best SIR. The HFHMM-S-E and Hu and Wang also demonstrate reasonable SAR. However, the best SAR and SDR performance was achieved by the MIXMAX algorithm. It is also indicated that using the Nyström extension leads to a degraded performance,

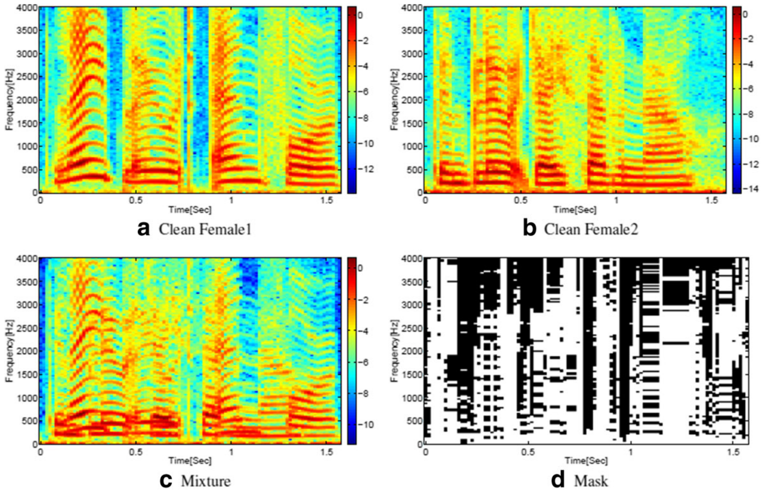


Fig. 8 DFHMM, illustrative mask for a female-female mixture. Masking results for the female-female mixture using the DFHMM-H scheme. Sonograms of the clean speech of the first female, the second female and the corresponding mixture signal. The resulting hard mask is also depicted with white regions correspond to speaker 2 active

which might explain the relatively disappointing scores of the DFHMM schemes.

Informal listening tests of all estimators and scenarios, demonstrate that there is a large room for improvement. While we notice a slight advantage to the MIXMAX and Hu and Wang algorithms over the proposed algorithms, we claim that the performance differences are rather marginal. Several examples can be found in our website[3].

Analyzing both objective results and the informal listening test, we can also observe that better results are obtained for the female-male mixtures, as compared with the female-female and male-male mixtures. This may be attributed to the higher spectral content overlap of the latter two.

We attribute the superior performance of MIXMAX algorithm with respect to the SAR and SDR metrics to the different separation scheme it utilizes. The proposed DFHMM and HFHMM schemes, as well as [36], estimate a *single* dominant latent state per time-frame, to yield the separation mask, making it susceptible to state estimation errors. In contrast, the MIXMAX estimator utilizes a weighted sum of state estimates

$$\hat{\mathbf{a}}_\ell = \sum_{ij} p\left(s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b} | \mathbf{z}_\ell\right) \hat{\mathbf{a}}_\ell^{ij} \tag{21}$$

where $\hat{\mathbf{a}}_\ell^{ij}$ is the estimation of $\hat{\mathbf{a}}_\ell$ given $\mathbf{z}_\ell$ and $\left(s_\ell^a = s^{i,a}, s_\ell^b = s^{j,b}\right)$. The iterative algorithm of Hu and Wang also utilizes a more sophisticated soft masking procedure (with respect to the procedure discussed in Section 4.3) and hence yields good SAR and SDR. The iterative adaptation of the pre-trained HMMs of the speakers might explain the good SIR performance of Hu and Wang algorithm.

6 Computational complexity of the DFHMM

One of the main attributes of the DFHMM algorithm is its computational efficiency (with respect to [36, 38]) due to the use of the low-dimensional embedding. The HFHMM has identical complexity as in [36], since they differ only in the training stage.

The application of the DFHMM algorithm consists of the following steps: spectral analysis and logarithm calculation, Nyström extension, factorial Viterbi algorithm, filtering (masking), and spectral synthesis. The procedure in [36] is similar, with the Nyström extension omitted. Another difference is the dimensionality of the factorial Viterbi algorithm. In the DFHMM, it is applied in the (low-dimensional) embedded domain of the mixed signal, whereas in [36] in the (high-dimensional) log-spectrum domain.

It therefore suffices to analyze the computational requirements of the Nyström extension for the DFHMM, and the factorial Viterbi algorithm in order to compare the computational requirements of both techniques. The number of HMM states for each speaker is S. The analysis refers to a single log-spectral vector of the mixed signal.

6.1 Nyström extension

The Nyström Extension is used to embed a log-spectral vector of the mixed signal using (5), (11), and (12)

$$\lambda_l \overline{\psi}_l(\tilde{\mathbf{x}}) = \sum_{i=1}^{L} p(\tilde{\mathbf{x}}, \mathbf{x}_i) \psi_l(\mathbf{x}_i). \tag{22}$$

It follows that the number of operations depends on the number of samples in $\Omega = \{\mathbf{x}_i\}_{i=1}^L$, namely LD additions and multiplications.

The computation of the embedding for each point in the dataset $\mathbf{x}_i \in \Omega$ involves the computation of the kernel (5), requiring d multiplications and additions. The exponent can be computed using a lookup table (LUT). Hence, (5) is implemented by Ld multiplications and additions, and L LUT indexing operations. Finally, note that in (12), the denominator is the same for every $\mathbf{x}_i \in \Omega$. Consequently, only additional L multiplications and additions are required.

The total number of operations of the Nyström extension is therefore $L(d+D+1)$ multiplications and additions, and L LUT indexing operations.

6.2 Factorial Viterbi algorithm

The factorial Viterbi is utilized by both the DFHMM and [36], and applied to data of different dimensionality. We start by analyzing the number of operations required by Roweis's approach [36], as summarized in Algorithm 1. At the preprocessing phase, all expressions are evaluated in advance, except for the p.d.f. $p(\mathbf{z}_\ell | s_\ell^a = i, s_\ell^b = j)$. Writing this p.d.f. explicitly, we have

$$\begin{aligned} p\left(\mathbf{z}_\ell | s_\ell^a = i, s_\ell^b = j\right) \\ = \frac{1}{\sqrt{(2\pi)^D |Q|}} \exp\left\{-\frac{1}{2}\mathbf{h}_{\ell ij}^T \mathbf{Q} \mathbf{h}_{\ell ij}\right\} \end{aligned} \tag{23}$$

where $|\mathbf{Q}|$ is the determinant of the covariance matrix, and

$$\mathbf{h}_{\ell ij} = \mathbf{z}_\ell - \boldsymbol{\mu}_{ij}.$$

This analysis relates to each time instant ℓ. The normalization of the Gaussian can be discarded, as it does not affect the maximization. The computational complexity can be further reduced by maximizing the logarithm of the p.d.f.. Hence, only $\mathbf{h}_{\ell ij}^T \mathbf{Q} \mathbf{h}_{\ell ij}$, $i, j = 1, 2, \ldots, S$ should be calculated, requiring $S^2 \times (d^2 + d)$ multiplications and additions. In the forward stage, the computation of $\nu_\ell(i,j)$ for all states requires approximately S^2 additions. Note that the term $\tilde{p}_{ri}^a + \tilde{p}_{qj}^b$ can be calculated in advance and consequently does not require additional calculations. The Backward phase does not involve any additional calculations.

The DFHMM also applies the factorial Viterbi algorithm to decode the states. However, it is executed in a lower-dimensional space $D \ll d$. It is applied twice, once for FHMM$_a$ and once for FHMM$_b$. By substituting d by D, the total number of operations in the preprocessing stage is therefore reduced to $2 \times S^2 \times (D^2 + D)$ multiplications and additions. The forward stage is independent of the dimensionality, and hence requires $2 \times S^2$ additions. The major computational saving is thus attributed to the lower dimensionality of the embedded space.

Table 3 summarizes the total number of operations for the DFHMM, Roweis [36], Hu and Wang [38] and the MIXMAX. The HFHMM has an identical computational burden as [36]. Algorithms for the two stages previously discussed.

To demonstrate the differences in the computational burden of the algorithm we show the results for the nominal values used for obtaining the results in Section 5: $S = 70$, $L = 3000$, $d = 129$, and $D = 30$. The parameters for [38] also include the beam width $W = 16$ and the iterations number $I = 4$.

With the above parameter settings, the number of multiplications of the DFHMM algorithm is lower by close to two orders of magnitude as compared with Roweis' algorithm [36], and the number of additions is, respectively, lower by an order of magnitude. The number of operations required by [38] is double the number of operations required by the DFHMM.

The MIXMAX is an exception, since it only uses GMMs instead of the HMMs and therefore avoids the computationally expensive factorial search.

The computational burden of the DFHMM can be further reduced by decreasing the value of D. This, however, might result in degraded performance. Thus, there is a tradeoff between performance and computational complexity of the DFHMM.

Extended computational saving might be obtained by adopting a grammar model. As shown in [39], various methods for complexity reduction can be applied. These methods can be adopted by the proposed DFHMM algorithm to reduce the computational complexity without sacrificing performance. However, we preferred to leave these extensions for a later study, and to only focus in this paper on dimensionality reduction.

7 Conclusions

In this work, we presented two novel approaches for estimating temporal FHMM on manifolds based on the diffusion framework, that are non-iterative and rigorously accurate. The core of our approach is to utilize the asymptotics of the Markov random walk, induced on the graph representation of a high-dimensional data source, to decouple the estimation of the latent state space (states and transition probabilities), and the estimation of the emission (observation) p.d.f.s. We applied the proposed schemes to the task of separating two speakers using a single-microphone, that provides a viable baseline to validate the effectiveness of the proposed scheme. In particular, we derived two FHMM-based separation schemes, where the first estimates the HMM of each speaker in the diffusion domain, and then utilizes the log-max approximation to infer the FHMM model. The second, formulates the speech separation problem entirely in the embedded domain, as the derivation of two FHMM models, each adapted to the diffusion embedding of a particular speaker. The inferred states are used to construct masking functions to unmix the speech signal. Two masking schemes are presented, utilizing either soft or hard masks. We experimentally evaluated the proposed schemes using both objective metrics and informal subjective listening tests, for male-female, male-male, and female-female mixtures. The HFHMM scheme is shown to yield comparable and even slightly better performance than [36], while the DFHMM scheme exhibits performance degradation, probably due to sub-optimal embedding (that uses the Nyström extension). The MIXMAX [35] and the iterative algorithm by Hu and Wang [38] had the best SAR and SDR score, although the HFHMM and Roweis methods with soft masks obtained good SAR as well. The proposed HFHMM scheme obtained the best SIR scores on average among all tested algorithms, with insignificant advantage over Hu and Wang method. Informal listening tests demonstrate the insufficiency of the current solution to fully recover the two speakers. Although the separation capabilities of the MIXMAX and Hu and Wang algorithms are slightly better than those of the proposed schemes, the differences are quite marginal, according to our subjective evaluation. Several sound clips are available on our website.

Table 3 Number of operations per output frame for the DFHMM, Roweis, MIXMAX, and the iterative separation by Hu and Wang (with I iterations)

Algorithm	Additions	Multiplications	LUT indexing
DFHMM	$L(d + D + 1) + 2S^2(D^2 + D)$	$L(d + 1) + D(L + 2S^2 + D)$	L
Roweis	$S^2(d^2 + d + 1)$	$S^2(d^2 + d)$	–
MIXMAX	$3dS^2$	$2dS^2$	–
Hu and Wang	$(6d + W + 2)S^2I$	$(3d + W + 1)S^2I + dSI$	$2dSI$

Finally, the proposed training-based methods, and in particular the DFHMM scheme, can be considered as a computationally efficient alternative to the inference of time-series modeled by a large number of states. The performance of the proposed methods is comparable to contemporary methods, and we anticipate that careful examination of the relation between the log-spectral domain and the embedded domain might lead to further improvements.

Endnotes

[1] The authors are grateful to Ke Hu and Deliang Wang for the assistance in applying their algorithm [38] to the reported dataset.

[2] The implementation of the MIXMAX [35] algorithm was done by Yeminy, Keller and Gannot.

[3] See http://www.eng.biu.ac.il/gannot/speech-enhancement/single-microphone-speech-separation-by-diffusion-based-hmm-estimation.

Appendix

Example of learning HMMs with diffusion

In order to demonstrate the learning process of an HMM with diffusion maps, we take a simple HMM with 3 states as an example. The mean vectors of the HMM are $\boldsymbol{\mu}_1 = -10 \cdot \mathbf{1}_7, \boldsymbol{\mu}_2 = 2 \cdot \mathbf{1}_7$ and $\boldsymbol{\mu}_3 = 0 \cdot \mathbf{1}_7$, where $\mathbf{1}_D$ is the $D \times 1$ vector whose all elements are ones. Each state has a diagonal covariance matrix, with $\mathbf{Q}_1 = 0.5 \cdot \mathbf{I}_7, \mathbf{Q}_2 = \mathbf{Q}_3 = 0.1 \cdot \mathbf{I}_7$, where $\mathbf{Q}_i$ is the covariance matrix of the ith state, and $\mathbf{I}_D$ is the identity matrix of dimension D.

The matrix of transition probabilities, $\mathbf{P}$, is

$$\begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.55 & 0.15 \\ 0.35 & 0.25 & 0.4 \end{pmatrix}$$

and the initial probabilities are $p_1 = 0.3, p_2 = 0.2, p_3 = 0.5$. A sequence of 1000 samples of the process is generated, and the diffusion mapping is applied to the data. Since there are 3 states, the kernel bandwidth was chosen so as to have each sample connected to ≈ 20 % of the data. There are states with high probability and low variance and vice-versa, so such a connectivity suits the scenario. A value of $\varepsilon = 1.6$ met the requirement. Only the two most leading eigenvectors and eigenvalues were used, so in the diffusion space each sample has two coordinates.

By applying the K-means algorithm to the embedded samples, the meta-states were found. The transitions probabilities between states were subsequently assessed like detailed above, namely constructing a 3×3 transitions histogram. It resulted in the following estimated transition probabilities matrix:

$$\begin{pmatrix} 0.376 & 0.300 & 0.324 \\ 0.302 & 0.554 & 0.144 \\ 0.356 & 0.248 & 0.396 \end{pmatrix}.$$

which is very close to the original matrix, $\mathbf{P}$. The mean vectors and the variances were also very close to the real values. The estimated initial probabilities are $\hat{p}_1 = 0.343$, $\hat{p}_2 = 0.382$, and $\hat{p}_3 = 0.275$, which are not consistent with the true value. Nevertheless, the initial probabilities are used only for the first frame of each sentence during the inference process. Together with the fact the sentence is composed of tens of frames, it is reasonable to consider their impact negligible.

Abbreviations

BSS: Blind source separation; CASA: Computational auditory scene analysis; DFHMM: Dual FHMM; DFHMM-H: DFHMM with hard mask; DFHMM-S: DFHMM with soft mask; EM: Estimation-maximization; FHMM: Factorial hidden Markov model; FS-HMM: Factorial scaled hidden Markov model; GMM: Gaussian mixture model; HFHMM: Hybrid FHMM; HFHMM-H-E: HFHMM with hard mask and using exact embedding; HFHMM-H-N: HFHMM with hard mask and using Nyström extension; HFHMM-S-E: HFHMM with soft mask and using exact embedding; HFHMM-S-N: HFHMM with soft mask and using Nyström extension; HMM: Hidden Markov model; ICA: Independent component analysis; IdBM: Ideal binary mask; iDFHMM: Ideal DFHMM; JADE: Joint approximate diagonalization of eigen-matrices; k-NN: k nearest neighbors; MAP: Maximum-a-posteriori; MFCC: Mel frequency cepstral coefficients; ML: Maximum likelihood; MMSE: Minimum mean square error; N-HMM: Non-negative hidden Markov model; NMF: Non-negative matrix factorization; p.d.f.: Probability density function; PSD: Power spectral density; RBF: Radial basis function; Roweis-H: Roweis with hard mask; Roweis-S: Roweis with soft mask; SAR: Source to artifacts ratio; SCSS: Single-channel speech separation; SDR: Source to distortion ratio; SIR: Source to interference ratio; SNR: Source to noise ratio; SOBI: Second order blind identification; STFT: Short-time Fourier transform

Competing interests

The authors declare that they have no competing interests.

Endnotes

Signals in endnote *a* are available at http://www.eng.biu.ac.il/gannot/speech-enhancement.

Received: 1 February 2016 Accepted: 4 October 2016
Published online: 18 October 2016

References

1. M Weintraub, *A theory and computational model of auditory monaural sound separation (stream, speech enhancement, selective attention, pitch perception, noise cancellation)*. PhD thesis. (Stanford University, Palo Alto, 1985)
2. TW Parsons, Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. **60**(4), 911–918 (1976)
3. GJ Brown, M Cooke, Computational auditory scene analysis. Comput. Speech Lang. **8**(4), 297–336 (1994)
4. D Wang, GJ Brown, Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Netw. **10**(3), 684–697 (1999)
5. G Hu, D Wang, A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio Speech Lang. Process. **18**(8), 2067–2079 (2010)
6. D Wang, GJ Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. (Wiley-IEEE Press, Hoboken, 2006)
7. E Vincent, MD Plumbley, in *Proceedings of Independent Component Analysis (ICA)*. Single-channel mixture decomposition using Bayesian harmonic models (Springer, Charleston, 2006), pp. 722–730
8. Y Li, D Wang, On the optimality of ideal binary timefrequency masks. Speech Comm. **51**, 230–239 (2009)

9. O Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Sig. Process. **52**(7), 1830–1847 (2004)
10. G-J Jang, T-W Lee, A maximum likelihood approach to single-channel source separation. J. Mach. Learn. Res. **4**, 1365–1392 (2003)
11. G-J Jang, T-W Lee, in *NIPS*. A probabilistic approach to single channel blind signal separation (MIT Press, Vancouver, 2002), pp. 1173–1180
12. M Zibulevsky, BA Pearlmutter, blind source separation by sparse decomposition in a signal dictionary. Neural Comp. **13**(4), 863–882 (2001)
13. AJW van der Kouwe, D Wang, GJ Brown, A comparison of auditory and blind separation techniques for speech segregation. IEEE Trans. Speech Audio Process. **9**(3), 189–195 (2001)
14. JF Cardoso, High-order contrasts for independent component analysis. Neural Comput. **11**, 157–192 (1999)
15. A Belouchrani, K Abed-Meraim, JF Cardoso, E Moulines, A blind source separation technique using second order statistics. IEEE Trans. Signal Process. **45**, 434–444 (1997)
16. T Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**, 1066–1074 (2007)
17. P Smaragdis, Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. **15**(1), 1–12 (2007)
18. C Joder, F Weninger, F Eyben, D Virette, B Schuller, in *Latent Variable Analysis and Signal Separation*, ed. by F Theis, A Cichocki, A Yeredor, and M Zibulevsky. Real-time speech separation by semi-supervised nonnegative matrix factorization. Lecture Notes in Computer Science, vol. 7191 (Springer, Tel-Aviv, 2012), pp. 322–329
19. L Benaroya, L McDonagh, F Bimbot, R Gribonval, in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Nonnegative sparse representation for Wiener based source separation with a single sensor (IEEE, Hong-Kong, 2003), pp. 613–616
20. R Blouet, I Cohen, in *Speech Processing in Modern Communication*, ed. by I Cohen, J Benesty, and S Gannot. Codebook approaches for single sensor speech/music separation (Springer, Berlin, 2009), pp. 183–198
21. P Mowlaee, MG Christensen, SH Jensen, New results on single-channel speech separation using sinusoidal modeling. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1265–1277 (2011)
22. P Mowlaee, R Saeidi, MG Christensen, Z-H Tan, T Kinnunen, P Franti, SH Jensen, A joint approach for single-channel speaker identification and speech separation. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2586–2601 (2012)
23. FR Bach, M Jordan, in *NIPS*. Blind one-microphone speech separation: A spectral learning approach (MIT Press, Vancouver, 2005), pp. 65–72
24. P-S Huang, M Kim, M Hasegawa-Johnson, P Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
25. D Yu, M Kolbæk, Z-H Tan, J Jensen, in *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*. Permutation invariant training of deep models for speaker-independent multi-talker speech separation (IEEE, Shanghai, 2016)
26. JR Hershey, Z Chen, J Le Roux, S Watanabe, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Deep clustering: Discriminative embeddings for segmentation and separation (IEEE, Shanghai, 2016), pp. 31–35
27. X-L Zhang, D Wang, A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(5), 967–977 (2016)
28. B Raj, P Smaragdis, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Latent variable decomposition of spectrograms for single channel speaker separation (IEEE, New York, 2005), pp. 17–20
29. T Kristjansson, H Attias, J Hershey, in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Single microphone source separation using high resolution signal reconstruction, vol. 2 (IEEE, Montreal, 2004), pp. 817–820
30. L Benaroya, F Bimbot, R Gribonval, Audio source separation with a single sensor. IEEE Trans. Audio Speech Lang. Process. **14**(1), 191–199 (2006)
31. A Nádas, D Nahamoo, MA Picheny, Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoust. Speech Sig. Process. **37**(10), 1495–1503 (1989)
32. D Burshtein, S Gannot, Speech enhancement using a mixture-maximum model. IEEE Trans. Speech Audio Process. **10**(6), 341–351 (2002)
33. Y Yeminy, S Gannot, Y Keller, in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Speech enhancement using a multidimensional Mixture-Maximum model, vol. 2 (IWAENC, Tel Aviv, 2010)
34. AM Reddy, B Raj, Soft mask methods for single-channel speaker separation. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1766–1776 (2007)
35. MH Radfar, RM Dansereau, Single-channel speech separation using soft mask filtering. IEEE Trans. Audio Speech Lang. Process. **15**(8), 2299–2310 (2007)
36. ST Roweis, One microphone source separation. Adv. neural Inf. Process. Syst. **13**, 793–799 (2001)
37. MH Radfar, W Wong, RM Dansereau, W-Y Chan, in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Scaled factorial hidden Markov models: A new technique for compensating gain differences in model-based single channel speech separation (IEEE, Dallas, 2010), pp. 1918–1921
38. K Hu, D Wang, An iterative model-based approach to cochannel speech separation. EURASIP J. Audio Speech Music Process. **2013**(1), 1–11 (2013)
39. JR Hershey, SJ Rennie, PA Olsen, TT Kristjansson, Super-human multi-talker speech recognition: a graphical modeling approach. Elsevier Comput. Speech Lang. **24**(1), 45–66 (2010)
40. RJ Weiss, DPW Ellis, Speech separation using speaker-adapted eigenvoice speech models. Elsevier Comput. Speech Lang. **24**(1), 16–29 (2010)
41. J Ming, R Srinivasan, D Crookes, A Jafari, Close–a data-driven approach to speech separation. IEEE Trans. Audio Speech Lang. Process. **21**(7), 1355–1368 (2013)
42. GJ Mysore, P Smaragdis, B Raj, in *Latent Variable Analysis and Signal Separation - 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010. Proceedings*. Non-negative hidden Markov modeling of audio with application to source separation (Springer, St. Malo, 2010), pp. 140–148
43. A Ozerov, C Févotte, M Charbit, in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Factorial scaled hidden Markov model for polyphonic audio representation and source separation (IEEE, New York, 2009), pp. 121–124
44. S Russell, P Norvig, *Artificial Intelligence: a Modern Approach*, 3rd edn. (Prentice Hall Press, Upper Saddle River, 2009)
45. S Rennie, P Olsen, J Hershey, T Kristjansson, in *Workshop on Statistical and Perceptual Audio Processing (SAPA)*. The iroquois model: using temporal dynamics to separate speakers (ISCA, Pittsburgh, 2006)
46. SJ Rennie, JR Hershey, PA Olsen, in *IEEE Workshop on Automatic Speech Recognition and Understanding*. Hierarchical variational loopy belief propagation for multi-talker speech recognition (IEEE, Merano, 2009), pp. 176–181
47. M Wohlmayr, M Stark, F Pernkopf, A probabilistic interaction model for multipitch tracking with factorial hidden Markov models. IEEE Trans. Audio Speech Lang. Process. **19**, 799–810 (2011)
48. MJ Reyes-gomez, DPW Ellis, N Jojic, in *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*. Multiband audio modeling for single-channel acoustic source separation (IEEE, Montreal, 2004)
49. Y Michalevsky, R Talmon, I Cohen, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Speaker identification using diffusion maps (IEEE, Barcelona, 2011), pp. 4029–4032
50. RR Coifman, S Lafon, Diffusion maps. Appl. Comput. Harmon. Anal: Spec. Iss. Diffus. Maps Wavelets. **22**, 5–30 (2006)
51. WH Press, BP Flannery, SA Teukolsky, WT Vetterling, *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*, 2nd edn. (Cambridge University Press, Cambridge, 1992)
52. JR Hershey, SJ Rennie, J Le Roux, *Techniques for Noise Robustness in Automatic Speech Recognition*. (Wiley, Chichester, 2012). Chap. 12
53. Y Keller, RR Coifman, S Lafon, SW Zucker, Audio-visual group recognition using diffusion maps. IEEE Trans. Signal Process. **58**(1), 403–413 (2010)
54. S Lafon, Y Keller, RR Coifman, Data fusion and multicue data matching by diffusion maps. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1784–1797 (2006)
55. B Nadler, S Lafon, RR Coifman, IG Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Appl. Comput. Harmon. Anal. **21**(1), 113–127 (2006)

56. B Nadler, S Lafon, RR Coifman, IG Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. Adv. Neural Inf. Process. Syst. **18**, 955–962 (2005)
57. S Lafon, AB Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning and data set parameterization. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1393–1403 (2006)
58. M Cooke, TW Lee, The speech separation challenge. (2006). http://laslab.org/SpeechSeparationChallenge/. Accessed 2012
59. Y Keller, Y Gur, A diffusion approach to network localization. IEEE Trans. Signal Process. **59**(6), 2642–2654 (2011). doi:10.1109/TSP.2011.2122261
60. E Vincent, R Gribonval, C Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
61. C Févotte, R Gribonval, E Vincent, *BSS EVAL Toolbox User Guide*, (Rennes, 2005). http://bass-db.gforge.inria.fr/bss_eval/. Accessed 2012