Integration of neural networks and probabilistic spatial models for acoustic blind source separation

Lukas Drude, *Student Member, IEEE,* and Reinhold Haeb-Umbach, *Senior Member, IEEE*

***Abstract*—We formulate a generic framework for blind source separation (BSS), which allows integrating data-driven spectro-temporal methods, such as deep clustering and deep attractor networks, with physically motivated probabilistic spatial methods, such as complex angular central Gaussian mixture models. The integrated model exploits the complementary strengths of the two approaches to BSS: the strong modeling power of neural networks, which, however, is based on supervised learning, and the ease of unsupervised learning of the spatial mixture models whose few parameters can be estimated on as little as a single segment of a real mixture of speech. Experiments are carried out on both artificially mixed speech as well as true recordings of speech mixtures. The experiments verify that the integrated models consistently outperform the individual components. We further extend the models to cope with noisy, reverberant speech and introduce a cross-domain teacher-student training where the mixture model serves as the teacher to provide training targets for the student neural network.**

***Index Terms*—blind source separation, speech processing, beamforming, deep clustering, neural networks, teacher-student.**

I. Introduction

Acoustic blind source separation (BSS) deals with algorithmic solutions to extract the speech of each concurrent speaker from an audio recording. The problem at hand is often coined the *Cocktail Party Problem* [1] envisioning people to discuss simultaneously in a fairly uncontrolled setting in contrast to, e.g., telephone speech where close-talk, low noise and no cross-talk conditions are common. In recent years, a number of neural network-based blind source separation systems emerged. This raises the question whether statistical model-based clustering has still its justification in modern systems. Therefore, we first revisit data-driven single-channel source separation algorithms and then review statistical model-based multi-channel approaches. Finally, we guide to integration variants, which allow to use both modalities, permit to incorporate prior knowledge, enable unsupervised training and possibly provide more insight than the fearfully named *black box* neural networks.

Over the years many conceptually quite different algorithmic approaches to BSS emerged, which either focus on single-channel observations mainly leveraging spectral cues such as pitch and common onset times or – on the contrary – are designed for multi-channel observations mainly leveraging spatial cues such as phase and level differences between the microphones but often neglecting temporal and spectral correlation altogether.

Shallow blind decomposition techniques for single channel separation which do not use any kind of deep neural network (DNN) have only led to limited success. Although Computational Auditory Scene Analysis can separate speech to some degree, it relies on complicated hand-tuned grouping rules (e.g., harmonicity or common onsets) which do not involve any automated learning [2]. Non-negative matrix factorization (NMF) is a well-studied approach to separate e.g. single channel mixtures by using previously learned signal-dependent dictionaries [3]–[5]. A quite different approach is to use factorial hidden Markov models (HMMs) [6] such that the temporal structure of each source is modeled with a speaker-dependent HMM which led to promising results on a fairly narrow automatic speech recognition (ASR) task [7]. An overview of pre-DNN single-channel BSS with probabilistic models can be found in [8].

Early deep neural network-based approaches demonstrated much better separation performance but still relied on speaker-dependent networks [9]–[12]. In contrast, deep clustering (DC) broke with all these drawbacks and turned out to be a great step forward towards single-channel speech separation [13], [14]: A neural network is trained to learn embeddings from the time-frequency representation of the signal, such that embeddings belonging to the same source form clusters. This latent structure can then be used to obtain masks by using, e.g., k-means clustering. An attractive property of DC is the fact, that the network is not fixed to a predefined number of speakers. In fact, the network can be trained with two speakers and evaluated on three speakers [13]. An interesting alternative to avoid the speaker counting issue is to train a neural network to output masks one by one [15]. Deep attractor networks (DANs) are a notable variant of DC in the sense that they allow to train with a signal reconstruction criterion while still creating a latent representation for clustering [16]. Source Contrastive Estimation modified the training recipe such that the clusters in the latent space tend to be more compact [17]. Permutation invariant training (PIT) is an alternative to DC and has proven to be a successful tool to train a neural network to separate a predefined number of target speakers [18]–[20]. In contrast to DC it does not require an additional clustering step. In the spirit of multi-task learning, it was demonstrated that a PIT network can achieve remarkably better separation performance when trained with an additional DC loss function [21].

While single-channel source separation relies on spectro-temporal properties of the speech signal, multi-channel statistical model-based solutions exploit the spatial diversity of the sources. Traditionally, multi-channel blind speech separation is either tackled by independent component analysis (ICA) [22], [23] or with statistical model-based clustering. In particular,

L. Drude is with the Communications Engineering Group, Paderborn University, Paderborn, 33098 Germany e-mail: mail@lukas-drude.de.

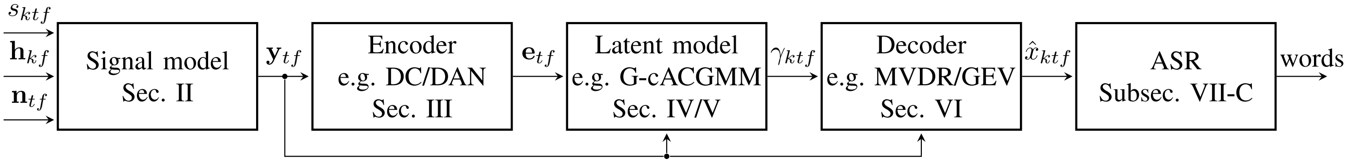

Fig. 1. Block diagram of the entire framework including signal model and speech recognition back-end. The observed signal $\mathbf{y}_{tf}$ or at least one channel of it is available to the encoder (a module which encodes the input into a latent representation), the latent model and the decoder (a module which uses the latent representation to extract the sources). Depending on the model choice, either the encoder can be omitted or the latent model does not make use of $\mathbf{y}_{tf}$.

spatial model-based approaches that exploit the sparseness of speech in the short time Fourier transform (STFT) domain have become very popular [24]–[31]. The majority of these techniques treats each time-frequency bin as statistically independent and neglects frequency dependencies. Carrying out separation on each frequency separately leads to the frequency permutation problem: Even if the source separation were perfect for each frequency bin, it is likely, that component one of a given frequency bin does not correspond to the same speaker as component one of another frequency bin [32]. Notable exceptions either apply a frequency normalization [33] or estimate statistics which are shared across frequencies [34]. An extensive overview of multi-channel speech enhancement and separation can be found in [35].

Although the spectral, as well as the spatial features, are very informative for speech enhancement and separation tasks, the number of systems integrating both modalities is limited. Woodruff et al. integrate both modalities using CASA and binaural clustering [36]. Nakatani et al. proposed DOLPHIN, which integrates factorial spectral models with spatial clustering [37]. Tran Vu et al. use 2D-HMMs to model temporal and spectral dependencies while using a spatial observation model to capture spatial cues [38]. In [10] spectral features are modeled by an NMF, while spatial features are modeled by a full rank covariance model. More recently, [39] proposed the integration of a DNN-based mask estimator and a complex angular central Gaussian mixture model (cACGMM) to extract a single source. In [40] a DNN refines the source estimate in each expectation maximization (EM) iteration. In [41] we proposed modeling spectral features with a DC model and spatial features with a time-variant complex Gaussian mixture model (TV-cGMM). In [42] we presented a DAN+TV-cGMM.

Here, we extend prior work along several different dimensions. First, we formulate the integration framework in a generic sense as an encoder decoder structure. Second, we introduce the von-Mises-Fisher complex angular central Gaussian mixture model (vMF-cACGMM) and the Gaussian complex angular central Gaussian mixture model (G-cACGMM). Third, we expand it to noisy reverberant environments by introducing an additional noise class both for the spectro-temporal encoder as well as the probabilistic integration model. Forth, we theoretically justify, why an integration weight [41] is now obsolete. Fifth, we evaluate DC, DAN, and the integration models for the first time on real recordings. Finally, in contrast to [41] we employ a state of the art acoustic model (AM) to ensure that gains in the front-end are not eaten up by a strong back-end.

This work is organized as follows, where Fig. 1 can be seen as a visual table of contents: The signal model, as well as the assumptions underlying this work, are explained in Sec. II. Sec. III introduces DC and DANs as particular examples of neural network-based source separation techniques. Sec. III also introduces the concept of encoder, latent model, and decoder which will help to formalize the different approaches. Sec. IV revisits probabilistic spatial models and the corresponding solution in the form of update equations of an EM algorithm. Sec. V introduces the integration framework by generalizing the probabilistic spatial mixture models to handle an observation model for each cue. Sec. VI briefly reviews source extraction methods relevant for this work while Sec. VII consists of a thorough evaluation of different aspects of the proposed framework. Most notably, we evaluate unsupervised training of DC, multi-channel features for the neural network encoder and BSS, as well as ASR experiments on real recordings.

II. Signal model and objective

A convolutive mixture in time domain captured by D sensors is approximated by an instantaneous mixture in the STFT domain, where s_{ktf} represents K independent source signals:

$$\begin{aligned}\mathbf{y}_{tf} &= \sum_k \mathbf{h}_{kf}\, s_{ktf} + \mathbf{n}_{tf} \\ &= \sum_k \mathbf{x}_{ktf} + \mathbf{n}_{tf},\end{aligned} \tag{1}$$

where $\mathbf{y}_{tf}$, $\mathbf{h}_{kf}$, $\mathbf{n}_{tf}$, and $\mathbf{x}_{ktf}$ are the D-dimensional complex-valued observed signal vector, the complex-valued unknown acoustic transfer function vector of source $k \in \{1, \dots K\}$, the complex-valued noise vector, and the complex-valued source images at the sensors, respectively. Furthermore, $t \in \{1, \dots T\}$ and $f \in \{1, \dots F\}$ specify the time frame index and the frequency bin index, respectively. This narrowband approximation ignores inter-frame and inter-band convolution effects [43]. Consequently, it is assumed that the impulse response is short enough to approximately fit in a single frame. Since speech signals are sparse in the STFT domain [25], [44], we may assume that a time frequency slot is dominated by a single source or occupied by noise only; i.e., we assume that the sources are sufficiently disjoint in the STFT domain [25].

The goal of all methods presented subsequently is to obtain an estimate $\hat{\mathbf{x}}_{ktf}$ for the speech image $\mathbf{x}_{ktf}$, e.g., at a particular reference microphone. Consequently, the focus is not on dereverberation and not on obtaining and estimate for the source signal s_{ktf}.

III. Neural network-based source separation

In this section, we review neural network-based source separation. In contrast to speaker-dependent source separation neural networks, e.g., [12], we will here focus on more recent developments, which provide methods to separate speakers, which were never seen during training. Namely, we will introduce DC and DANs and relate them to the vocabulary commonly used in variational autoencoders [45].

A. Deep clustering

DC is a technique which aims to blindly separate unseen speakers in a single-channel mixture. The training procedure described in the original work [13], [14] assumes, that ideal binary masks for each speaker are available to train a multi-layer bidirectional long short term memory network (BLSTM) [46] to map from $T \cdot F$ spectral features (e.g., log-spectral amplitude) to the same number of E-dimensional embedding vectors $\mathbf{e}_{tf}$, where $\|\mathbf{e}_{tf}\| = 1$. This network can be seen as an encoder such that the embedding vectors are some kind of latent code. The objective during training is to minimize the Frobenius norm of the difference between the estimated and true affinity matrix:

$$\ell = \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\mathrm{F}}^2 = \left\|\mathbf{E}\mathbf{E}^\mathsf{T} - \mathbf{C}\mathbf{C}^\mathsf{T}\right\|_{\mathrm{F}}^2, \tag{2}$$

where $\hat{\mathbf{A}}$ and $\mathbf{A}$ are the estimated and ground truth affinity matrices (for a discussion of improved loss functions see [47]). The entries $A_{n,n'}$ encode, whether observation n and n' belong to the same source ($A_{n,n'} = 1$, and zero else; n indexes $T \cdot F$ rows/ columns). Correspondingly, the embeddings are stacked in a single matrix $\mathbf{E}$ with shape $(TF \times E)$ and the ground truth one-hot vectors describing which time frequency slot belongs to which source are stacked in a single matrix $\mathbf{C}$ with shape $(TF \times K)$, such that $C_{nk} = 1$, if observation n belongs to source k and $C_{nk} = 0$ otherwise. During training, the loss as defined in Eq. 2 encourages the network to move embeddings belonging to the same source closer together while pushing embeddings which belong to different sources further apart.

After training, the embeddings, which are normalized to unit-length, can be clustered to obtain time frequency masks for each source. This can be related to the latent probabilistic models in structured variational autoencoders [48]. The original work on DC used k-means clustering which then yields masks for a subsequent source extraction scheme, e.g., masking (compare Fig. 2). To again relate it to variational autoencoders, since this source extraction uses the latent structure to predict signals in the domain of the observation, it can be seen as some kind of decoder structure. However, binary masks often lead to musical tones harming ASR performance which can be reduced using an additional DNN as a decoder neural network [14].

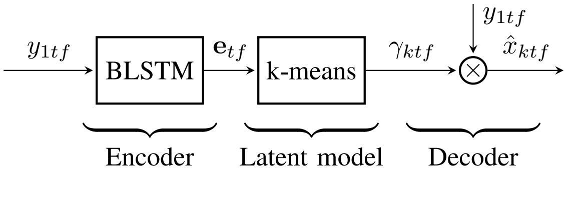

Fig. 2. Schematic view of a DC system with a masking decoder.

TABLE I

Viewing DC and related methods as a structure with an encoder, a latent model and a decoder. All entries represent the configuration at test time.

Reference	Encoder	Latent model	Decoder
[13], [17], [49]	BLSTM	k-means	Masking
[14]	BLSTM	weighted k-means	BLSTM + Masking
[16], [50]	BLSTM	k-means	DAN-Masking

B. Deep attractor networks

DANs [16] consist of an encoder just as the DC system. However, during training, attractors $\boldsymbol{\mu}_k$ are calculated similarly to the M-step of an EM algorithm for Gaussian mixture models (GMMs) by using the target mask $M_{ktf}^{(\text{oracle})}$ as supervision in a weighted mean:

$$\boldsymbol{\mu}_k = \sum_{tf} M_{ktf}^{(\text{oracle})} \mathbf{e}_{tf} \Bigg/ \sum_{tf} M_{ktf}^{(\text{oracle})} . \tag{3}$$

These can then be used to estimate a soft mask as an inner product for signal reconstruction (DAN-Masking in Tbl. I):

$$\gamma_{ktf} = \operatorname*{softmax}_k \left(\boldsymbol{\mu}_k^\mathsf{T} \mathbf{e}_{tf}\right) = \mathrm{e}^{\boldsymbol{\mu}_k^\mathsf{T} \mathbf{e}_{tf}} \Bigg/ \sum_{k'=1}^{K} \mathrm{e}^{\boldsymbol{\mu}_{k'}^\mathsf{T} \mathbf{e}_{tf}} . \tag{4}$$

First of all, this allows defining a loss function which includes the mask and therefore avoids a surrogate loss function such as in Eq. 2 and is faster to evaluate. In particular, a reconstruction loss has proven to be successful when the application later uses a DAN-Masking decoder for signal reconstruction:

$$\ell_{\text{MSE}} = \operatorname*{MSE}_{ktf}\left(\hat{x}_{ktf}, x_{ktf}\right), \qquad \hat{x}_{ktf} = \gamma_{ktf} \cdot y_{1tf}, \tag{5}$$

where $\mathrm{MSE}()$ is the mean squared difference of the arguments.

C. Extension with additional noise class

Most DC and DAN related papers so far do neither test in noisy conditions nor do the models specifically account for a noise class. For any more realistic scenario, background noise will always be present and it is worth addressing it already during training.

Four possible ways to train the network are then:

1) Treat the noise mask just like a speaker mask. The network is then forced to produce $K + 1$ clusters at arbitrary locations.
2) Provide a fixed attractor for the noise class. The speaker clusters can move anywhere. The noise attractor is already known at test time, so this avoids the problem to confuse noise with any of the speakers.
3) Since it is unclear, how to set the fixed attractor, the attractor can also be a trainable network parameter.
4) Alternatively, the network can be trained to output an additional noise presence probability mask.

Preliminary experiments showed that treating the noise mask just like an additional speaker works sufficiently well. Therefore, we leave a detailed investigation of the other options for future research.

IV. Probabilistic spatial mixture models

This section focuses on probabilistic models to capture spatial characteristics of multi-channel observations in the STFT domain. Based on the assumption that speech is a sufficiently sparse signal in the STFT domain [25], [44] one can model the observations with a mixture model.

In its generic form, the distribution of the multi-channel observations can be formulated as a marginalization over all class labels with the assumption that all observations are conditionally i.i.d.:

$$p(\mathbf{y}_{tf}) = \sum_k \pi_{kf} p(\mathbf{y}_{tf}|\boldsymbol{\theta}_k), \tag{6}$$

where π_{kf} is the a-priori probability, that an observation belongs to mixture component k, and $p(\mathbf{y}_{tf}|\boldsymbol{\theta}_k)$ is any appropriate class conditional distribution which can model $\mathbf{y}_{tf}$ and $\boldsymbol{\theta}_k$ captures all class-dependent parameters. Independent of the particular choice of the mixture weight and $p(\mathbf{y}_{tf}|\boldsymbol{\theta}_k)$, the class affiliation posterior is obtained as follows:

$$\gamma_{ktf} = P(c_{ktf}{=}1|\mathbf{y}_{tf}) = \frac{\pi_{kf} p(\mathbf{y}_{tf}|\boldsymbol{\theta}_k)}{\sum_{k'} \pi_{k'f} p(\mathbf{y}_{tf}|\boldsymbol{\theta}_{k'})}, \tag{7}$$

where c_{ktf} represents the class affiliation of $\mathbf{y}_{tf}$.

One instance of this generic mixture model is a complex Watson mixture model (cWMM) [28], [51], where the class conditional distribution is a complex Watson distribution [52]. A full-band cWMM is analyzed in [53] and a variational inference approach to cWMMs with complex Bingham priors is proposed in [54]. The complex Bingham distribution [55] can of course also be used as a class conditional distribution yielding the complex Bingham mixture model [56]. Another alternative is a complex Gaussian mixture model [29], of which a variational approach is presented in [57].

All spatial clustering approaches mentioned in this section do not rely on any pre-trained values or speaker-dependent codebooks. Each parameter can be estimated and each latent variable inferred with an EM algorithm on a single mixture. Therefore, these models may serve as a teacher for neural network-based separation systems as discussed in Subsec. VII-F.

In the remainder of this section we will focus on a specific example of a spatial mixture model and further introduce issues related to initialization and frequency permutation often encountered in STFT domain spatial clustering. Finally, we mention specific model choices and highlight guided source separation.

A. Complex Angular central Gaussian mixture model

The cACGMM [58] uses a complex Angular central Gaussian distribution [59] as a class conditional distribution:

$$p(\tilde{\mathbf{y}}_{tf}|\mathbf{B}_{kf}) = \frac{(D-1)!}{2\pi^D \det \mathbf{B}_{kf}} \frac{1}{(\tilde{\mathbf{y}}_{tf}^{\mathsf{H}} \mathbf{B}_{kf}^{-1} \tilde{\mathbf{y}}_{tf})^D}, \tag{8}$$

where $\tilde{\mathbf{y}}_{tf} = \mathbf{y}_{tf}/\|\mathbf{y}_{tf}\|$. Due to this normalization, the model can only capture intra-channel level differences but does not account for the power of an observation. Additionally, it is worth noting, that $\tilde{\mathbf{y}}_{tf}^{\mathsf{H}} \mathbf{B}_{kf}^{-1} \tilde{\mathbf{y}}_{tf}$ is invariant of the absolute phase, thus $p(\tilde{\mathbf{y}}_{tf}) = p(\tilde{\mathbf{y}}_{tf} \mathrm{e}^{\mathrm{j}\phi})$. Therefore, the model only captures intra-channel phase differences, but not the absolute phase.

The parameters and therefore the distribution of the random variables can be estimated using maximum likelihood. This leads to an iterative solution, where the class affiliation posterior γ_{ktf} is updated during the E-step as in Eq. 7 and all remaining parameters are updated during the M-step. It is worth noting, that the update equations of the cACGMM coincide with the update equations of the TV-cGMM [58] although the probabilistic model (Fig. 3) is actually simpler.

We decided to use a cACGMM since it performed better than all other tested mixture models in previous informal experiments on the datasets used in Sec VII.

B. Frequency permutation problem

The aforementioned spatial mixture models neglect frequency dependencies. Thus, when used without any kind of guidance, it will yield a solution where the speaker index is inconsistent over frequency bins. This issue is the so called frequency permutation problem [32]. It can be addressed by calculating that permutation alignment (PA) (bin by bin) which maximizes the correlation of the masks along neighboring frequencies [32].

C. Initialization and influence of the mixture weight

Probabilistic spatial mixture models tend to be very susceptible to initialization. First of all, initializing the class affiliation posteriors γ_{ktf} i.i.d. is suboptimal, since this will result in almost equal class-dependent model parameters after the first M-step and will thus lead to slow convergence. A better initialization is to randomly assign a few consecutive frames exclusively to each of the classes, which alleviates the frequency permutation problem to some degree and encourages that the class-dependent parameters are initially more spread out. A more elaborate initialization scheme is presented in [60].

It is a common choice to use frequency-dependent mixture weights π_{kf} in Eq. 7. However, other alternatives are also possible: a constant mixture weight $1/K$ tends to lead to clusters more evenly populated; a frequency-independent but time-dependent mixture weight π_{kt} is more rarely seen but has the nice property, that it alleviates the permutation problem to some degree and better represents noise-only segments [61].

D. Guided source separation

In case external information is available (see [62] for an overview of guided source separation), probabilistic spatial mixture models provide an easy and intuitive way to integrate this information either as a prior or by fixing part of the parameters during an update, e.g., a guided cACGMM [63] uses this concept to fix the possible values of γ_{ktf} to incorporate external timing annotations for the CHiME 5 challenge [64].

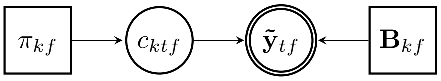

Fig. 3. Probabilistic dependencies in a cACGMM. Circles depict random variables, where doubly circled elements are observable random variables. Boxes are model parameters which are estimated during test time. Arrows indicate statistical dependencies, e.g., c_{ktf} is categorically distributed with the mixture weights $\pi_{1f}, \ldots \pi_{Kf}$.

V. Integration models

An obvious drawback of the probabilistic spatial mixture models described in Sec. IV is, that they do neither account for temporal or spectral relations between the observations nor do they include any information about the energy in a time-frequency slot. Likewise, the single-channel approaches in Sec. III do not access cross-channel information at all, although phase differences and level differences have proven to be important features.

Therefore, an interesting way to make use of both modalities is to use an encoder network (trained according to, e.g., the DC or DAN recipe) to produce spectral features which can then be used in a statistical model which jointly characterizes the distribution of both modalities. The generic formulation of such an integrated probabilistic model assuming conditional independence of both features is given as follows [41]:

$$\begin{aligned} p(\mathbf{y}_{tf}) &= p(\mathbf{e}_{tf}, \tilde{\mathbf{y}}_{tf}) \\ &= \sum_k \pi_{kt} p(\mathbf{e}_{tf} | \boldsymbol{\theta}_k^{(\text{spectral})}) p(\tilde{\mathbf{y}}_{tf} | \boldsymbol{\theta}_k^{(\text{spatial})}), \end{aligned} \tag{9}$$

where, similar to Eq. 6, the parameter π_{kt} is a time-dependent mixture weight while $\boldsymbol{\theta}_k^{(\text{spectral})}$ and $\boldsymbol{\theta}_k^{(\text{spatial})}$ capture all class-dependent spectral and spatial parameters.

It is now possible to derive an EM algorithm which jointly estimates all unknown parameters and allows to infer the a posteriori distribution of the latent class labels:

$$\gamma_{ktf} = \frac{\pi_{kt} p(\mathbf{e}_{tf} | \boldsymbol{\theta}_k^{(\text{spectral})}) p(\tilde{\mathbf{y}}_{tf} | \boldsymbol{\theta}_k^{(\text{spatial})})}{\sum_{k'} \pi_{k't} p(\mathbf{e}_{tf} | \boldsymbol{\theta}_{k'}^{(\text{spectral})}) p(\tilde{\mathbf{y}}_{tf} | \boldsymbol{\theta}_{k'}^{(\text{spatial})})}. \tag{10}$$

Fig. 4 contains an algorithm, outlining the entire separation process. In the remaining part, we now describe two particular instances of this generic model formulation in Eq. 9 and address issues related to the choice of fixed parameters.

A. vMF complex angular central Gaussian mixture model

The vMF-cACGMM jointly models the embedding vectors $\mathbf{e}_{tf}$ of a DC encoder and the normalized spatial observations $\tilde{\mathbf{y}}_{tf} = \mathbf{y}_{tf} / \|\mathbf{y}_{tf}\|$. Since the DC embeddings are unit normalized, a distribution defined on the surface of a unit-sphere is a suitable choice. Therefore, the vMF-cACGMM makes use of a von-Mises-Fisher (vMF) distribution [65] as a spectral observation model $p(\mathbf{e}_{tf} | \boldsymbol{\theta}_k^{(\text{spectral})})$ with the normalization term here represented by $c_{\text{vMF}}(\kappa_k)$:

$$p(\mathbf{e}_{tf} | \boldsymbol{\mu}_k, \kappa_k) = \frac{1}{c_{\text{vMF}}(\kappa_k)} \mathrm{e}^{\kappa_k \boldsymbol{\mu}_k^{\mathsf{T}} \mathbf{e}_{tf}}, \tag{11}$$

while the spatial observation model is a cACGMM. All statistical dependencies are visualized in Fig. 5.

It is worth contrasting the spectral observation model to the proposed approach in [13]. In [13] a k-means is used for clustering which can be seen as a GMM with shared scaled identity covariance matrices and a binary decision in the E-step. Here, we avoid this binary decision and use a vMF distribution. However, when compared in [41], the choice of the exact spectral observation model was not crucial.

```
Calculate DC/DAN embeddings $\mathbf{e}_{tf}$.
Initialize affiliations $\gamma_{ktf}$ with k-means clustering on $\mathbf{e}_{tf}$.
while not converged do
    M-step: Update class conditional parameters.
    E-step: Obtain masks $\gamma_{ktf}$ with Eq. 10.
end while
Run source extraction (Sec. VI).
```

Fig. 4. Source separation algorithm for the generic integration framework.

B. Gaussian complex angular central Gaussian mixture model

The embeddings calculated by a DAN encoder are not normalized to unit norm. Consequently, the vMF distribution is inappropriate as an observation model for the embedding vectors. Therefore, we resort to a Gaussian distribution as a spectral observation model, i.e. replacing $p(\mathbf{e}_{tf} | \boldsymbol{\theta}_k^{(\text{spectral})})$:

$$p(\mathbf{e}_{tf} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}} \mathrm{e}^{-\frac{1}{2}(\mathbf{e}_{tf} - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{tf} - \boldsymbol{\mu}_k)}. \tag{12}$$

Again, without restricting the framework in general, a cACGMM can be used as a spatial observation model.

C. Estimated vs. fixed parameters

In the vMF-cACGMM and the G-cACGMM all parameters can be estimated on the current mixture. An alternative is to fix, e.g., the concentration parameter $\kappa_k := \kappa^{(\text{fix})}$ or even to constrain the covariance matrix of the spectral model of a G-cACGMM to a scaled identity matrix $\boldsymbol{\Sigma}_k := \sigma^{(\text{fix})} \mathbf{I}_E$, where E is number of embedding dimensions. Using a fixed parameter has the advantage, to choose the parameter such that an additional weighting factor as used in [41] (similar to a language model weight in acoustic modeling) is not necessary anymore. The spectral weight can be factored into, e.g., the concentration parameter:

$$\left(p(\mathbf{e}_{tf} | \boldsymbol{\mu}_k, \kappa_k)\right)^{\alpha} \propto \mathrm{e}^{\alpha \kappa_k \boldsymbol{\mu}_k^{\mathsf{T}} \mathbf{e}_{tf}} = \mathrm{e}^{\kappa_k' \boldsymbol{\mu}_k^{\mathsf{T}} \mathbf{e}_{tf}} \text{ with } \kappa_k' = \alpha \kappa_k.$$

This parameter can then be obtained on a separate development set (see Subsec. VII-D for details). Additionally, choosing a scaled identity matrix $\sigma^{(\text{fix})} \mathbf{I}_E$ instead of a full covariance model in a G-cACGMM has the advantage, that it is a bit closer to the training conditions in which all embedding dimensions were treated equally.

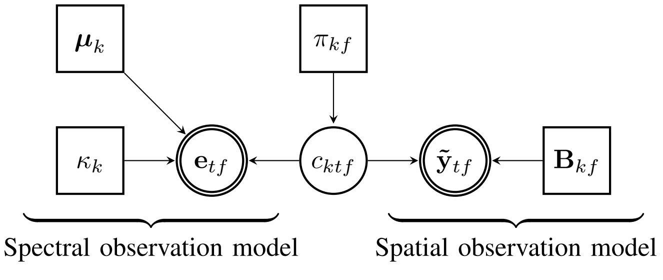

Fig. 5. Probabilistic dependencies in a vMF-cACGMM. Circles depict random variables, where doubly circled elements are observable random variables. Boxes are model parameters which are estimated during test time. Arrows indicate statistical dependencies, e.g., c_{ktf} is categorically distributed with the mixture weights $\pi_{1f}, \dots \pi_{Kf}$.

VI. Source extraction

This section details, how the clustering result of any of the aforementioned methods can be used to actually extract the sources. To remain in the variational autoencoder vocabulary, this processing step corresponds to the decoder.

One obvious choice is to use the class affiliation posterior γ_{ktf} (i.e. mask) to directly mask the observation:

$$\hat{x}_{ktf} = \gamma_{ktf} \cdot y_{1tf}, \tag{13}$$

where y_{1tf} is either an arbitrary reference channel or, in case of single-channel source separation, the only available channel. This is a common choice especially for single-channel approaches (see Tbl. I). When the masks are trained with some kind of reconstruction loss, it can lead to great interference suppression (see, e.g., [66] for a comparison). However, this may lead to musical tones, e.g. when the masks obtained with k-means are used directly.

An alternative is to use mask-based beamforming. This approach is common practice when dealing with multi-channel mixture models and has more recently become an even more competitive approach for multi-channel noise reduction when the masks are estimated with a neural network [67], [68].

For beamforming, we first calculate covariance matrices for each target speaker using the posterior masks γ_{ktf}:

$$\mathbf{\Phi}_{kf}^{(\text{target})} = \sum_t \gamma_{ktf} \mathbf{y}_{tf} \mathbf{y}_{tf}^{\mathsf{H}} \bigg/ \sum_t \gamma_{ktf} \ . \tag{14}$$

Similarly, the covariance matrix of all interferences and noise is calculated with $(1 - \gamma_{ktf})$ instead of γ_{ktf}. Once the beamforming vector $\mathbf{w}_{kf}$ is obtained according to, e.g., Subsec. VI-A below, a linear projection of the observation STFT vector $\mathbf{y}_{tf}$ yields the estimate:

$$\hat{x}_{ktf} = \mathbf{w}_{kf}^{\mathsf{H}} \mathbf{y}_{tf}. \tag{15}$$

A combination of beamforming with a subsequent masking step as a single-channel postfilter is rather obvious, but will not be addressed further in this work.

In the remainder of this section we briefly introduce the statistical beamformers which found application in this work.

A. Minimum variance distortionless response beamformer

In its original form, the minimum variance distortionless response (MVDR) beamformer is designed to minimize the expected output variance while avoiding distortions given a fixed steering vector [69], [70]. Since we intend to use masks from the previous processing step, a formulation which does not require explicit knowledge of the steering vector is preferred. We therefore use the formulation by Souden et al. [71, Eq. 24] with a blind approach to estimate a reference channel [68].

B. Generalized eigenvalue beamformer/ MaxSNR beamformer

A Generalized eigenvalue (GEV) beamformer [72] or also called maximum SNR (MaxSNR) beamformer [73] maximizes the expected signal to noise ratio (SNR) gain for a given target k at the beamformer output. Since GEV beamforming is known to introduce some frequency-dependent distortions, a blind analytic normalization (BAN) can optionally be used [72].

VII. Experimental evaluation

To thoroughly evaluate the proposed integration framework, we evaluate on artificial mixtures as well as on real mixture recordings. To get a good impression of the system performance, we present results in terms of BSS-Eval signal to distortion ratio (SDR) gain [74], invasively calculated SDR gain [28], PESQ gain [75], STOI gain [76] and finally word error rates (WERs). We report invasively calculated SDR gains, since these do not rely on any projection method and cannot be fooled by scaling effects.

This section first introduces both databases and then focuses on a detailed analysis of the integration approach on the simulated databases. We subsequently highlight, how a spatial mixture model can be a sufficient supervision for a DC network. Finally, we compare the integration methods with multi-channel DC on the simulated database and on real recordings.

A. Database design

For the simulated database, we artificially generated 30 000, 500 and 1500 six-channel mixtures with a sampling rate of 8 kHz with source signals obtained from three non-overlapping Wall Street Journal (WSJ) sets (train: `si284`, develop: `dev93`, test: `eval92`) [77], [78]. We padded or cut the second speaker to match the length of the first speaker. Room impulse responses were generated with the Image Method [79], where the room dimensions, the position of the circular array with radius 10 cm and the position of two concurring speakers were randomly sampled. The minimum angular distance was set to 15°. The reverberation time (T60) was uniformly sampled between 200 and 500 ms. White Gaussian noise with 20 to 30 dB SNR was added to the mixture. We here deviated from the file lists provided by [13] since the speakers of their training set and development set overlap and although their training set consists of 20 000 mixtures it only includes 6842 unique utterances from `si84` which turned out to be insufficient when training an acoustic model on that list.

Real recordings were taken from the multi-channel WSJ audio visual corpus [80]. Specifically, we used 8 channels of array 1 of the `olap_dev_5k` dataset (178 mixtures) for development and the `olap_ev1_5k` (142 mixtures) for test. No training set was available. To evaluate objective performance gains, we used the fairly clean headset signal. It is worth mentioning, that the recorded speech stems from British English speakers in contrast to our simulated database.

B. Separation neural network topology and training

The DC networks and DANs in this work all consist of two BLSTM layers with 600 forward and 600 backward units and a final linear layer mapping to embeddings with $E = 20$ dimensions. The forward and back streams of the BLSTM are concatenated before entering the next layer. All layers contain a sequence normalization and use dropout with a ratio of 0.5 during training. The DC networks employ a unit-norm normalization on the embeddings, while the DAN uses a tanh non-linearity on its embedding output. We train each network for 200 000 steps where each step consumes a mini-batch of 4 mixtures with ADAM [81].

TABLE II
COMPARISON OF DIFFERENT ENCODER VARIANTS FOR A FIXED DECODER ON THE SIMULATED DATABASE. BASELINE SYSTEMS ARE SET IN GRAY.

	Encoder	Latent model	Weight	Parameter	Decoder	SDR gain / dB		PESQ	STOI	WER / %	
						BSS-Eval	Invasive	gain	gain	Clean	Image
1		cACGMM	π_{kf}		MVDR	5.1	12.7	0.37	0.09	40.9	28.2
2	DC($K'=K$)	k-means	$1/K'$		MVDR	5.8	14.0	0.45	0.12	45.8	27.9
3	DC($K'=K+1$)	k-means	$1/K'$		MVDR	6.0	14.3	0.48	0.12	42.5	26.6
4	DC($K'=K+1$)	vMF-cACGMM	π_{kt}	$\kappa^{(\text{fix})}=5$	MVDR	**6.8**	**16.5**	**0.60**	**0.15**	**33.4**	**18.9**
5	DAN($K'=K$)	k-means	$1/K'$		MVDR	6.4	15.3	0.52	0.13	41.9	23.7
6	DAN($K'=K+1$)	k-means	$1/K'$		MVDR	5.8	14.0	0.46	0.11	44.9	28.1
7	DAN($K'=K$)	G-cACGMM	π_{kt}	$\sigma^{(\text{fix})}=0.6$	MVDR	**6.8**	16.4	0.59	0.14	35.8	19.9
8	DAN($K'=K+1$)	G-cACGMM	π_{kt}	$\sigma^{(\text{fix})}=0.2$	MVDR	5.9	14.6	0.49	0.11	42.1	26.4
9					Oracle					31.1	10.7

C. Acoustic model training

The hybrid AM consists of a combination of a Wide Residual Network to model local context and a BLSTM to model long term dependencies. The AM is thus dubbed wide bi-directional residual network (WBRN) [82]. The choice fell to a WBRN since it is considered state of the art on the single-channel track with baseline RNNLM rescoring during the CHiME 4 challenge. We train different acoustic models. The `clean` AM is trained directly on WSJ utterances with alignments extracted with a vanilla DNN-HMM recipe from Kaldi [83]. The `image` AM is trained on artificially reverberated WSJ utterances without any interfering speaker or noise. To obtain reliable alignments, we extracted these on the same utterances reverberated with a truncated room impulse response. Thus, both AMs never saw mixed speech and never saw possible artifacts produced by any kind of separation system. Therefore, we train a third AM (here named `match`) directly on the separation results of each algorithm of interest. Although warm-starting with a pre-trained AM is possible in this context, we trained each model from scratch without any significant degradation. For decoding we use the trigram language model delivered with the WSJ database without additional rescoring. All WERs are evaluated for one of the two speakers.

D. Analysis of integration methods

The BSS algorithms operate on an STFT with a discrete Fourier transform (DFT) size of 512 and a shift of 128. The AM uses 40 Mel filterbank features extracted with a DFT size of 256, a window size of 200 and a shift of 80.

Tbl. II compares different encoders on the simulated database with MVDR beamforming as a decoder. The first row shows the cACGMM result, which is entirely unsupervised. In contrast, row 2 and 5 show the vanilla DC and DAN systems both trained without an additional noise class where the DAN results in better WERs. The first step is now to introduce an additional noise class (row 3 and 6). This improves the DC result but severely harms the DAN performance. Therefore, we use the additional noise class only together with the DC system from now on. The best WERs with both the `clean` as well as the `image` AM are obtained with the DC encoder and a vMF-cACGMM latent model with additional noise class. Just as in [61], we used a time-variant mixture weight π_{kt} for all integration variants. The oracle results in Tbl. II are speech recognition results directly on the reverberated speech without interference or noise. It can be observed that the best PESQ and STOI scores are obtained with the vMF-cACGMM which leads to the conclusion that the integration is helpful both for ASR as well as speech enhancement.

The integration results in Tbl. II are obtained with a fixed concentration parameter for the vMF-cACGMM and a fixed scale parameter for the G-cACGMM. These values can be obtained on the development set as visualized in Fig. 6. It can be observed that the maxima in terms of invasive SDR gains coincide on the development and on the test set. The symbols on the right border indicate the development set performance when the parameter is estimated on the speech mixture. It can be observed that in almost all cases this is much worse than fixing the parameter using the development set.

E. Comparison of different decoders

It is often discussed which decoder variant is most suitable for speech recognition. Therefore, we compare different decoders in Tbl. III with a fixed DC encoder and a vMF-cACGMM latent model. It can be observed that the best BSS-

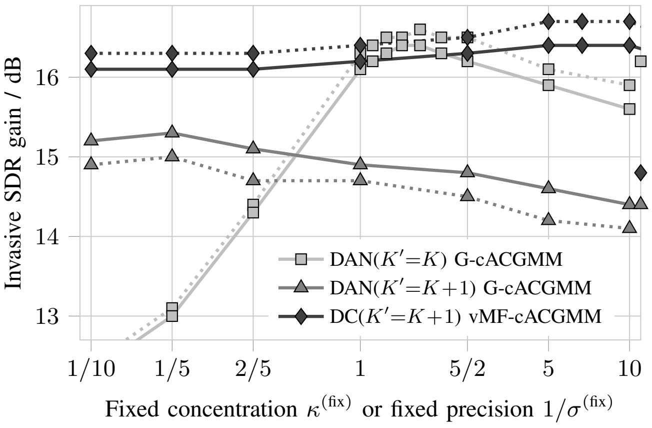

Fig. 6. Invasive SDR gain for different integration models and different parameter choices. Results on the development set are dashed while test results are solid lines. The symbols at the right border indicate development set results when κ or σ is estimated on each speech mixture instead of fixed.

TABLE III
COMPARISON OF DIFFERENT DECODER (SOURCE EXTRACTION) VARIANTS FOR A FIXED ENCODER AND LATENT MODEL: EACH VARIANT USES DC AS AN ENCODER AND A vMF-cACGMM AS A LATENT MODEL.

Decoder	SDR gain / dB		PESQ	STOI	WER / %		
	BSS-Eval	Invasive	gain	gain	Clean	Image	Match
Masking	**10.6**	14.2	0.47	**0.19**	62.9	39.5	18.6
GEV	4.5	14.6	**0.60**	0.11	39.9	21.6	14.7
GEV+BAN	7.4	16.0	0.57	0.14	34.3	**18.5**	15.4
MVDR	6.8	**16.5**	**0.60**	0.15	**33.4**	18.9	**13.9**
Oracle					31.1	10.7	10.7

Eval SDR gains are obtained using masking. However, the beamforming variants yield higher invasive SDR gains, higher perceptual gains (PESQ, STOI) and better word error rates with all three acoustic models. Nevertheless, it is worth noting that the masking decoder profits most from retraining a matched AM. Using the MVDR formulation as proposed by Souden et al. [71, Eq. 24] the proposed integration with a vMF-cACGMM yields a WER of 13.9 %, which is not much higher than the WER of a matched AM on oracle images with 10.7 %.

F. Unsupervised training for deep clustering

In many circumstances, artificial mixtures are not available or are not close enough to real recordings. It is therefore desirable to train, e.g., a DC neural network without external supervision. Probabilistic spatial mixture models have the advantage, that they can infer masks on a mixture without any training data. Therefore, it is a valid question, if an encoder such as DC can be trained without parallel data or oracle masks.

To do so, we first infer class affiliation posteriors (masks) with a cACGMM on a given mixture. Then, we apply a permutation alignment step (compare Subsec. IV-B) to obtain masks, which can be used for the affinity loss in Eq. 2. This can be seen as some kind of teacher-student training. Zhou and Qian suggested a similar scheme with complex Gaussian mixture model on multi-channel mixtures to fine-tune a (single-speaker) mask estimator [84].

Fig. 7 (left) nicely illustrates that the predicted posteriors (masks) from the spatial mixture model are rather rough and speckled. Furthermore, the predicted masks often contain frequency permutation errors, when the permutation alignment step did not resolve all permutations. Tbl. IV compares the performance of an encoder trained with supervision (DC) with the corresponding training without supervision (U-DC) by using the cACGMM as a trainer. The student initializing the cACGMM (row 3) beats the teacher (row 1) by 28 % relative WER and significantly improves SDR, PESQ and STOI gains. It can be deduced that indeed the DC encoder can be trained without parallel data, which is also illustrated by a mask produced by the student in Fig. 7 (right). Additionally, even if parallel data is partially available, using this approach to fine-tune on real recordings can be a viable option.

TABLE IV
COMPARISON OF DEEP CLUSTERING WITH SUPERVISION (DC) AND WITHOUT SUPERVISION (U-DC), WHERE THE cACGMM REPRESENTED IN THE FIRST ROW WAS USED AS A TEACHER FOR ROW 2 AND 3 INSTEAD OF IDEAL MASKS. BASELINE SYSTEMS ARE SET IN GRAY.

Encoder	Latent	SDR gain / dB		PESQ	STOI	WER / %	
	model	BSS-Eval	Invasive	gain	gain	Clean	Image
	cACGMM	5.1	12.7	0.37	0.09	40.9	28.2
U-DC	k-means	5.7	13.6	0.43	0.11	41.7	29.1
U-DC	cACGMM	**6.4**	**15.3**	**0.52**	**0.13**	**33.1**	**20.4**
DC	k-means	6.0	14.3	0.48	0.12	42.5	26.6
DC	cACGMM	6.1	14.9	0.50	0.12	34.4	21.6

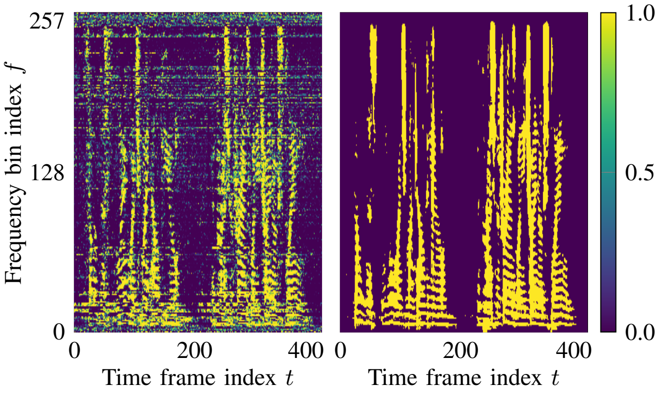

Fig. 7. Intermediate masks generated by the cACGMM (left) guide the neural network training which results in k-means clustering result with less artifacts (right). Especially the lower frequencies are resolved better.

G. Encoders with spatial features

It is of course possible to supply additional spatial features to the encoder network (here named spatial encoders) such that the neural network figures out spatial diversity by itself. For example, [49] additionally used the sine and cosine of inter-channel phase differences. Therefore, Tbl. V compares integration methods (row 1 and 2) with DC/DAN using spatial features as in [49] (row 3 and 4) and finally with integration methods which use a spatial encoder (row 5 and 6). Again, the fixed parameters are obtained on the development set.

First of all, it can be observed that the integration methods and DC/DAN using spatial features without integration do not differ greatly in terms of all reported measures. Although the integration methods already yield slightly higher invasive SDR gains, it is still beneficial to use an encoder with additional spatial features (row 5 and 6). To name an example, the Spatial-DC encoder with vMF-cACGMM yields a 3.7 % relative WER improvement over the Spatial-DC encoder with k-means. The perceptual gains (PESQ, STOI) are fairly similar with slight improvements by integrating a spatial encoder.

H. Evaluation on real recordings

To evaluate how the systems trained on artificial mixtures generalize to unseen real recordings and an entirely different microphone array geometry Tbl. VI compares separation results in terms of BSS-Eval SDR gain, PESQ gain and STOI gain measured against the headset microphone of the multi-channel

TABLE V
COMPARISON OF DIFFERENT INTEGRATION MODELS WITH A SINGLE-CHANNEL ENCODER VS. MODELS WITH A MULTI-CHANNEL ENCODER. THE MULTI-CHANNEL ENCODERS USE SINE AND COSINE OF INTER-CHANNEL PHASE DIFFERENCES AS SPATIAL FEATURES AS IN [49].

	Encoder	Latent model	Weight	Parameter	Decoder	SDR gain / dB		PESQ	STOI	WER / %	
						BSS-Eval	Invasive	gain	gain	Clean	Image
1	DC($K'=K+1$)	vMF-cACGMM	π_{kt}	$\kappa^{(\mathrm{fix})}=5$	MVDR	6.8	16.5	0.60	0.15	33.4	18.9
2	DAN($K'=K$)	G-cACGMM	π_{kt}	$\sigma^{(\mathrm{fix})}=0.6$	MVDR	6.8	16.4	0.59	0.14	35.8	19.9
3	Spatial-DC($K'=K+1$)	k-means			MVDR	6.7	16.2	0.59	0.15	34.5	18.7
4	Spatial-DAN($K'=K$)	k-means			MVDR	**6.9**	16.3	0.60	0.15	36.1	19.9
5	Spatial-DC($K'=K+1$)	vMF-cACGMM	π_{kt}	$\kappa^{(\mathrm{fix})}=5$	MVDR	**6.9**	**16.8**	**0.62**	**0.16**	**32.7**	**18.0**
6	Spatial-DAN($K'=K$)	G-cACGMM	π_{kt}	$\sigma^{(\mathrm{fix})}=2.5$	MVDR	**6.9**	**16.8**	**0.62**	0.15	33.5	18.8
7					Oracle					31.1	10.7

WSJ audio visual corpus. The headset microphone is not an optimal reference but may serve here as a proxy. Invasive SDR gains are unavailable on real recordings since the calculation requires oracle source images and noise images. One potential reason why the WERs are substantially worse than the previous results is the mismatch between an American English AM and British English observations. They can therefore only be compared relative to each other.

First of all the cACGMM does not suffer from any training mismatch because it estimates all parameters ad-hoc on the current mixture. The single-channel DC here performs worse than the cACGMM. If we compare this to row 1 and 2 in Tbl. II, we may conclude that the single-channel DC indeed has mismatch issues. This mismatch is very well compensated when using the integrated vMF-cACGMM with the single-channel DC encoder (row 3).

However, it turns out that the encoder with spatial features generalizes surprisingly well to the unseen microphone geometry (row 4). The DC encoder with k-means clustering performs just as well as the integrated approach with a single-channel encoder (row 3). Nevertheless, the best results both in terms of reported signal level measures as well as WER are obtained using an integrated vMF-cACGMM with a spatial encoder.

VIII. CONCLUSIONS

In this work, we presented an integrated approach to blind source separation combining neural network-based methods (i.e. deep clustering and deep attractor networks) with probabilistic spatial mixture models. The integration was achieved by defining a mixture model with two kinds of observation distributions, one corresponding to the embedding vectors obtained by the neural network and one for the vector of microphone signals, while both modalities share the same latent class affiliation variable. Our key findings from our experimental evaluation are (a) the integration model consistently outperforms the individual components, (b) the integration model, but also a neural network-based separation method with spatial features, are fairly robust to microphone mismatch even when evaluating on real recordings, (c) a student neural network trained with supervision from an unsupervised spatial mixture model is able to separate speech and outperform the teacher.

TABLE VI
EVALUATION ON REAL RECORDINGS TAKEN FROM THE MULTI-CHANNEL WSJ DATABASE WITH BRITISH ENGLISH. GAINS ARE MEASURED WITH RESPECT TO HEADSET MICROPHONE. ALL SYSTEMS HAVE A NOISE CLASS AND EXTRACT THE SOURCES WITH MVDR BEAMFORMING.

Encoder	Latent	SDR gain / dB	PESQ	STOI	WER / %	
	model	BSS-Eval	gain	gain	Clean	Image
	cACGMM	6.9	0.41	0.15	54.3	49.7
DC	k-means	6.3	0.32	0.13	67.1	58.6
DC	vMF-cACGMM	8.5	0.55	0.20	43.8	41.8
Spatial-DC	k-means	8.7	0.54	0.20	48.3	41.8
Spatial-DC	vMF-cACGMM	**9.0**	**0.58**	**0.21**	**43.0**	**39.7**

ACKNOWLEDGMENT

Computational resources were provided by the Paderborn Center for Parallel Computing.

APPENDIX
REPRODUCABILITY INSTRUCTIONS

To be able to reproduce the results of our implementation of the probabilistic spatial models including models not analyzed here can be found at `https://github.com/fgnt/pb_bss`. The code also includes a permutation alignment algorithm. The room impulse responses can be generated with the implementation found at `https://github.com/ehabets/RIR-Generator`. A Python implementation of the BSS-Eval SDR performance measure [74] is available at `https://github.com/craffel/mir_eval`. The file lists for the simulated database as well as the TensorFlow code for the DC and DAN training are available upon request.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF–half-baked or well done?" *Mitsubishi Electric Research Labs (MERL), Technical Report*, 2015.

[6] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," in *Advances in Neural Information Processing Systems (NIPS)*, 1996, pp. 472–478.

[7] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *International Conference on Spoken Language Processing (SLT)*, 2006, pp. 97–100.

[8] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[10] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[11] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Deep neural network based speech separation for robust speech recognition," in *International Conference on Signal Processing (ICSP)*. IEEE, 2014, pp. 532–536.

[12] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Communication*, vol. 95, pp. 28–39, 2017.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[14] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016.

[15] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5064–5068.

[16] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[17] C. Stephenson, P. Callier, A. Ganesh, and K. Ni, "Monaural audio speaker separation with source contrastive estimation," *arXiv preprint arXiv:1705.04662*, 2017.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[19] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[20] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[21] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[22] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[23] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[24] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

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

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

[27] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

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

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

[30] L. Drude, C. Boeddeker, and R. Haeb-Umbach, "Blind speech separation based on complex spherical k-mode clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 141–145.

[31] N. Ito and T. Nakatani, "FastFCA-AS: Joint diagonalization based acceleration of full-rank spatial covariance analysis for separating any number of sources," *arXiv preprint arXiv:1805.09498*, 2018.

[32] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2007, pp. 3247–3250.

[33] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Normalized observation vector clustering approach for sparse source separation," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2006, pp. 1–5.

[34] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[35] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[36] J. Woodruff and D. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1856–1866, 2010.

[37] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2516–2531, 2013.

[38] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation exploiting temporal and spectral correlations using 2D-HMMs," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2013, pp. 1–5.

[39] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[40] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[41] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[42] L. Drude, T. Higuchi, K. Kinoshita, T. Nakatani, and R. Haeb-Umbach, "Dual frequency- and block-permutation alignment for deep learning based block-online blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[43] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.

[44] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[45] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[46] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[47] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[48] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams, "Structured vaes: Composing probabilistic graphical models and variational autoencoders," *arXiv preprint arXiv:1603.06277*, 2016.

[49] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[50] Z. Chen, Y. Luo, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *arXiv preprint arXiv:1707.03634*, 2017.

[51] D. H. Tran Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2010.

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

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

[54] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6834–6838.

[55] J. T. Kent, "The complex Bingham distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56, no. 2, pp. 285–299, 1994.

[56] N. Ito, S. Araki, and T. Nakatani, "Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 465–468.

[57] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 253–256.

[58] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1153–1157.

[59] J. T. Kent, "Data analysis for shapes and images," *Journal of statistical planning and inference*, vol. 57, no. 2, pp. 181–193, 1997.

[60] D. H. Tran Vu and R. Haeb-Umbach, "On initial seed selection for frequency domain blind speech separation," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1757–1760.

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

[62] E. Vincent, N. Bertin, R. Gribonval, F. Bimbot *et al.*, "From blind to guided audio source separation," *IEEE Signal Processing Magazine*, 2013.

[63] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop*, 2018.

[64] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

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

[66] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[67] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 196–200.

[68] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 1981–1985.

[69] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[70] B. Van Veen and K. Buckley, "Beamforming techniques for spatial filtering," *Digital Signal Processing Handbook*, pp. 61–1, 1997.

[71] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[72] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[73] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2007, pp. 41–44.

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

[75] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.

[76] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[77] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Workshop on Speech and Natural Language (HLT)*. Association for Computational Linguistics, 1992, pp. 357–362.

[78] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[79] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[80] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2005, pp. 357–362.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[82] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *International Workshop on Speech Processing in Everyday Environments (CHiME16)*, 2016, pp. 12–17.

[83] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[84] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *(unpublished)*, 2018.

Lukas Drude is a Ph.D. student at Paderborn University since 2014 where he also pursued his Bachelor's and Masters's degree in Electrical Engineering. His research interests range from blind speech separation to automatic speech recognition with a focus on integrating probabilistic graphical models and neural networks. In 2015 he was a visiting researcher at Carnegie Mellon University, USA. In 2017 he pursued a research internship with NTT Communication Science Laboratories, Kyoto, Japan.

Reinhold Haeb-Umbach is a professor of Communications Engineering at Paderborn University, Germany. He holds a Dr.-Ing. degree from RWTH Aachen University, and has a background in speech research both in an industrial and academic research environment. His main research interests are in the fields of statistical signal processing and pattern recognition, with applications to speech enhancement, acoustic beamforming and source separation, as well as automatic speech recognition and unsupervised learning from speech and audio. He has more than 200 scientific publications, and recently co-authored the book Robust Automatic Speech Recognition – a Bridge to Practical Applications (Academic Press, 2015). From 2015 – 2020 he is/has been a member of the IEEE Signal Processing Society Speech and Language Technical Committee. He is a fellow of the International Speech Communication Association (ISCA), class of 2015.