SOURCE COUNTING IN SPEECH MIXTURES BY NONPARAMETRIC BAYESIAN ESTIMATION OF AN INFINITE GAUSSIAN MIXTURE MODEL

Oliver Walter, Lukas Drude and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany

ABSTRACT

In this paper we present a source counting algorithm to determine the number of speakers in a speech mixture. In our proposed method, we model the histogram of estimated directions of arrival with a nonparametric Bayesian infinite Gaussian mixture model. As an alternative to classical model selection criteria and to avoid specifying the maximum number of mixture components in advance, a Dirichlet process prior is employed over the mixture components. This allows to automatically determine the optimal number of mixture components that most probably model the observations. We demonstrate by experiments that this model outperforms a parametric approach using a finite Gaussian mixture model with a Dirichlet distribution prior over the mixture weights.

Index Terms— Source counting, Blind source separation, nonparametric Bayesian methods, Chinese restaurant process

1. INTRODUCTION

Multichannel blind source separation (BSS) algorithms are used to extract source signals, observing only a mixture of them at multiple microphones. BSS algorithms in general consider the source locations, transfer functions and source signals to be unknown. On the other hand, knowledge of the number of active sources is often assumed. In a practical application, however, the number of sources is usually unknown and has to be estimated.

Several algorithms have been proposed to estimate the number of sources. Model selection algorithms, that compare models of different order with respect to criteria such as penalized likelihood or Bayesian information [1], are usually computationally expensive. An alternative are finite mixture models, where the number of component densities is determined in the course of the mixture parameter estimation. In [2] a variational Expectation Maximization (EM) algorithm for complex Gaussian mixture models is employed. Starting from an assumed maximum number of sources the variational EM iterates until only a few mixture weights remain significantly larger than zero. Their number constitutes then the estimate of active speakers. While this approach modeled the distribution of the microphone signals in the short-time Fourier transform (STFT) domain, the distribution of the directions of arrival (DoA), which are computed from the microphone signals, is modeled in [3]. They employed a real-valued Gaussian mixture model (GMM) and proposed to impose a Dirichlet distribution as a prior on the mixture weights. The Dirichlet distribution prior is parametrized such, that only dominant mixture components retain a high mixture weight in the course of the iterations. The number of speakers is then determined based on a weight threshold. In [4] the phase, frequency and amplitude normalized vector of microphone signals was modeled by a complex Watson mixture model. A variational EM algorithm was derived, and the number of active speakers was determined by a stepwise source counting algorithm: after identifying a source, its contribution to the microphone signal is eliminated before searching for the next source. This process is repeated until a maximum number of potential sources is reached. The final source number estimate was then determined by a threshold on the mixture weights and the concentration parameters of the complex Watson distributions.

All these aforementioned approaches have in common that the maximum number of sources has to be specified in advance. In this contribution we propose an approach, which avoids exactly this. Instead of using a parametric model in terms of the number of mixture components of a mixture model, we employ a Dirichlet process (DP) prior over the mixture components. The resulting model is also known as an infinite mixture model [5], which allows the number of mixture components to adapt to the observations and be potentially infinite. While this approach can be used for any kind of representations of the microphone signals, we employ here DoA estimates computed from the multichannel microphone input. The DoAs are modeled by a mixture of wrapped Gaussians as in [3]. We propose to use the Chinese restaurant process (CRP) representation of the Dirichlet process, where the mixture components of the GMM and its weights are assigned by the CRP. The parameters of the Gaussians of each mixture component, namely the means and variances, are assumed to be drawn from a normal-gamma distribution. Employing the predictive posterior distribution, this effectively results in a mixture of Student's t-distributions, where the number of mixture components is learned automatically. In [6] an infinite Gaussian mixture model (IGMM) is used to estimate models for nonstationary noise. We extend the approach of [6] by the wrapped phase Gaussian model of [3]. Moreover we modify the CRP formulation to allow weighted counts for customers instead of a fixed count of one per customer. This allows to place more emphasis on observations which are presumably less corrupted by noise. A selection or weighting of observations for source counting has also been proposed in [3, 7, 8].

The paper is organized as follows: In section 2 we describe the signal model, followed by a description of the generative process of the IGMM in section 3. Section 4 gives an overview over the parameter estimation procedure while section 5 describes the application to source counting. Experimental results are reported in section 6 and conclusions are drawn in section 7.

2. SIGNAL MODEL

Consider a convolutive mixture model of K independent source signals $S_k(\tau, f)$ captured by D microphones yielding the sensor signals

Supported by Deutsche Forschungsgemeinschaft under contract no. Ha 3455/9-1 within the Priority Program SPP1527 "Autonomous Learning".

$X_d(\tau, f)$ in the STFT domain [9]:

$$\mathbf{X}(\tau, f) = \sum_{k=1}^{K} \mathbf{H}_k(f) S_k(\tau, f) + \mathbf{N}(\tau, f), \quad (1)$$

where $\mathbf{X} = (X_1, \ldots, X_D)^{\mathrm{T}}$ is the vector of sensor signals, $\mathbf{H}_k = (H_{1,k}, \ldots, H_{D,k})^{\mathrm{T}}$ is the vector of multiplicative transfer functions associated to source k, and $\mathbf{N} = (N_1, \ldots, N_D)^{\mathrm{T}}$ is the noise vector, with time frames τ from 1 to T and frequency bins f from 1 to F. A time difference of arrival (TDoA) vector $\mathbf{q}(\tau, f) = (\ldots, q_{ij'}, \ldots)^{\mathrm{T}}$ for all microphone pairs j-j' is calculated as $q_{ij'}(\tau, f) = \frac{1}{2\pi f_{\text{real}}} \arg\left(X_j(\tau, f) X_{j'}^*(\tau, f)\right)$, with $f_{\text{real}} = \frac{f}{M} f_{\text{s}}$, where $F = M/2 - 1$. With the matrix $\mathbf{D} = [\ldots, \mathbf{p}_j - \mathbf{p}_{j'}, \ldots]^{\mathrm{T}}$ of differences of two-dimensional position vectors $\mathbf{p}_j$ of all microphone pairs and the sound velocity v, employing the Moore-Penrose pdeudo-inverse, denoted by $^+$, we can estimate the DoAs $\psi(\tau, f)$:

$$\begin{bmatrix} \cos\psi(\tau, f) \\ \sin\psi(\tau, f) \end{bmatrix} = v\mathbf{D}^+\mathbf{q}(\tau, f). \quad (2)$$

We stack all DoAs $\psi(\tau, f)$ in a vector $\mathbf{d} = [d_n]$, with the n-th DoA observation $d_n := \psi(\tau, f) \forall \tau, f \wedge n = (\tau - 1)F + f$ and $n \in 1, \ldots, T \cdot F$. This allows us to consider all DoAs together.

3. INFINITE GAUSSIAN MIXTURE MODEL

The input to our algorithm are the DoA measurements as described in the last section. Since the range of the DoAs is $[-\pi, \pi)$, for observations close to $-\pi$ and π, the observed distribution would become bimodal. To overcome this effect, all possible shifts of an observation by 2π have to be accounted for [3].

Let k_n be the shift of the n-th DoA observation. A mixture component and shift conditional distribution, where μ_l is the mean and σ_l^2 the variance of the l-th Gaussian component, is then given by:

$$p\left(d_n|\mu_l, \sigma_l^2, k_n\right) = \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(\frac{-(d_n + 2\pi k_n - \mu_l)^2}{2\sigma_l^2}\right). \quad (3)$$

Each component of the mixture distribution is assumed to model one of the sources, with its mean being the direction of the source. Let $z_n = l$ be an indicator variable, indicating that the n-th observation belongs to the l-th mixture component, $P(z_n = l)$ the weight of the mixture component and L the total number of mixture components. For a finite mixture of wrapped Gaussians we sum over the weighted mixture components and the shifts, assuming that each shift has the same probability. Denoting the set of all means as $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_L\}$ and the set of all variances as $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_L\}$, we get:

$$p\left(d_n|\boldsymbol{\mu}, \boldsymbol{\sigma}^2\right) = \sum_{l=1}^{L} P(z_n = l) \sum_{k_n=-\infty}^{\infty} p\left(d_n|\mu_l, \sigma_l^2, k_n\right). \quad (4)$$

The generative process of the IGMM, using the CRP-based representation, is now assumed as follows: First, for all mixture components $l \in \{1, \ldots, \infty\}$ the parameters $\boldsymbol{\Theta}_l = \{\mu_l, \sigma_l^2\}$ are sampled from the same normal-gamma distribution with mean $m^{(0)}$, concentration $\xi^{(0)}$, shape $\eta^{(0)}$ and scale $r^{(0)}$:

$$\boldsymbol{\Theta}_l = \{\mu_l, \sigma_l^2\} \sim \mathcal{N}\left(\mu; m^{(0)}, \sigma^2/\xi^{(0)}\right) \mathcal{G}\left(\sigma^{-2}; \eta^{(0)}, r^{(0)}\right) \quad (5)$$

Next, a CRP is used to generate the indicator variables z_n. A CRP is a sequential process where each z_n is generated, considering the previous $z_1, \ldots, z_{n-1}$ indicators. Let n_l be the number of observations assigned to the l-th mixture component and N the total number of observations generated so far. For a new observation d_{N+1}, an existing mixture component l with $n_l > 0$ is chosen with probability

$$P(z_{N+1} = l|z_1, \ldots, z_N) = n_l/(N + \gamma). \quad (6)$$

Alternatively, a new mixture component is created with probability

$$P(z_{N+1} = l_{\text{new}}|z_1, \ldots, z_N) = \gamma/(N + \gamma), \quad (7)$$

where γ is the concentration parameter of the Dirichlet process.

If a new mixture component is created, a $\boldsymbol{\Theta}_l$ with $n_l = 0$ is assigned to the observation. The process starts with generating the first observation and therefore the first mixture component. For every subsequent observation either an existing mixture component is chosen or a new one is created. Since only a finite number of observations is generated, the number of created mixture components will also be finite and upper bounded by the number of observations. Next, a shift k_n is sampled for each observation. For simplicity we limit the shifts to be between $-K$ and K and assume equal prior probability for the possible shifts. Finally the observations d_n are generated using the distribution (3).

4. PARAMETER ESTIMATION

The problem of learning is now to determine the number of mixture components and their parameters $\boldsymbol{\Theta}_l$. For the parameter estimation we use a hybrid approach consisting of Gibbs sampling and maximum a posteriori (MAP) estimation.

4.1. Parameter Update

First, the assignment of an observation to a mixture component is sampled using Gibbs sampling. The CRP has the property that its random variables are interchangeable, which means that the order of the variables does not matter when calculating the probability for a certain variable. Using this property it is easy to devise a Gibbs sampler. In Gibbs sampling the value of one variable is resampled given the values of all other random variables. Therefore, to resample a certain indicator variable, we need the posterior probability of $z_n = l$, given all observations and the set of hyper parameters $\boldsymbol{\Theta}^{(0)} = \{m^{(0)}, \xi^{(0)}, \eta^{(0)}, r^{(0)}\}$ of the prior distribution. We denote $\mathbf{z}_{\setminus n}$ as the set of indicator variables without the n-th one, since we want to resample the n-th one. The set of all observations without the current one is denoted by $\mathbf{d}_{\setminus n}$. The set of all shifts without the current one is $\mathbf{k}_{\setminus n}$. The posterior probability of $z_n = l$ can then be calculated as follows:

$$P(z_n = l|d_n, \mathbf{d}_{\setminus n}, \mathbf{z}_{\setminus n}, \mathbf{k}_{\setminus n}, \boldsymbol{\Theta}^{(0)}) \propto$$
$$P(z_n = l|\mathbf{z}_{\setminus n}) p(d_n|\mathbf{d}_{\setminus n}, z_n = l, \mathbf{z}_{\setminus n}, \mathbf{k}_{\setminus n}, \boldsymbol{\Theta}^{(0)}). \quad (8)$$

For the mixture component weights $P(z_n = l|\mathbf{z}_{\setminus n})$ we use equations (6) and (7). For the second term we use the predictive probability for an observation d_n belonging to the l-th class. We have to integrate out the mixture component parameters $\boldsymbol{\Theta}_l$ and sum over the shifts k_n in equation (3):

$$p(d_n|\mathbf{d}_{\setminus n}, z_n = l, \mathbf{z}_{\setminus n}, \mathbf{k}_{\setminus n}, \boldsymbol{\Theta}^{(0)}) \propto$$
$$\sum_{k_n=-K}^{K} \int p(d_n|\boldsymbol{\Theta}_l, k_n) p(\boldsymbol{\Theta}_l|\mathbf{d}_{\setminus n}, z_n = l, \mathbf{z}_{\setminus n}, \mathbf{k}_{\setminus n}, \boldsymbol{\Theta}^{(0)}) \mathrm{d}\boldsymbol{\Theta}_l \quad (9)$$

$$\propto \sum_{k_n=-K}^{K} \mathcal{T}(d_n + 2\pi k_n; m_l, \xi_l, \eta_l, r_l). \quad (10)$$

Integrating out the parameter $\mathbf{\Theta}_l$, using the posterior distribution of the parameter, delivers the predictive posterior distribution, which is a Student's t-distribution:

$$\mathcal{T}(x; m, \xi, \eta, r) = \left(\frac{\xi}{2\pi r(\xi+1)}\right)^{\frac{1}{2}} \frac{\Gamma(\eta+\frac{1}{2})}{\Gamma(\eta)} \left(1+\frac{\xi(x-m)^2}{2r(\xi+1)}\right)^{-(\eta+\frac{1}{2})}. \quad (11)$$

Summing over k_n, assuming a uniform prior and limiting the possible shifts between $-K$ and K, results in a wrapped Student's t-distributions according to (10). In case of $n_l > 0$, which is a mixture component with observations assigned to it, the parameters m_l, ξ_l, η_l, r_l of the predictive posterior distribution are used. For the case $z_n = l_{\text{new}}$, the prior parameters $\mathbf{\Theta}^{(0)}$ are used instead.

Next, we have to estimate the shift k_n for the observation d_n. To do this we employ the posterior probability of the shift k_n given all other variables. We then decide for the shift, that maximizes the posterior probability. To obtain the posterior we again integrated out $\mathbf{\Theta}_l$ in equation (3). Assuming a uniform prior probability for the shifts k_n, the posterior is also a Student's-t distribution:

$$P(k_n|d_n, \mathbf{d}_{\setminus n}, z_n = l, \mathbf{z}_{\setminus n}, \mathbf{k}_{\setminus n}, \mathbf{\Theta}^{(0)}) \propto \mathcal{T}(d_n + 2\pi k_n; m_l, \xi_l, \eta_l, r_l). \quad (12)$$

Having the class assignments z_n and shifts k_n we can update the parameters m_l, ξ_l, η_l and r_l of the predictive posterior distribution.

$$\xi_l = \xi^{(0)} + s_{0,l}, \quad (13)$$

$$m_l = (\xi^{(0)} m^{(0)} + s_{1,l})/\xi_l, \quad (14)$$

$$\eta_l = \eta^{(0)} + s_{0,l}/2, \quad (15)$$

$$r_l = r^{(0)} + \left(s_{2,l} + \xi^{(0)}(m^{(0)})^2 - \xi_l \mu_l^2\right)/2, \quad (16)$$

where the parameters are updated employing the sufficient statistics $s_{0,l}, s_{1,l}$ and $s_{2,l}$ of each mixture component:

$$s_{0,l} = n_l, \quad (17)$$

$$s_{1,l} = \sum_{n:z_n=l} (d_n - 2\pi k_n), \quad (18)$$

$$s_{2,l} = \sum_{n:z_n=l}^{N} (d_n - 2\pi k_n)^2. \quad (19)$$

Summarizing all steps, the parameters are estimated by iterating over the observations in a random order. First, an observation is removed from the sufficient statistics of its mixture component, then the indicator variable z_n is resampled, the shift k_n estimated and finally the observation added back to the sufficient statistics of the chosen mixture component. The parameters of the predictive posterior distributions are updated after every change of the sufficient statistics. This is repeated for several iterations. The parameter set after the final iteration is then used for the source counting.

The algorithm starts with no observations assigned to a mixture component and no mixture components created. Observations are assigned to chosen or created mixture components in the first iteration and then added to the corresponding sufficient statistics.

4.2. Parameter Update with Weighted Observations

We want to emphasize the observations of active speakers with high power by power weighting. The rationale behind this is that observations with high power are probably little affected by noise and thus the DoA estimates obtained from them are supposedly more reliable than those obtained from samples with low power.

Therefore we introduce a weight a_n corresponding to the power of the n-th observation, similar to [3]:

$$a_n = c|X_{1,n}|^2 / \sum_{\tilde{n}=1}^{T \cdot F} |X_{1,\tilde{n}}|^2. \quad (20)$$

Where the summation is over all frequency bins and time frames. The constant c is a scaling factor, which is set to $c = 1000$, and $X_{1,n}$ is the signal at the first microphone. Another beneficial effect of the above weighting of the DoAs is that the histogram of the weighted DoAs is more Gaussian like than of the unweighted DoAs.

The calculation of the sufficient statistics is modified to include the power weighting. The sufficient statistics are iteratively updated by subtraction before each Gibbs sampling step and addition afterwards:

$$s_{0,l=z_n} \leftarrow s_{0,l=z_n} \pm a_n, \quad (21)$$

$$s_{1,l=z_n} \leftarrow s_{1,l=z_n} \pm a_n(d_n - 2\pi k_n) \quad (22)$$

$$s_{2,l=z_n} \leftarrow s_{2,l=z_n} \pm a_n(d_n - 2\pi k_n)^2 \quad (23)$$

All sufficient statistics of the mixture components with no observations assigned are initialized to zero.

Note, that with the introduction of power weights, we also change the CRP to a weighted CRP. Instead of counting an observation with a weight of one, we count it with its corresponding weight. Using the sufficient statistics $s_{0,l}$, the equations (6) and (7) become

$$P(z_n = l|\mathbf{z}_{\setminus n}) \propto s_{0,l} \text{ and } P(z_n = l_{\text{new}}|\mathbf{z}_{\setminus n}) \propto \gamma. \quad (24)$$

Before the first iteration we initialize the hyper parameters of the prior distribution. Due to the power weighting, observations can have a small weight. We therefore set the concentration parameter $\xi^{(0)}$ and the shape parameter $\eta^{(0)}$ to small values to obtain prior weights in the same scale of the observation weights:

$$\xi_0 = 5 \times 10^{-3}, \eta_0 = 5 \times 10^{-3}, \quad (25)$$

$$m^{(0)} = \frac{1}{c} \sum_{n=1}^{T \cdot F} a_n d_n, \quad (26)$$

$$r^{(0)} = \eta_0 \left(\frac{1}{c} \sum_{n=1}^{T \cdot F} a_n \left(d_n - m^{(0)}\right)^2\right). \quad (27)$$

5. SOURCE COUNTING

To obtain the final number of sources we apply further steps in the estimation procedure.

As a first step we reduce the value of Dirichlet process concentration parameter γ by a factor of 100 after a burn-in period. This leads to the removal of mixture components with low weights and reduces the amount of newly created mixture components. Experiments showed, that the length of the burn-in period should be chosen long enough for an initialization, and the factor of the reduction high enough to reduce the number of newly created classes per iteration. We did not observe changes in performance with higher reductions.

After finishing the iterations, mixture components with a mean close to the mean of mixture components with higher weights are removed by only keeping those mixture components whose means have the highest probability under their own distribution.

Finally, mixture components mainly modeling the noise floor, with a variance $\sigma_l^2 = r_l/\eta_l$ higher than 10 times the minimal variance of all mixture components, are removed. In the experimental results we show that the performance of speaker counting algorithm is rather insensitive to the exact value of this factor and that it can be chosen independently of the SNR and the minimal speaker spacing.

The remaining mixture components then model the distribution of the individual DoAs of the speakers. The number of mixture components is the estimate for the number of speakers.

6. EXPERIMENTAL RESULTS

In a simulation environment, up to six speech sources are placed on a circle of radius 1 m around an array of $D = 3$ omnidirectional microphones arranged in a triangular shape with 2 cm edge length. The number of possible source positions is varied from four to up to eight with an equal spacing of 90°, 60° and 45°. The sources and the sensor array share the same height of 1.5 m.

Speech samples of 5 s length and sampling frequency f_s=16 kHz are chosen at random from the training utterances of the TIMIT database [10]. Speech samples of zero to up to six speakers without speech pauses are convolved with impulse responses of a simulated non-reverberant room of dimension 4 m $\times$ 4 m $\times$ 3 m and mixed.

An STFT with frame size $M = 1024$ and a frame shift of 256 is applied to each sensor signal. The SNR for white Gaussian noise is varied from 10 dB to 30 dB.

For the Gibbs sampling algorithm about 6 % of the observations with highest power are chosen for the parameter estimation. The concentration parameter of the Dirichlet process prior is set to $\gamma = 100$ for the first 75 burn in iterations and then reduced to $\gamma = 1$ for the remaining 225 iterations, where no change in performance was observed with more iterations. Our unoptimized C++ implementation took about 15 s per recording on a Core i7 960.

Figure 1 shows the results of 100 trials for our Dirichlet process prior-based algorithm and, for comparison, an algorithm with Dirichlet distribution prior. The parameters and initial values for the Dirichlet distribution prior-based counting algorithm are taken directly from the corresponding publication [3]. As a performance measure we used the accuracy, i.e., the percentage of times the number of sources is correctly estimated. It can be seen that our algorithm delivers a better performance. For example, for a minimum speaker spacing of 90° the proposed algorithm almost always counts the number of active speakers correctly for all three SNR values, while the method with the Dirichlet distribution prior does so only if the number of active speakers is small and the SNR is high. Furthermore, the results are comparable to a state of the art algorithm using complex Watson mixture models that we published in [8].

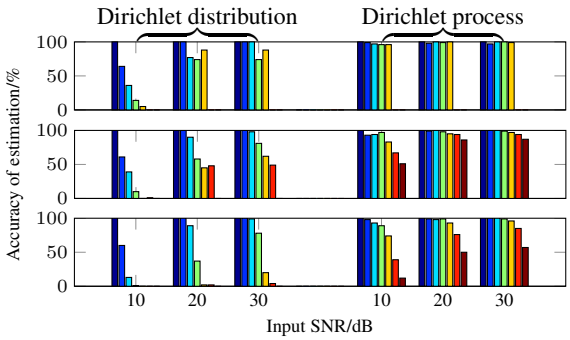

Fig. 1. Speaker counting accuracy in % with Dirichlet distribution prior and Dirichlet process prior over different number of active speakers with different minimal speaker spacings (Top to bottom: up to 4 speakers - 90°, up to 6 speakers - 60°, up to 6 speakers - 45°, different input SNRs (10 dB, 20 dB, 30 dB) and different number of active speakers (left to right or blue to red: 0, 1, 2, 3, 4, 5, 6).

We observed, that the variances of the estimated mixture components increase with decreasing SNR. On the other hand, mixture components corresponding to valid speakers always had the lowest variances among the mixture components. Therefore we decided to use an adaptive thresholding value on the mixture component variances to discard mixture components with high variance. The thresholding value is automatically determined from the data by finding the minimal variance in all mixture components and discarding all mixture components that have a 10 times higher variance. Figure 2 shows the sensitivity of our algorithm against this scaling factor. It can be seen that the scaling factor is independent of the SNR and speaker spacing. The best results are achieved for a scaling factor between 5 and 15, with the optimum around 10, the value that has been chosen for the results presented in Fig. 1.

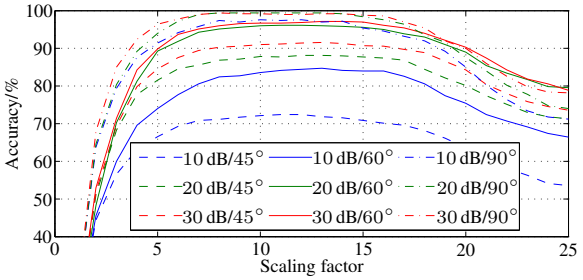

Fig. 2. Speaker counting accuracy in % vs. variance scaling factor at different SNRs and minimal speaker spacings

It is important to note that we present results for an anechoic environment only. In the presence of reverberation the frequency normalization described in Section 2, which assumes a linear dependency of the phase on the frequency, and the small microphone spacing of 2 cm are unfavorable. We employed this normalization to allow for a fair comparison with the algorithms of [3] and [8]. For the presented algorithm, the normalization is in principle not necessary and could be eliminated, however at the expense of introducing a permutation problem, for which, however, solutions are given in the literature.

7. CONCLUSIONS AND OUTLOOK

We have presented a source counting algorithm based on the Bayesian estimation of an infinite Gaussian mixture model. Unlike earlier approaches to source counting based on estimating the number of components of a mixture model, the maximum number of active sources does not need to be specified in advance. The experimental results show that the proposed Dirichlet process prior-based algorithm outperforms a comparable algorithm using a Dirichlet distribution prior. A dynamic thresholding based on a variance floor learned from the data was proposed to reduce the dependency of the flooring parameter on the SNR.

Based on these results we believe that the nonparametric modeling in terms of mixture components can help to increase the performance over parametric models with a fixed number of mixture components. Since the CRP formulation for infinite mixture models can be applied to almost any kind of mixture models, it can also be applied to a complex Watson mixture model (cWMM). Recently we have shown [4, 8] that a cWMM is an appropriate model of the observation vector computed from the microphone signals and can be used to derive source counting, beamforming and source separation algorithms. An exension of the parametric cWMM to a nonparametric infinite cWMM is therefore a promising topic for future research.

8. REFERENCES

[1] C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.

[2] 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 *Proc. ICASSP*, 2012, pp. 253–256.

[3] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. ICASSP*, 2009, pp. 33–36.

[4] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6834–6838.

[5] C. E. Rasmussen, "The infinite Gaussian mixture model.," in *NIPS*, 1999, vol. 12, pp. 554–560.

[6] M. Fujimoto, Y. Kubo, and T. Nakatani, "Unsupervised non-parametric Bayesian modeling of non-stationary noise for model-based noise suppression," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 5562–5566.

[7] B. Lösch, *Complex Blind Source Separation with Audio Applications*, Ph.D. thesis, Stuttgart University, 2013.

[8] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, "Towards online source counting in speech mixtures applying a variational EM for complex Watson mixture models," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan Les Pins, France, Sept. 2014.

[9] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.