TOWARDS ONLINE SOURCE COUNTING IN SPEECH MIXTURES APPLYING A VARIATIONAL EM FOR COMPLEX WATSON MIXTURE MODELS

Lukas Drude, Aleksej Chinaev, Dang Hai Tran Vu, Reinhold Haeb-Umbach

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

ABSTRACT

This contribution describes a step-wise source counting algorithm to determine the number of speakers in an offline scenario. Each speaker is identified by a variational expectation maximization (VEM) algorithm for complex Watson mixture models and therefore directly yields beamforming vectors for a subsequent speech separation process. An observation selection criterion is proposed which improves the robustness of the source counting in noise. The algorithm is compared to an alternative VEM approach with Gaussian mixture models based on directions of arrival and shown to deliver improved source counting accuracy. The article concludes by extending the offline algorithm towards a low-latency online estimation of the number of active sources from the streaming input data.

Index Terms— Blind source separation, Bayes methods, Directional statistics, Number of speakers, Speaker diarization

1. INTRODUCTION

Blind speech separation (BSS) algorithms are designed to improve signal enhancement metrics. While these algorithms in general consider the source locations and the source signals to be unknown, they do generally assume knowledge of the number of speakers to be found [1, 2]. Although the number of sources is unknown in almost every practical application, relatively few articles are concerned with estimating the number of active sources for BSS algorithms.

In [3], time frequency (tf) slots are identified first, which are presumably dominated by a single speaker. A histogram of directions of arrival (DOA) is computed from these slots and the number of sources is determined by counting the number of significant peaks.

The authors of [4] employ a variational Expectation Maximization (VEM) algorithm to complex Gaussian Mixture Models (GMM), which model the distribution of the microphone array signals in the Short Time Fourier Transform (STFT) domain. Starting from an assumed maximum number of sources the VEM is iterated after which only as many mixture weights remain significantly larger than zero as there are simultaneously active sources. Since this is done for each frequency bin separately, the permutation problem has to be solved afterwards.

In [5] the permutation problem is avoided by phase and frequency normalization. As a result the DOAs of a speaker form a single cluster, irrespective of the considered frequency bin. The DOAs are modeled by a real-valued GMM and a VEM is run, which reveals, after a sufficiently large number of iterations, the number of active speakers by the number of mixture weights remaining significantly larger than zero.

In [6] we introduced a source counting algorithm which employs the vector of microphone signals directly, rather than estimating DOAs from it. After phase, frequency and norm normalization, the direction vectors form clusters on the complex-valued unit hypersphere and were modeled by a complex Watson Mixture Model (cWMM). A VEM algorithm was derived which estimated the posterior distributions of the Watson mode vectors, which can be used for beamforming and source separation. Furthermore, the number of active speakers was determined by a step-wise source counting algorithm which is reminiscent of successive interference cancellation in multi-user digital communications: after identifying a source, its contribution to the microphone signal was deemphasized before searching for the next source. This process was repeated until a maximum number of potential sources was reached. The final number of sources was then determined by thresholding the mixture weights and concentration parameters of the cWMM and enforcing a minimum angular distance between sources.

In this contribution we present an extension of this work with respect to the following aspects: First, an improved criterion is presented for emphasizing time-frequency bins that are most likely dominated by a single source. This results in increased robustness to additive noise compared to [6]. Furthermore, we show that reliable source counting can be done by either thresholding the mixture weights or the concentration parameters, with no need to require both. The experiments demonstrate that this approach outperforms the DOA-based VEM on GMMs. Finally, we propose an algorithm for online source counting, which allows the estimation of the number of simultaneously active sources on a per-frame basis with low latency.

The paper is organized as follows. In the next section we

The work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/8-1.

describe our signal model, followed by a description of the cWMM in Section 3. The offline and online source counting algorithms are presented in Sections 4 and 5, respectively. Experiments are reported in Sections 6 and 7 and conclusions are drawn in Section 8.

2. SIGNAL MODEL

Consider a convolutive mixture model of K independent source signals $S_k(t,f)$ captured by D microphones yielding the sensor signals $X_d(t,f)$ in STFT domain [7]:

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

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 t from 1 to T and frequency bins f from 1 to F.

According to the normalized observation vector approach, vectors $\mathbf{X}$ are phase normalized, frequency normalized and unit-norm normalized selecting an arbitrary sensor signal as reference, e.g., $X_1(t,f)$ [8]:

$$\begin{aligned} \tilde{X}_d(t,f) &= |X_d(t,f)| \exp\left(\mathrm{j}\frac{\arg\left(X_d(t,f)X_1^*(t,f)\right)}{2 f_{\mathrm{real}} c^{-1} d_{\max}}\right), \\ \mathbf{Y}(t,f) &= \tilde{\mathbf{X}}(t,f)/\|\tilde{\mathbf{X}}(t,f)\|, \end{aligned} \tag{2}$$

where $f_{\mathrm{real}} = f/(LT_{\mathrm{s}})$ is the frequency corresponding to frequency index f, FFT-length L and sampling rate $1/T_{\mathrm{s}}$. The constant c is the speed of sound and $d_{\max}$ is the maximum distance between the sensors. The number of used frequencies F is set to $L/2 - 1$.

The frequency normalization allows to increase the number of observations per estimation step. Rather than treating each frequency bin separately, all frequency bins can be used jointly in the algorithm. This is particularly beneficial for online source counting, since now even in a single time frame there are enough observations to carry out a block EM algorithm.

Note that Equation (2) maintains component-wise amplitude differences and the frequency normalization assumes a linear frequency dependency for $\mathbf{H}(f)$.

3. STATISTICAL MODEL

Since the extracted features are vectors on the complex D-dimensional unit hypersphere, a probability density function able to model clusters in this domain is desirable. Tran Vu *et al.* proposed to model these observations by a cWMM [1]:

$$p(\mathcal{Y}|\mathcal{C}) = \prod_{t=1}^{T} \prod_{f=1}^{F} \prod_{k=1}^{K+1} \left(\frac{1}{c_{\mathrm{W}}(\kappa_k)} \mathrm{e}^{\kappa_k \left|\mathbf{W}_k^{\mathrm{H}} \mathbf{Y}(t,f)\right|^2}\right)^{c_k(t,f)}. \tag{3}$$

Implying sparseness of the source signals $c_{k=l}(t,f) = 1$ and $c_{k \neq l}(t,f) = 0$ indicates that the l-th speaker is dominant in the given tf slot (t,f). The additional component $K+1$ models noise only slots.

In (3), $c_{\mathrm{W}}(\kappa_k)$ is a normalization constant given by

$$c_{\mathrm{W}}^{-1}(\kappa_k) = \frac{(D-1)!}{2\pi^D M(1, D, \kappa_k)}, \tag{4}$$

where $M(\,\cdot\,)$ is the confluent hypergeometric function [9].

The choice of a cWMM can be justified as follows: Firstly, conventional approaches reduce the feature space [5] or approximate distributions on the complex hypersphere by a GMM on a planar projection [2]. Using a cWMM maintains all spatial information contained in the signal. Secondly, an a priori distribution for the complex Watson mode vectors is known. Thirdly, the distance measure $\mathbf{W}^{\mathrm{H}}\mathbf{Y}$ resembles a spatial correlation and, thus, fits to the beamforming concept [1].

4. SOURCE COUNTING IN AN OFFLINE SCENARIO

At first, tf slots which are most likely dominated by a single speaker and are thus well suited for source counting are emphasized. These slots are identified by a power based criterion similar to the one in [3]. We employ a quantile instead of applying a power threshold to obtain a fixed number of observations independent of the signals themselves:

$$P = \text{quantile}(\mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f), q), \tag{5}$$

where $q \in [0,1]$. Subsequently, observation weights are defined:

$$A^{(1)}(t,f) = \begin{cases} \frac{1}{2} + \frac{1}{2}a, & \mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f) > P, \\ \frac{1}{2} - \frac{1}{2}a, & \mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f) < P, \end{cases} \tag{6}$$

where $a \in [0,1[$ controls how much tf slots, which are considered to be dominated by a single speaker, are emphasized over others.

EM algorithms are known to be very sensitive with respect to initial values. To relax this sensitivity, the mode vector $\mathbf{W}$ of the most dominant source is found by employing a VEM algorithm for a cWMM with two mixture components. One mixture component is intended to model the dominant speaker while the second component is fixed to a uniform distribution on the complex hypershere and captures noise and all remaining speakers. Motivated by the use of uninformative priors for model complexity estimation for GMMs in [10] a uninformative Bingham prior is used for the speaker component.

The update equations for the cWMM are given in brief. A more detailed derivation can be found in [6]. The observation weights $A^{(\nu)}(t,f)$ are incorporated into the VEM heuristically. At first the class responsibilities for the dominant speaker and the noise are updated:

$$\begin{aligned}\ln \gamma_1(t,f) &= \ln A^{(\nu)}(t,f) - \ln M(1,D,\kappa_1)\\ &\quad + \kappa_1\, \mathbb{E}_{\mathbf{W}_1}\left\{\mathbf{W}_1^{\mathrm{H}}\mathbf{Y}(t,f)\mathbf{Y}^{\mathrm{H}}(t,f)\mathbf{W}_1\right\}\\ &\quad + \mathbb{E}_{\pi_1}\left\{\ln \pi_1\right\} + \text{const.} \qquad (7)\\ \ln \gamma_2(t,f) &= \ln(1 - A^{(\nu)}(t,f)) - \ln M(1,D,0)\\ &\quad + \mathbb{E}_{\pi_2}\left\{\ln \pi_2\right\} + \text{const.}\end{aligned}$$

Then the distributions are refined by

$$\mathbf{B}_k^{(i)} = \kappa_k^{(i-1)} N_k^{(i)} \mathbf{\Phi}_{YY,k}^{(i)} + \mathbf{B}_{0,k}, \qquad (8)$$

$$\mathbf{\Phi}_{YY,k}^{(i)} = \frac{1}{N_k^{(i)}} \sum_{t=1}^{T}\sum_{f=1}^{F} \gamma_k^{(i)}(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\mathrm{H}}(t,f), \qquad (9)$$

$$\pi_k^{(i)} = N_k^{(i)} / \sum_{k=1}^{K+1} N_k^{(i)} \text{with } N_k^{(i)} = \sum_{t=1}^{T}\sum_{f=1}^{F} \gamma_k^{(i)}(t,f), \quad (10)$$

$$\frac{M(2,D+1,\kappa_k^{(i)})}{D \cdot M(1,D,\kappa_k^{(i)})} = \mathbb{E}_{\mathbf{W}_k}\left\{\mathbf{W}_k^{\mathrm{H}} \mathbf{\Phi}_{YY,k}^{(i)} \mathbf{W}_k\right\}. \qquad (11)$$

After successfully estimating the parameters of the cWMM for the first speaker, the mode vector $\mathbf{W}_1$ for this speaker is calculated as the principal component of the parameter matrix $\mathbf{B}_1$ of the a posteriori Bingham distribution.

Inspired by the successive interference cancellation method used in multi-user digital communications the influence of the identified dominant source on the microphone signal is removed by reweighting each observation depending on its distance to the already found mode vector [11]:

$$A^{(\nu+1)}(t,f) = A^{(\nu)}(t,f)\left(1 - \mathrm{e}^{\kappa_{\mathrm{Re}}\left(\left|\hat{\mathbf{W}}_\nu^{\mathrm{H}}\mathbf{Y}(t,f)\right|^2 - 1\right)}\right), \quad (12)$$

where κ_{Re} governs the concentration of the reweighting function.

The searching and reweighting is then repeated until a maximum number of expected speakers $\nu_{\max}$ is reached. We now compute the angular distance between mode vectors and determine the number of active speakers by thresholding either the mixture weights π_ν or the concentration parameters κ_ν, while at the same time requiring that the scalar product $\mathbf{W}_\nu^{\mathrm{H}}\mathbf{W}_{\nu'}$ between mode vectors is below a threshold s_{Th}. This is equal to requiring a minimal angular distance between mode vectors.

5. SOURCE COUNTING IN AN ONLINE SCENARIO

In the online scenario we maintain a set $\mathcal{B}$ of Bingham matrices for already found but possibly inactive speakers, which is reviewed in every frame. At the beginning, the set is initialized with the empty set.

For each frame, at first, observation weights are calculated according to Equation (6). Secondly, if the set $\mathcal{B}$ is not empty, the main VEM algorithm updates each $\mathbf{B}_k$ in $\mathcal{B}$ and estimates $\kappa_k(t)$ and $\pi_k(t)$ for each frame. These estimates are thresholded for a frame based speaker activity decision similar as in the offline counting algorithm. Thirdly, the observation weights are reweighted according to Equation (12). Finally, a seeker VEM searches for possible new speakers in the remaining signal and adds the corresponding a posteriori Bingham matrix to the set $\mathcal{B}$, if the threshold conditions are fulfilled.

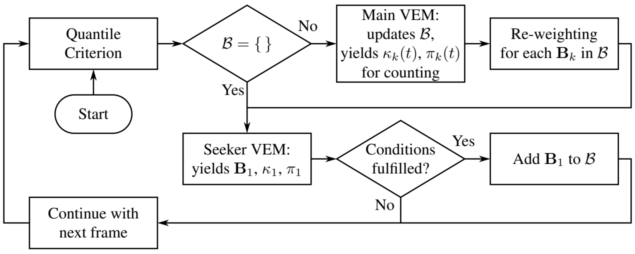

Fig. 1: Flowchart of the online algorithm.

To summarize, a decision is made on a frame-by-frame basis. The seeker VEM needs one frame to find a candidate for a new speaker. The main VEM needs another frame to validate the speaker, amounting to an overall latency of two frames.

6. RESULTS FOR THE OFFLINE SCENARIO

In a simulation environment, up to six speakers 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 sources and the sensor array share the same height of 1.5 m.

Speech samples of 5 s length and sampling frequency of 16 kHz are chosen at random from the training utterances of the TIMIT database [12]. Speech samples of zero to up to six speakers without speech pauses are convolved with impulse responses of a simulated room of dimension $4\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$ using the image method [13] and mixed.

An STFT with frame size 1024 and a frame shift of 256 is applied to each sensor signal. The maximum sensor distance is set to $d_{\max} = 1.2 \cdot 2\,\mathrm{cm}$ to increase the distance of periodic repetitions of clusters in the feature space. The emphasis parameter is set to $a = 0.6$ and the fraction of deemphasized observations is set to $q = 0.9$. The reweighting factor in equation (12) is set to $\kappa_{\mathrm{Re}} = 20$ and maximum number of expected speakers is $\nu_{\max} = 8$. The SNR for white Gaussian noise is varied from 10 to 30 dB.

The parameters and initial values for the DOA-based counting algorithm, which is given for comparison, are taken directly from the corresponding publication [5].

Figure 2 compares the proposed step-wise deletion algorithm, using a threshold on the mixture weights or a threshold on the concentration parameters, with the DOA-based algorithm [5]. The proposed algorithm performs arguably well for up to five speakers and high SNR values, whereas the DOA-

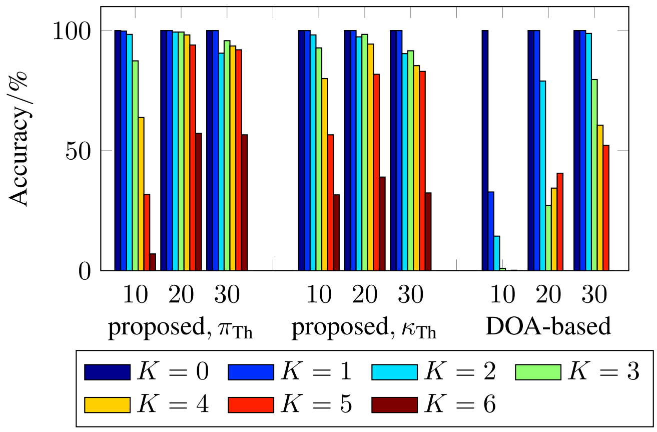

Fig. 2: Comparison of the proposed counting algorithm with the DOA-based algorithm with respect to different thresholds and SNR conditions.

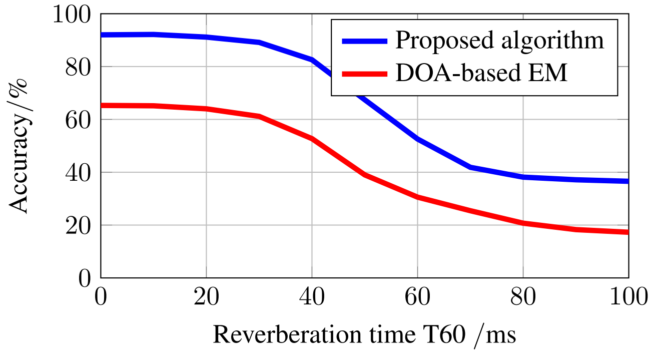

Fig. 3: Comparison of the proposed counting algorithm with the DOA-based algorithm with respect to reverberation time for SNR = 20 dB and $\pi_{\text{Th}} = 0.0033$, $s_{\text{Th}} = 0.65$.

based algorithm already shows limited performance for more than three speakers. Both algorithms struggle with an SNR of 10 dB. The DOA-based algorithm is hardly capable of finding more than one speaker at all.

The overall counting accuracy of the proposed algorithm is 84.0 % when thresholding the mixture weights and 83.5 % when thresholding the concentration parameters, while the DOA-based algorithm achieved 48.6 % accuracy.

Figure 3 shows the counting performance as a function of the room reverberation time. Due to the fact that both algorithms rely on the assumption of a linear phase of the impulse responses and both algorithms perform a frequency normalization, their counting accuracy declines already for small impulse response lengths. Working on each frequency bin independently may overcome this shortcoming, however at the price of introducing the permutation problem.

7. RESULTS FOR THE ONLINE SCENARIO

A ten second speech activity pattern for three speakers was generated containing different source activity situations, see Fig. 4. In contrast to the offline scenario, individual speakers stop and restart speaking. The thresholds are set to $\kappa_{\text{Th}} = 1$, $\pi_{\text{Th}} = 0.0025$ and $s_{\text{Th}} = 0.7$.

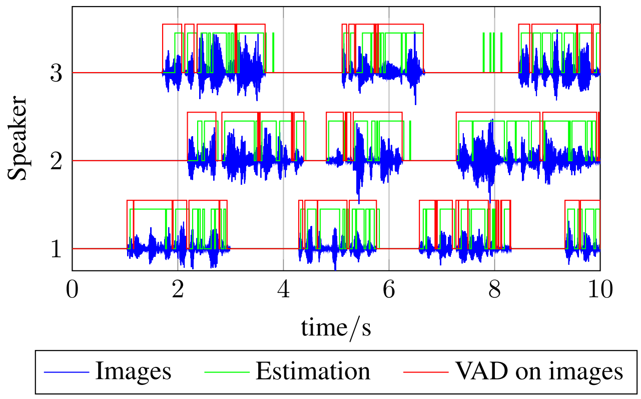

Fig. 4: Speaker activity estimation of the proposed online algorithm in comparison to the images and a VAD on images with an SNR of 20 dB.

As a reference, a voice activity detection (VAD) is calculated on the noise-free and unmixed speech signals to obtain ground truth speaker activity information.

The estimated active speakers are depicted in green. The first active speaker is detected on time. Subsequent new speaker onsets are detected with a small delay which increases with the number of simultaneously active speakers. Already identified speakers are found much faster. Note that the proposed algorithm is not only able to count the number of active speakers. It is also capable of rediscovering speakers who have been speaking earlier.

According to the NIST Rich Transcription 2007 evaluation the diarization error rate (DER) is defined as the ratio of the time a speaker is misidentified to the total duration of speech [14]. Since speech pauses occur, the DER can be above 100 %. The achieved DER in the given example is 30 %, whereas the worst possible DER for the given example is 187 %.

8. CONCLUSIONS

We have presented an active source counting algorithm based on a variational EM algorithm for complex Watson Mixture Models. A key component was the identification and emphasis of tf slots which are assumed to contain a single speaker and thus are appropriate for source counting. In tests in an offline scenario it was shown to outperform a DOA-based source counting algorithm. We then proposed an algorithm for online estimation of the number of active source, which has a latency of only two frames. Since the proposed algorithm estimates mode vectors which capture the spatial properties of a source, it is not only able to count the number of sources but to rediscover sources seen before. Furthermore, the mode vectors can be immediately used for beamforming and source separation.

9. REFERENCES

[1] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*. IEEE, 2010, pp. 241–244.

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

[3] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, 2008.

[4] 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.

[5] 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.

[6] 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 *Proc. ICASSP*, May 2014, p. in press.

[7] S. Makino, T.W. Lee, and H. Sawada, *Blind Speech Separation*, Signals and communication technology. Springer Science+Business Media B.V., 2007.

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Normalized observation vector clustering approach for sparse source separation," *Proc. EUSIPCO*, 2006.

[9] F. W. J. Olver and National Institute of Standards and Technology (U.S.), *NIST Handbook of Mathematical Functions*, Cambridge University Press, 2010.

[10] A. Corduneanu and Christopher M. Bishop, "Variational bayesian model selection for mixture distributions," in *Artificial intelligence and Statistics*. Morgan Kaufmann Waltham, MA, 2001, vol. 2001, pp. 27–34.

[11] D. H. Tran Vu and R. Haeb-Umbach, "On initial seed selection for frequency domain blind speech separation.," in *Proc. INTERSPEECH*, 2011, pp. 1757–1760.

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

[13] 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, 1979.

[14] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, pp. 373–389. Springer, 2008.