Acoustic Vector Sensor based Speech Source Separation with Mixed Gaussian-Laplacian Distributions

Xiaoyi Chen*†, Atiyeh Alinaghi†, Xionghu Zhong‡ and Wenwu Wang†
*Department of Acoustic Engineering, School of Marine Technology
Northwestern Polytechnical University, China, 710072
Email: xiaoyi.chen@surrey.ac.uk
†Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering
University of Surrey, UK, GU2 7XH
Emails: {w.wang and A.Alinaghi}@surrey.ac.uk
‡School of Computer Engineering, College of Engineering
Nanyang Technological University, Singapore, 639798
Email: xhzhong@ntu.edu.sg

***Abstract*—Acoustic vector sensor (AVS) based convolutive blind source separation problem has been recently addressed under the framework of probabilistic time-frequency (T-F) masking, where both the DOA and the mixing vector cues are modelled by Gaussian distributions. In this paper, we show that the distributions of these cues vary with room acoustics, such as reverberation. Motivated by this observation, we propose a mixed model of Laplacian and Gaussian distributions to provide a better fit for these cues. The parameters of the mixed model are estimated and refined iteratively by an expectation-maximization (EM) algorithm. Experiments performed on the speech mixtures in simulated room environments show that the mixed model offers an average of about 0.68 dB and 1.18 dB improvements in signal-to-distotion (SDR) over the Gaussian and Laplacian model, respectively.**

***Index Terms*—Acoustic vector sensor, mixed model, direction of arrival, EM algorithm, blind source separation.**

I. Introduction

Blind source separation (BSS) aims to estimate the individual sound signals in the presence of interferences. Direction of arrival (DOA) information was exploited in the source separation problem by using a beamforming technique [1]. However, the spatial selectivity of such beamformers is not sufficient for good separation due to the limited number of microphones, especially under the reverberant environment. Recently, a technique that relies on the use of a compact coincident microphone array, called acoustic vector sensor (AVS), was reported to provide highly accurate estimation of source directions [2], especially under 2-dimension geometries and small numbers of sensors. In [3], the DOA (based on AVS) and the mixing vector cues (as in [4]) were combined to achieve separation in reverberant environments, where these cues are modelled by (complex) Gaussian distributions. In anechoic situations, however, a Laplacian distribution has been shown to perform well for modelling the mixing vectors [5]. It can be noticed from these works that the distribution of both the DOA and mixing vectors varies from Laplacian to Gaussian depending on the level of reverberation, but it is not solely Laplacian or Gaussian under different environments.

In this paper, a weighted Gaussian-Laplacian distribution is proposed for statistically modelling the DOA and the mixing vector cues at each time-frequency (T-F) point of speech mixtures under both the anechoic and the reverberant environments. The weighted distribution is able to adapt to the room reverberation and to fit these two kinds of cues in a more accurate way. The model parameters and the assigned T-F regions of the speech mixtures are refined iteratively using the EM algorithm. In the E-step, the mixed probability distribution functions are applied to calculate the likelihood in each spectrogram point. In the M-step, the parameters of each source model are re-estimated according to the T-F regions of the mixtures that are most likely to be dominated by that source. It is noticed from [6] that the EM algorithm is sensitive to the initialization value because of the non-convex characteristics of the total log likelihood, so the accurate DOA information obtained by AVS which is used as the initialization value in the EM algorithm has the potential to improve the separation performance. Moreover, the mixed model proposed in this paper shows benefits for fit the data distribution in a more reliable way and hence achieves better performance under different room environments.

The remainder of this paper is organized as follows. Section II shows a brief summary of estimating the DOAs and the mixing vector from the mixtures that are acquired by an AVS in the T-F domain. The distribution of these two kinds of cues are tested under diverse situations in Section III. Section IV explains firstly the mixed model employed for the DOAs and the mixing vector, then, the EM algorithm to maximize the combined log likelihood and to estimate the mixed model parameters of these two cues. The experimental results which

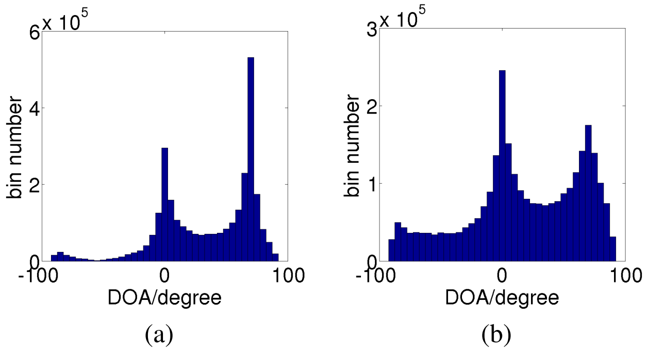

Fig. 1. The histogram of DOAs under (a) anechoic and (b) reverberant ($T_{60} = 0.3$ s) environments. Two speech sources are located at $0°$ and $70°$ respectively.

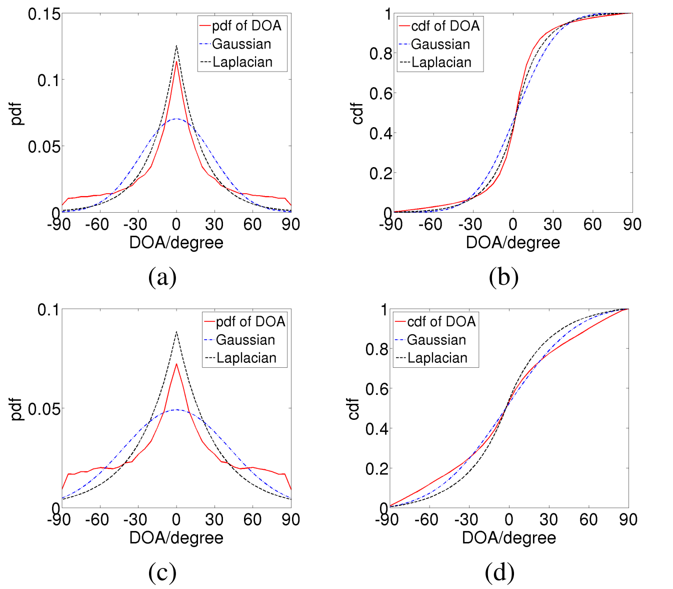

Fig. 2. The PDF (a,c) and CDF (b,d) of DOA distributions from a single source which is located at $0°$ and then modelled by the Gaussian and Laplacian distribution when $T_{60} = 0.3$ s (a) and 0.5 s (c).

include the comparisons with Gaussian- and Laplacian-model based methods respectively are shown in Section V, and finally Section VI gives the conclusions.

II. AVS Spatial Cues for Source Separation

As mentioned earlier, the DOA information of the sources which is carried by the data collected from the AVS can be employed to separate the speech signals [3]. Due to the sparse nature of an audio signal in the T-F domain, the clustering can also be performed on the mixtures using the information on the filter coefficients, which is known as the mixing vectors in BSS methods [4]. However, both the DOA and the mixing vector cues show increasing ambiguities as the increase of the reverberation. Therefore, a method which combines these two cues was proposed in [3] to improve the performance of source separation.

It was assumed in [3] that the sources and the sensors are strictly located at a 2-D $(x-y)$ space as they are all located at 0 degrees in elevation. Acoustic vector sensor is constructed by three microphones for the measurement of acoustic pressure and the calculation of pressure gradient. The received mixtures from the source signals $s_i(t)$, $i = 1, ...I$, in both anechoic and reverberant conditions can thus be expressed as

$$\begin{bmatrix} p_0(t) \\ p_x(t) \\ p_y(t) \end{bmatrix} = \sum_{i=1}^{I} \begin{bmatrix} h_0^i(t) \\ h_x^i(t) \\ h_y^i(t) \end{bmatrix} \otimes s_i(t) \tag{1}$$

where I is the number of sources, t is the discrete time index, $\otimes$ denotes convolution, and $p_0(t)$, $p_x(t)$ and $p_y(t)$ are the acoustic pressure signal received from the sensors located at the origin, x-coordinate and y-coordinate respectively. $h_0^i(t)$, $h_x^i(t)$ and $h_y^i(t)$ represent the corresponding room impulse response (RIR) from the ith source.

The pressure gradient can then be obtained from the acoustic pressure as

$$\mathbf{g}(t) = \begin{bmatrix} g_x(t) \\ g_y(t) \end{bmatrix} = \begin{bmatrix} p_x(t) - p_0(t) \\ p_y(t) - p_0(t) \end{bmatrix} \tag{2}$$

where $g_x(t)$ and $g_y(t)$ is the pressure gradient corresponding to the x- and y- coordinates, respectively. The general form of the speech mixtures at the output of a single AVS can thus be constructed as $[p_0(t), \mathbf{g}(t)^T]^T$.

The resulting direction can thus be obtained by

$$\theta(\omega, k) = \arctan\left[\frac{\Re\{P_0^*(\omega,k) G_y(\omega,k)\}}{\Re\{P_0^*(\omega,k) G_x(\omega,k)\}}\right] \tag{3}$$

where the superscript $*$ is the conjugate, $\Re\{\cdot\}$ means taking the real part of its argument, ω and k are the frequency bins and time frame indices, $P_0(\omega,k), G_x(\omega,k), G_y(\omega,k)$ are the short-time Fourier transforms (STFT) of $p_0(t), p_x(t), p_y(t)$ respectively.

Assuming the audio signals are sparse, which means only one source is dominant at each T-F unit. The STFT of the observations can be represented in a vector form as

$$\begin{aligned} \mathbf{m}(\omega,k) &= \sum_{i=1}^{I} \hat{\mathbf{h}}_i(\omega)\mathbf{s}_i(\omega,k) \\ &\approx \hat{\mathbf{h}}_{i^\star}(\omega)\mathbf{s}_{i^\star}(\omega,k), \forall i \in [1, \ldots, I] \end{aligned} \tag{4}$$

where $\mathbf{m}(\omega,k) = [G_x(\omega,k), G_y(\omega,k)]^T$, $\hat{\mathbf{h}}_i(\omega) = [H_x^i(\omega) - H_0^i(\omega), H_y^i(\omega) - H_0^i(\omega)]^T$, and $H_0^i(\omega), H_x^i(\omega), H_y^i(\omega)$ are the STFTs of the $h_0^i(t), h_x^i(t), h_y^i(t)$ respectively. The $i^\star$ is the index of the most dominant source for each T-F point. To avoid the effect of the source amplitude, the mixtures are normalized in the T-F domain to have a unit norm.

III. Distribution of the DOA and Mixing Vector Cues under Different Reverberation

In order to quantify the effects of reverberation on the distribution of DOA and mixing vector cues, the Kolmogorov-Smirnov (KS) distance test [7] is employed to compare the real distribution with both the Gaussian and Laplacian distributions, respectively. The KS distance is based on the

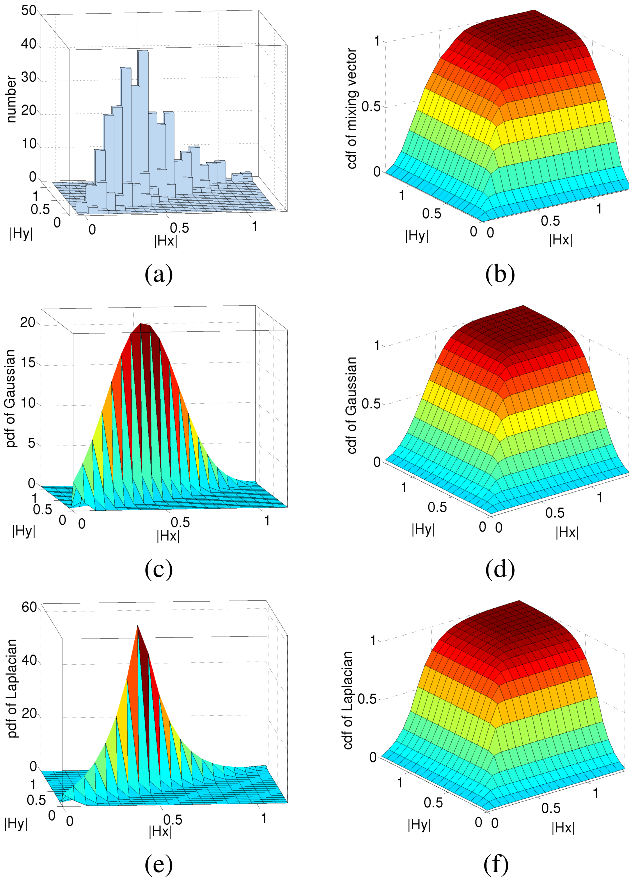

Fig. 3. The histogram (a) and CDF (b) of the mixing vector and then modelled by a multivariate Gaussian (c,d) and Laplacian distribution (e,f) when $T_{60} = 0.1$ s.

calculation of maximum distance T between the cumulative density function (CDF) of real data F_X and the proposed theoretical distribution F, which implies how close it is between the true value and the theoretical model.

$$T = max|F_X - F| \tag{5}$$

To evaluate the distribution of the DOA and the mixing vector cues in the T-F domain, the data of single source which is located at 0° is simulated using the same method as described in Section V, and these two cues are then tested under different reverberation conditions.

From eq. (3), the probability density function (PDF) of DOA cues can be obtained as the histogram of the $\theta(\omega, k)$, as shown in Fig. 1. With the assumption that the speech signals are sparse in the T-F domain, the directions estimated in each T-F point will correspond to a unique source. However, the DOA information will be blurred and the shape of DOA distribution will change due to the room reverberation. In Fig. 2, the PDF of DOA is estimated and then fitted by Gaussian and Laplacian distribution with the same mean and variance for $T_{60} = 0.3$ s and 0.5 s, respectively. The CDF is then calculated to estimate the KS distance with each distribution.

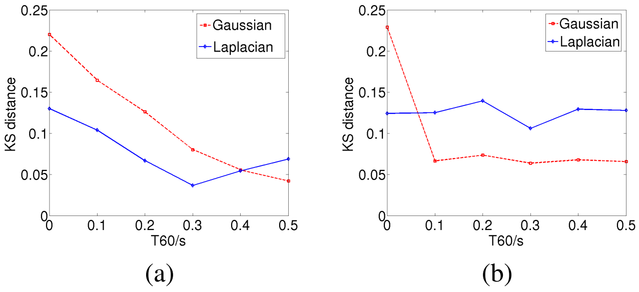

Fig. 4. KS distance test for the DOAs (a) and the mixing vectors (b) with the Gaussian and Laplacian distribution, respectively ($T_{60} = 0 : 0.5$ s).

Since the mixing vector is complex-valued in the T-F domain, in this section, the absolute values of the RIRs at $x-$ and $y-$ coordinates are used to show the distribution of the mixing vector. In Fig. 3, the histogram of the mixing vector (a) is estimated and then modelled by a multivariate Gaussian (c) and Laplacian distribution (e) when $T_{60} = 0.1$ s, (b, d, f) are the corresponding CDF of each distribution. After a paired-sample KS test [8], it is observed that $|H_x(\omega)|$ and $|H_y(\omega)|$ arise from the same distribution, therefore a one dimensional KS distance test is proposed for the mixing vector in the same way as for DOAs.

Fig. 4 shows the KS distance estimated from the two theoretical distributions for the direction information (a) and the mixing vectors (b) under various reverberation times. It can be observed that the distributions of both the DOAs and the mixing vector vary from Laplacian to Gaussian with the increase of reverberation, but the DOA distribution is much closer to the Laplacian distribution under anechoic and low reverberation environments, while the mixing vector is more Gaussian-like when the reverberation exists.

IV. Intensity cue and mixing vector modelled by mixed distribution with an EM algorithm

Based on the KS distance tests under various reverberant environments, a mixed Gaussian-Laplacian distribution is employed to model both the DOA and the mixing vector cues.

A. Mixed distribution model

In the case of I sources, the probability of DOAs in each T-F point for the ith source can be calculated by

$$\begin{aligned} &p(\theta(\omega,k)|\lambda_d, \mu_i(\omega), \delta_i^2(\omega)) \\ &\quad = \lambda_d \cdot f_G^r(\theta(\omega,k)) + (1-\lambda_d) \cdot f_L^r(\theta(\omega,k)) \end{aligned} \tag{6}$$

in which

$$\begin{aligned} f_G^r(\theta(\omega,k)|\mu_i(\omega), \delta_i^2(\omega)) &= \frac{1}{\sqrt{2\pi\delta_i^2(\omega)}} \\ \times \exp\left(-\frac{(\theta(\omega,k)-\mu_i(\omega))^2}{2\delta_i^2(\omega)}\right) \end{aligned} \tag{7}$$

$$\begin{aligned} f_L^r(\theta(\omega,k)|\mu_i(\omega), \delta_i^2(\omega)) &= \frac{1}{\sqrt{2\delta_i^2(\omega)}} \\ \times \exp\left(-\frac{|\theta(\omega,k)-\mu_i(\omega)|}{\sqrt{\delta_i^2(\omega)/2}}\right) \end{aligned} \tag{8}$$

where f_G^r and f_L^r are the PDFs of the real-valued Gaussian and Laplacian distributions, respectively. $\lambda_d \in [0,1]$ is a weighting parameter to control the contribution of each distribution in the mixed model. μ_i and δ_i^2 are the mean and variance of the DOAs of the i-th source component.

Since the mixtures are transformed to the T-F domain, as shown in eq. (4), the mixing vectors $\hat{\mathbf{h}}_i$ are modelled as a complex mixed probabilities function, and evaluated for each observation. In this mixed model, the complex Gaussian density function f_G^c is calculated in the same way as in [4], then, following the line orientation idea in [9], a complex Laplacian density function f_L^c is employed to form the complex mixed model.

$$\begin{aligned} &p(\mathbf{m}(\omega,k)|\lambda_m, \hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)) \\ &\quad = \lambda_m \cdot f_G^c(\mathbf{m}(\omega,k)) + (1-\lambda_m)\cdot f_L^c(\mathbf{m}(\omega,k)) \end{aligned} \tag{9}$$

in which

$$\begin{aligned} &f_G^c(\mathbf{m}(\omega,k)|\hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)) = \frac{1}{\left(\pi\gamma_i^2(\omega)\right)^2} \\ &\times \exp\left(-\frac{||\mathbf{m}(\omega,k) - (\hat{\mathbf{h}}_i^H(\omega)\mathbf{m}(\omega,k))\hat{\mathbf{h}}_i(\omega)||^2}{\gamma_i^2(\omega)}\right) \end{aligned} \tag{10}$$

$$\begin{aligned} &f_L^c(\mathbf{m}(\omega,k)|\hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)) = \frac{1}{\gamma_i^2(\omega)} \\ &\times \exp\left(-\frac{||\mathbf{m}(\omega,k) - (\hat{\mathbf{h}}_i^H(\omega)\mathbf{m}(\omega,k))\hat{\mathbf{h}}_i(\omega)||}{\gamma_i(\omega)}\right) \end{aligned} \tag{11}$$

where $\hat{\mathbf{h}}_i$ and γ_i^2 are the mean and variance of the mixing vector of the i-th component. $\lambda_m \in [0,1]$ is a weighting parameter. When $\lambda_d = 1$ and $\lambda_m = 1$, the mixed model reduces to the traditional Gaussian mixture model as proposed in [3].

As mentioned in [10], the permutation ambiguity problems in bin-wise classification approach should be solved before estimating the overall probability:

$$L(\hat{\Theta}) = \max_{\Theta} \sum_{\omega,k} \log(p(\theta(\omega,k), \mathbf{m}(\omega,k)|\Theta) \tag{12}$$

The whole parameter set Θ is given by

$$\Theta = \{\psi_i(\omega), \lambda_d, \mu_i(\omega), \delta_i^2(\omega), \lambda_m, \hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)\} \tag{13}$$

B. EM algorithm

The EM algorithm operates iteratively, and at each iteration, the optimal parameters which increase locally the log-likelihood of the mixture are computed. Since the EM algorithm can start with the E-step or the M-step, this means it can be initialized with data from either the mixing vector or the T-F masks [10]. However, the information of the mixing filters usually cannot be estimated directly. Similar to [10], we propose to initialize the masks firstly, then, estimate the initial values of $\hat{\mathbf{h}}_i(\omega)$ and $\gamma_i(\omega)$ from the masked spectrogram. In order to initialize the masks properly, in the first iteration, we use the histogram of DOAs to initialize the parameters and to estimate the masks, and let the program run without the BSS contribution.

In the E-step, the occupation likelihood $\nu_i(\omega,k)$ is computed from the observations and $\hat{\Theta}$ which are estimated at the M-step. The probability at each T-F unit which is dominated by the source i at DOA θ with the mixed model in eq. (6) and eq. (9) is calculated as

$$\begin{aligned} \nu_i(\omega,k) \propto \psi_i(\omega)&\mathcal{M}(\theta(\omega,k)|\lambda_d, \mu_i(\omega), \sigma_i^2(\omega)) \\ &\times\mathcal{M}(\mathbf{m}(\omega,k)|\lambda_m, \hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)) \end{aligned} \tag{14}$$

where $\mathcal{M}$ denotes the mixed probability function.

In the M-step, the DOA parameters are re-estimated for each source using the observations and the expectation value $\nu_i(\omega,k)$. As mentioned earlier, we set $\mathcal{M}(\mathbf{m}(\omega,k)|\hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega)) = 1$ at the first iteration to remove the effect of the BSS contribution.

After one iteration, the mask $M_i(\omega,k) \equiv \nu_i$ is obtained based on only the information of DOA cues and then the parameters of the mixing vectors, $(\hat{\mathbf{h}}_i(\omega), \gamma_i^2(\omega))$ can be estimated from the next M-step without the permutation problem [4].

$$\mathbf{R}_i(\omega) = \sum_k \nu_i(\omega,k)\mathbf{m}(\omega,k)\mathbf{m}^H(\omega,k) \tag{15}$$

$$\gamma_i^2(\omega) = \frac{\sum_k \nu_i(\omega,k)||\mathbf{m}(\omega,k) - (\hat{\mathbf{h}}_i^H(\omega)\mathbf{m}(\omega,k))\hat{\mathbf{h}}_i(\omega)||^2}{\sum_k \nu_i(\omega,k)} \tag{16}$$

$$\psi_i(\omega) = \frac{1}{T}\sum_k \nu_i(\omega,k) \tag{17}$$

where T is the number of the time frames. The $\hat{\mathbf{h}}_i$ is optimized as the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_i$. The sources are finally reconstructed by using $M_i(\omega,k)$ and $\mathbf{m}(\omega,k)$ after the convergence of the EM algorithm.

V. Experiments and results

The proposed method is tested for two signals collected with a single AVS under various simulated room environments. Similar to [3], a shoe-box room with a dimension of $9\times5\times3$ m^3 is employed. The AVS is located at the center of the room with the same height (1.5 m) as the two speech sources. The microphones of AVS at $x-$ and $y-$ coordinates are 0.5 cm away from the one at the origin. 15 utterances with a length of 3 s are randomly chosen from the TIMIT dataset and then shortened to 2.5 s in order to avoid the silence at the end. Moreover, all the speech signals are normalized before convolving with the RIRs which are simulated by using the imaging method [11] with different reverberation times. 15 pairs of mixtures were chosen randomly from the 15 utterances. In each experimental condition, the target source was located at 0° and the interference signal at 70°, both of which are located at 1 m from the microphones.

At each reverberation time, the λ_d and λ_m are chosen as the values which achieve the best separation performance from 0

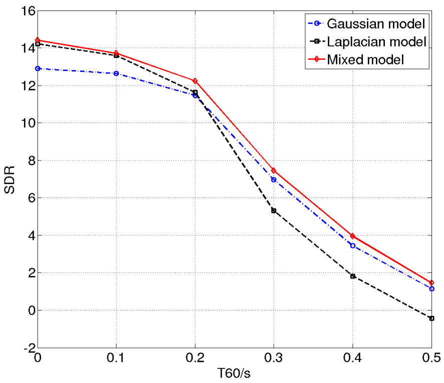

Fig. 5. The SDR comparison between the proposed method (based on the mixed Gaussian-Laplacian model) and the methods based respectively on the Gaussian and Laplacian models at different T_{60}s.

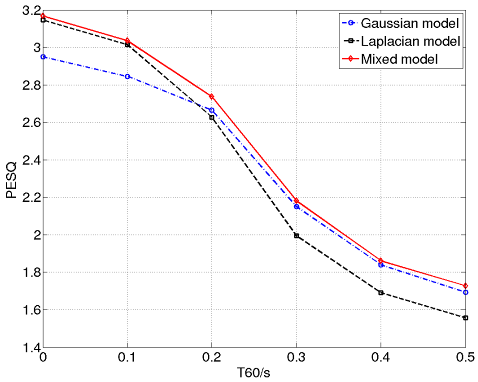

Fig. 6. The PESQ performance comparison between the proposed method (based on the mixed Gaussian-Laplacian model) and the methods based respectively on the Gaussian and Laplacian models at different T_{60}s.

to 1 with a step of 0.1, leading to 121 different combinations of λ_d and λ_m for each test, as shown in Table I. The separation performance was evaluated in terms of the signal-to-distortion ratio (SDR) [12] and perceptual evaluation of speech quality (PESQ) [13] which are averaged over all the 15 pairs of mixtures with T_{60}s from 0 s to 0.5 s with a step of 0.1 s. As shown in Fig. 5 and Fig. 6, the separation performance using the mixed model is consistently better than the baseline methods which only uses either the Gaussian or the Laplacian model. The improvement of the mixed model over the Gaussian model reduces when the reverberation becomes stronger, because in the high reverberation situation, both the DOA and the mixing vector cues show Gaussian-like distribution. The SDR results show an average of about 0.68 dB and 1.18 dB improvements, compared with the Gaussian- and Laplacian-model based algorithms, respectively. The PESQ improvements over the two baseline methods are both about 0.1.

VI. Conclusion

We have presented a weighted Gaussian-Laplacian distribution for modelling the spatial cues in probabilistic T-F masking for AVS based source separation. Simulation results show that the mixed model offers an improvement in SDR and PESQ over both the Gaussian and Laplacian based methods. It should be mentioned that the weighting parameter in this paper is chosen empirically, future work will concentrate on the method for calculating the weighting parameter in an analytical way.

TABLE I
Weighting parameters in the proposed method

$T60$	0.0	0.1	0.2	0.3	0.4	0.5
λ_d	0.1	0.5	0.7	0.8	0.9	0.9
λ_m	0.2	0.7	0.9	0.9	1.0	1.0

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council, mainly under the grant EP/I000992/1, and in part under the grants EP/H012842/1 and EP/H050000/1.

References

[1] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Transactions on, Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[2] B. Gunel, H. Hachabiboglu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Transactions on, Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 748–756, 2008.

[3] X. Zhong, X. Chen, W. Wang, A. Alinaghi, and A. B. Premkumar, "Acoustic vector sensor based reverberant speech separation with probabilistic time-frequency masking," in *European Signal Processing Conference*, 2013, (submitted).

[4] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proceeding of the IEEE Workshop on, Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 139–142.

[5] N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using laplacian mixture models," *IEEE Transactions on, Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1818–1832, 2007.

[6] M. I. Mandel, R. J. Weiss, and D. 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.

[7] A. G. Glen, L. M. Leemis, and D. R. Barr, "Order statistics in goodness-of-fit testing," *IEEE Transactions on, Reliability*, vol. 50, no. 2, pp. 209–213, 2001.

[8] J. Peacock, "Two-dimensional goodness-of-fit testing in astronomy," *Monthly Notices of the Royal Astronomical Society*, vol. 202, pp. 615–627, 1983.

[9] P. D O'Grady and B. A. Pearlmutter, "The lost algorithm: finding lines and separating speech mixtures," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. Article ID 784 296, 17 pp, 2008.

[10] A. Alinaghi, W. Wang, and P. J. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proceeding of the IEEE International Conference on, Acoustics, Speech and Signal Processing*. IEEE, 2011, pp. 209–212.

[11] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.

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

[13] L. D. Persia, D. Milone, H. Rufiner, and M.Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Process.*, vol. 88, no. 10, pp. 2578–2583, Oct. 2008.