UNDERDETERMINED SOURCE SEPARATION USING TIME-FREQUENCY MASKS AND AN ADAPTIVE COMBINED GAUSSIAN-STUDENT'S T PROBABILISTIC MODEL

Yang Sun, Waqas Rafique, Jonathon A. Chambers, Syed Mohsen Naqvi

Communications, Sensors, Signal and Information Processing Group,
School of Electrical and Electronics Engineering, Newcastle University, NE1 7RU, UK
{y.sun29, w.rafique2, jonathon.chambers, mohsen.naqvi}@newcastle.ac.uk

ABSTRACT

Time-frequency (T-F) masking algorithms are focused at separating multiple sound sources from binaural reverberant speech mixtures. The statistical modelling of binaural cues i.e. interaural phase difference (IPD) and interaural level difference (ILD) is a significant aspect of such algorithms. In this paper, a Gaussian-Student's t distribution combined mixture model is exploited for robust binaural speech separation. The weights of the distribution components are calculated adaptively with the energy of the speech mixtures. The expectation maximization (EM) algorithm is applied to calculate the parameters of the distributions. The speech signals from the TIMIT database are convolved with the real binaural room impulse responses (BRIRs) from two datasets for the evaluation of the proposed method. The objective performance measure signal to distortion ratio (SDR) confirms the improvement and robustness of the proposed method.

Index Terms— Source separation, Gaussian-Student's t combined mixture model, adaptive weights, real binaural room impulse responses

1. INTRODUCTION

The blind source separation (BSS) problem has drawn much attention from researchers during the past few decades and a robust solution for convolutive BSS (CBSS) for moving sources and the underdetermined case is still required [1] [2] [3] [4]. The well known statistical signal processing methods such as ICA [1] and IVA [5] are valid only for the exactly-determined (number of sources is equal to the number of sensors) and over-determined (number of sources is less than the number of sensors) cases [6]. We humans have the ability to separate up to six sources with only two ears i.e. solve the under-determined (number of sources greater than the number of sensors) case.

For machine learning, the T-F based methods are introduced under the framework of computational auditory scene analysis (CASA) e.g. the model-based expectation maximization source separation and localization (MESSL) algorithm [7] to mimic human auditory perception. The MESSL algorithm is based on the assumption of W-disjoint orthogonality [8], which implies that in the spectrogram, at most one source is active at each T-F point. On the basis of this result, non-linear modelling techniques such as those used to form T-F masks can be used to separate speech mixtures. In MESSL, the binaural cues are modelled by the Gaussian distribution mixture model (GMM). The EM algorithm can be used to determine the model parameters for the best fitting regions. Then, the probabilistic T-F masks for each of the sources are generated to separate the mixtures. However, within mixture distributions, the tails of the distributions also contain significant information [9]. In our previous work, the Student's t-distribution has been exploited to replace the Gaussian distribution to model the mixture and obtain more information from outliers [9]. Hence, the Student's t-distribution mixture model (SMM) was used to increase the robustness of separation performance. In the Student's t-distribution, the degree of freedom ν can control the tails; when ν goes to infinity, the distribution tends to be Gaussian [10]. However, once the mixture distribution is more Gaussian, the performance with SMM may decrease. Hence, using a single type of statistical distribution model may not be an accurate model.

In this paper, a combined approach is introduced to model both the IPD and ILD. The mixture distribution is jointly modelled with the GMM and SMM. The tails of the distribution are better modelled by the SMM whereas the lower amplitude information by the GMM. The proposed method is evaluated with real binaural room impulse responses (BRIRs) [11] [12]. The experimental results confirm the improvement and the robustness of the proposed approach.

The paper is organized as follows, in Section 2, the MESSL algorithm with the Student's t-distribution is described. In Section 3, the MESSL algorithm with the proposed combined model and adaptive weights are explained; experimental results are shown in Section 4. Finally, conclusions are drawn in Section 5.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307 and the MOD University Defence Research Collaboration in Signal Processing.

2. MESSL WITH STUDENT'S T-DISTRIBUTION

Assume the speech source is $s(t)$, and $l(t)$ and $r(t)$ are signals acquired by the left and right microphones, respectively. According to [7], a noise process is convolutive in the time domain, therefore, the signals are represented as $l(t) = s(t - \tau_l) * h_l(t) * n_l(t)$ and $r(t) = s(t - \tau_r) * h_r(t) * n_r(t)$, where $h_l(t)$ and $h_r(t)$ are the impulse responses of the left and right channels, $n(t)$ is the noise and τ is the time delay.

Therefore, in the frequency domain, the Fourier transform of the left and right channels are:

$$L(\omega, t) = |S(\omega, t)|e^{-j\omega\tau_l}\mathcal{F}\{h_l(t)\}\mathcal{F}\{n_l(t)\} \quad (1)$$

$$R(\omega, t) = |S(\omega, t)|e^{-j\omega\tau_r}\mathcal{F}\{h_r(t)\}\mathcal{F}\{n_r(t)\} \quad (2)$$

The ratio of $L(\omega, t)$ and $R(\omega, t)$ is the interaural spectrogram:

$$\frac{L(\omega, t)}{R(\omega, t)} = e^{-j\omega(\tau_l - \tau_r)}H(\omega)N(\omega, t) \quad (3)$$

where $N(\omega, t) = N_l(\omega, t)/N_r(\omega, t)$ represents the Fourier transform of the noise and $H(\omega) = \mathcal{F}\{h_l(t)\}/\mathcal{F}\{h_r(t)\}$ is the ratio of Fourier transforms of the impulse responses [7].

The interaural spectrogram is parametrized by the IPD $\phi(\omega, t)$ and ILD $\alpha(\omega, t)$ measured in dB. Because of the ambiguities and phase circularity problem, the IPD of the observation is defined as the phase residual $\hat{\phi}(\omega, t; \tau)$ [13]:

$$\hat{\phi}(\omega, t; \tau) = \arg(e^{j\phi(\omega,t)}e^{-j\omega\tau(\omega)}) \quad (4)$$

Both the IPD residual and ILD can be modelled approximately by Gaussian distributions:

$$p(\hat{\phi}(\omega, t) \mid \Theta_{Gp}) = \mathcal{N}(\hat{\phi}(\omega, t) \mid \xi(\omega), \sigma^2(\omega)) \quad (5)$$

$$p(\alpha(\omega, t) \mid \Theta_{Gl}) = \mathcal{N}(\alpha(\omega, t) \mid \mu(\omega), \eta^2(\omega)) \quad (6)$$

where $\Theta_{Gp} \equiv \{\xi(\omega), \sigma^2(\omega)\}$ and $\Theta_{Gl} \equiv \{\mu(\omega), \eta^2(\omega)\}$ are the parameter sets of the Gaussian models of IPD and ILD. The joint distribution of IPD and ILD is:

$$p(\phi(\omega, t), \alpha(\omega, t)|\Theta_G) = p(\hat{\phi}(\omega, t)|\Theta_{Gp}) \cdot p(\alpha(\omega, t)|\Theta_{Gl}) \quad (7)$$

where Θ_G is the entire parameter set for the joint Gaussian distribution.

By independently modelling IPD and ILD with Student's t distributions:

$$p(\hat{\phi}(\omega, t)|\Theta_{Sp}) = St(\hat{\phi}(\omega, t)|\Theta_{Sp})$$

$$= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda_p(\omega)}{\pi\nu}\right)^{\frac{1}{2}}\left(1 + \frac{\lambda_p(\omega)(\hat{\phi}(\omega, t) - \xi(\omega))^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (8)$$

and assuming the same degree of freedom ν for the IPD and ILD:

$$p(\alpha(\omega, t)|\Theta_{Sl}) = St(\alpha(\omega, t)|\Theta_{Sl})$$

$$= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda_l(\omega)}{\pi\nu}\right)^{\frac{1}{2}}\left(1 + \frac{\lambda_l(\omega)(\alpha(\omega, t) - \mu(\omega))^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (9)$$

where $\Theta_{Sp} \equiv \{\xi(\omega), \lambda_p, \nu\}$ and $\Theta_{Sl} \equiv \{\mu(\omega), \lambda_l, \nu\}$ represent the parameter sets for the Student's t distribution.

3. MESSL WITH COMBINED MODEL

3.1. MESSL with Combined Model

Since the shape of the distribution of the mixture is not fixed, we propose a combined probabilistic model of Student's t distribution and Gaussian distribution in this paper. Assume the weight of the SMM is k, where $k\epsilon[0, 1]$, is a weighting parameter to modify the contribution of each distribution in the combined model. The sum of the weighting parameters is one. Therefore, the new convex mixed distribution is:

$$k{\cdot}St(\hat{\phi}(\omega, t), \alpha(\omega, t)|\Theta_S)$$

$$+(1 - k){\cdot}\mathcal{N}(\hat{\phi}(\omega, t), \alpha(\omega, t)|\Theta_G) \quad (10)$$

where Θ_S denotes the set of all the parameters in ILD and IPD in the SMM, and includes the degree of freedom ν to control the shape of the Student's t-distribution. The weight k is a variable to find the best fit for the combined model, thereby better modelling the T-F points. By adapting the value of k, the shape of distributions can be changed.

From (10), the likelihood function of the observation with the combined model is represented as:

$$k \cdot \mathcal{L}(\Theta)_{St} + (1 - k) \cdot \mathcal{L}(\Theta)_G$$

$$= k \cdot \sum_{\omega,t} \log \sum_{i,\tau} St(\hat{\phi}(\omega, t)|\Theta_{Sp})St(\alpha(\omega, t)|\Theta_{Sl})$$

$$+(1-k){\cdot}\sum_{\omega,t} \log \sum_{i,\tau} \mathcal{N}(\hat{\phi}(\omega, t)|\Theta_{Gp})\mathcal{N}(\alpha(\omega, t)|\Theta_{Gl}) \quad (11)$$

The terms $\mathcal{L}(\Theta)_{St}$ and $\mathcal{L}(\Theta)_G$ represent the log likelihood function with SMM and GMM, respectively. Each component has independent parameters to be initialized and updated with the EM algorithm. Once the weights for each component are calculated, the EM algorithm is applied to estimate the parameters [14]. In the parameter estimation, after a number of iterations of the E and M steps, (11) converges. The resulting parameters are exploited to compute the responsibilities, namely the possibility of the active point of source i with delay τ. By using the GMM, the responsibilities are determined as:

$$\nu_{i\tau}(\omega, t) \equiv \psi_{i\tau} \cdot \mathcal{N}(\widehat{\phi}(\omega, t; \tau)|\Theta_{Gp})$$

$$\cdot\mathcal{N}(\alpha(\omega, t)|\Theta_{Gl}) \quad (12)$$

Moreover, by using the SMM, the responsibilities are calculated as:

$$\kappa_{i\tau}(\omega, t) \equiv$$

$$\frac{\psi_{i\tau} \cdot St(\widehat{\phi}(\omega, t; \tau)|\Theta_{Sp}){\cdot}St(\alpha(\omega, t)|\Theta_{Sl})}{\sum_{i,\tau} \psi_{i\tau} \cdot St(\widehat{\phi}(\omega, t; \tau)|\Theta_{Sp}){\cdot}St(\alpha(\omega, t)|\Theta_{Sl})} \quad (13)$$

where $\psi_{i\tau}$ is the mixing coefficient.

According to (11), (12) and (13), the probabilistic masks for each of the sources by marginalizing over delay can be obtained as:

$$M_i(\omega,t) \equiv k \cdot \sum_{\tau} \kappa_{i\tau}(\omega,t) + (1-k) \cdot \sum_{\tau} \nu_{i\tau}(\omega,t) \quad (14)$$

Finally, the above probabilistic T-F masks are applied to separate the target speech signals from the mixtures. The motivation being that the mixture can be fully modelled by the proposed method.

3.2. Adaptive Value of Weight in Combined Model

However, the way to find the appropriate weights for each T-F component is a challenge. The frequency range of the observed human speech signals is from 0 to 4kHz and the low frequency regions will generally contain more energy than the high frequency regions. If either region contains more energy in the mixture then the distribution shape will generally have heavier tails.

To obtain a robust and accurate separation performance, the whole spectrogram is divided into two components, one is the low frequency part (0-2kHz) and other is the high frequency part (2-4kHz). In the low frequency part, the weights for the SMM are generally more than for the GMM.

In order to enhance the robustness in the separation performance, an adaptive process is introduced to determine the value of the weight according to the mixture energy. In each component, the weights for the SMM and GMM are assigned according to the ratio of the low frequency part's energy to the total energy. The weight of the SMM for the low frequency component is calculated:

$$k = \left[\frac{\sum_{\omega=1}^{\frac{T}{4}} |X(\omega)|^2}{\sum_{\omega=1}^{\frac{T}{2}} |X(\omega)|^2} \right] \leqslant 1 \quad (15)$$

where k is the weight coefficient, T is the length of the Fourier transform and $X(\omega)$ is the single frequency bin of the mixtures. Hence, the weights for the GMM in each component are $1-k$. Thus, (15) gives the relationship between the weight of the SMM and the energy of the low frequency part. The value of k is used to determined the weight coefficients for the GMM and the SMM in (14) to obtain accurate T-F masks to separate the target speech source from the mixtures.

4. EXPERIMENTAL RESULTS

In this section, the proposed method with the combined model is evaluated with two types of real BRIRs [11] [12]. In all the experiments, speech signals are randomly selected from whole of the TIMIT database [15] to generate the mixtures. Every speech signal is approximately 2.5 seconds long to avoid ending silence. From our previous research, it can be known that once the sources are physically close to each other, the separation performance will drop significantly [7]. In the experiments, the target source is directed in the front of the sensors. To confirm the proposed method is valid for these challenging cases, the azimuth of the interfering source is selected as $15\,^\circ$, $30\,^\circ$ and $45\,^\circ$ to set the physical separation as a variable. In the underdetermined case, the second interferer is located symmetrically with the same azimuths. The degree of freedom ν in the SMM of the combined component is selected as 4, which is an appropriate choice [16]. In the combined model, the weighting parameter is adaptive to the proportion of the energy of the sub spectrogram to total energy. The initial parameters are IPD and ILD frequency dependent to obtain the best separation performance in MESSL compared with other complexities [7]. The SDR is exploited to evaluate the separation performance objectively [17].

4.1. Experiments with real BRIRs from Shinn [11]

In BRIRs from Shinn [11], the binaural impulse responses are recorded in a real classroom where $RT60 \approx 565ms$. The sampling frequency is 8kHz and the room size is $9m \times 5m \times 3.5m$.

The separation performance at each azimuth is averaged over five pairs of mixtures to improve the reliability of results. The averaged SDR values are shown in Figure 1.

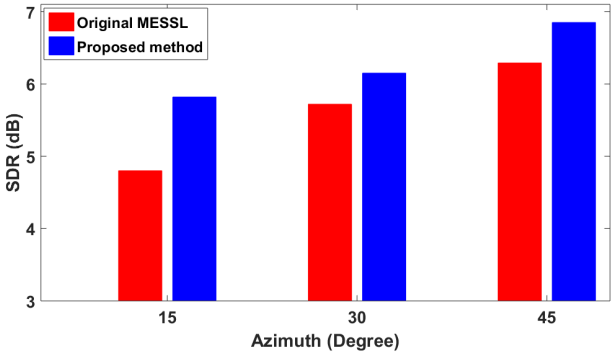

Fig. 1: Separation performance comparison in terms of averaged SDR (dB) over five different pairs of mixtures at each azimuths angle.

It is evident from Figure 1 that the proposed method improves separation performance particularly for the small azimuth. With the increase of azimuth, the original MESSL method also performs better, but even in that case, the proposed method has shown further performance improvement.

4.2. Experiments with real BRIRs from Hummersone [12]

In real BRIRs from Hummersone [12], there are four rooms with different reverberant environments named A, B, C and D. From [12], Room C has a higher direct to reverberant ratio (DRR). Therefore, we only use Room A, Room B and Room

D to compare the influence of $RT60$s on separation performance. Table 1 illustrates the parameters of these four rooms:

Table 1: Room settings for real BRIRs [12]

Room	Size	Dimension (m^3)	$RT60$ (s)
A	Medium	$5.7 \times 6.6 \times 2.3$	0.32
B	Small	$4.7 \times 4.7 \times 2.7$	0.47
C	Large	$23.5 \times 18.8 \times 4.6$	0.68
D	Medium	$8.0 \times 8.7 \times 4.3$	0.89

In the previous experiments with BRIRs from Shinn, the $RT60$ is fixed as $565ms$, whereas in these experiments, both the azimuth and $RT60$ are variable. The determined and underdetermined cases are evaluated with four rooms and three azimuths. The separation performance results shown below are the averaged value of five pairs of mixtures:

Table 2: Separation performance comparison in terms of averaged SDR (dB) for determined case with different types of rooms and azimuths.

Azimuth = 15 °	Room A	Room B	Room C	Room D
MESSL	5.25	5.08	5.99	3.54
Proposed	6.11	5.87	6.68	4.09
Improvement	16.4%	15.6%	11.5%	15.5%

Azimuth = 30 °	Room A	Room B	Room C	Room D
MESSL	9.37	7.71	9.54	6.08
Proposed	10.61	8.15	10.23	6.45
Improvement	13.2%	5.7%	7.2%	6.1%

Azimuth = 45 °	Room A	Room B	Room C	Room D
MESSL	10.91	8.09	10.85	7.16
Proposed	11.34	8.69	11.68	7.49
Improvement	4.0%	7.4%	7.6%	4.6%

Table 3: Separation performance comparison in terms of averaged SDR (dB) for underdetermined case with different types of rooms and azimuths.

Azimuth = 15 °	Room A	Room B	Room C	Room D
MESSL	0.2	1.97	1.62	0.13
Proposed	0.92	1.97	2.17	0.84
Improvement	360%	0%	34.0%	546.2%

Azimuth = 30 °	Room A	Room B	Room C	Room D
MESSL	4.82	5.22	5.75	3.69
Proposed	6.10	5.81	6.86	4.22
Improvement	26.6%	11.3%	19.3%	14.4%

Azimuth = 45 °	Room A	Room B	Room C	Room D
MESSL	7.24	6.32	8.42	5.68
Proposed	7.68	6.75	8.73	5.82
Improvement	6.1%	6.8%	3.7%	2.5%

It can be seen from Tables 2 & 3 that the increase of azimuth causes larger physical separation between the sources, which means the proportion of overlapping part in the mixture decreases and the separation performance is improved. With smaller azimuth case, the mixture distribution has heavier tails which can be better modelled by the combined model than just exploiting GMM or SMM. The results in Table 2 & 3 confirm the separation performance and robustness of the proposed method are improved. Besides, from comparing the separation performance of Rooms A, B and D in Tables 2 & 3, it is evident that the higher $RT60$ causes the mixture to be more complex, because the duration of the reverberation process becomes longer, more reflected signals are acquired in the mixtures. Hence, the value of SDR is decreased when the $RT60$ of environment is higher.

In the underdetermined case, an extra interfere is added into mixture. Table 3 shows that the proposed method is also efficient for solving underdetermined case than original MESSL method. However, the overall separation performance of the underdetermined case is less than the determined case. In the end, the separated speech signals are selected from the orignal MESSL and the proposed method and a listening test performed. The separated speech signals from the proposed method contain less noise from the interfering speech sources and are more clear than those from MESSL.

In the experiments, the proposed method is compared with the original MESSL method using two real BRIRs [11] [12]. These two BRIRs datasets provide three different $RT60$s and source location azimuths. The purpose of using BRIRs from Shinn is to find the relation between the separation performance and the position of the interference sources. In the experiments with [12], the azimuth, the number of interferes and $RT60$s are considered as variables. According to these experimental results, the proposed method outperforms the state of the art of the original MESSL algorithm.

5. RELATION TO PRIOR WORK AND CONCLUSIONS

In the MESSL algorithm, the mixture distribution is modelled by the GMM and the EM algorithm is used to calculate model parameters [7] [14]. However, modelling of the information in high amplitudes was still needed, which may not be accurately modelled by the Gaussian distribution. In this paper, Student's t and Gaussian distributions are combined together to model the mixture distribution. Besides, both GMM and SMM were assigned different weights in the sub spectrogram to fully model the mixture distribution. The weight parameters of each sub spectrogram are calculated adaptively by the proportion of the energy of low frequency part to the total energy. The parameters of the distributions are obtained after a number of E and M iterations and used to generate T-F masks to separate mixtures. The original MESSL and proposed method are evaluated with the TIMIT corpus [15] and real BRIRs [11] [12]. By comparing with experimental results from the original MESSL method, the separation performance and robustness of the proposed method are found to be improved.

6. REFERENCES

[1] A. Hyvarinen, and E. Oja, *Independent Component analysis*. Wiley, 2001.

[2] B. Rivet, W. Wang, S.M. Naqvi and J.A. Chambers, "Audiovisual Speech Source Separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 5, pp. 125-134, 2014.

[3] S.M. Naqvi, M. Yu and J.A. Chambers, "A Multimodal Approach to Blind Source Separation of Moving Sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895-910, 2010.

[4] M.S. Khan, S.M. Naqvi, A. Rehman ,W. Wang, and J.A. Chambers, "Video-Aided Model-Based Source Separation in Real Reverberant Rooms," *IEEE Transactions on Audio, Speech, Lang. Process*, vol. 21, pp. 1900-1912, 2013.

[5] T. Kim, H. Attias, S. Lee and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 70-79, 2007.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Processing and Speech Communication*, vol. 8, pp. 1-34, 2007.

[7] M.I. Mandel, R.J. Weiss and D.P.W. Ellis, "Model Based Expectation-Maximization Source Separation and Localization," *IEEE Transactions on Audio, Speech, Lang. Process*, vol. 18, pp. 382-394, 2010.

[8] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830-1847, 2004.

[9] Z.Y. Zohny, S.M. Naqvi and J.A. Chambers, "Enhancing MESSL Algorithm with Robust Clustering Based on Student's t-distribution," *Electronics Letters*, vol. 50, pp. 552-554, 2014.

[10] C.M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] B.S. Cunningham, N. Kopcp and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, pp. 3100-3115, 2005.

[12] C. Hummersone, "Binaural Room Impulse Response Measurements," *Surrey University*, 2011.

[13] M.I. Mandel, D.P.W. Ellis and T. Jebara, "An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments," in *Proc. Neural Information Processing System*, Canada, pp. 953-960, 2006.

[14] D. Peel, and G.J. Mclachlan, "Robust Mixture Modelling Using the T Distribution," *Statistics and Computing*, vol. 10, pp. 339-348, 2000.

[15] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, (Philadelphia), 1993.

[16] W. Rafique, S.M. Naqvi, P.J.B. Jackson and J.A. Chambers, "IVA Algorithm Using a Multivariate Student's t Source Prior for Speech Source Separation in Real Room Environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* , pp. 474-478, 2015.

[17] E. Vicent, R. Gribonval and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, pp. 1462-1469, 2006.