IJEW 国际工程期刊 International Journal of Engineering Works انجینئرنگ کاموں کا بین الاقوامی جرنل

International Journal of Engineering Works
ISSN-p: 2521-2419
ISSN-e: 2409-2770
Vol. 7, Issue 01, PP. 22-26, January 2020
https://www.ijew.io/
https://doi.org/10.34259/ijew.20.7012226

Speech Sources Separation Based on Models of Interaural Parameters and Spatial Properties of Room

Muhammad Israr[1], Muhammad Salman Khan[2], Khushal Khan[3]
[1,2,3] Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan
muhammad.israr@uetpeshawar.edu.pk[1]
Received: 07 January, Revised: 16 January, Accepted: 18 January

***Abstract*—** This paper presents the extended evaluation in different reverberant scenarios for different mixtures of speech having interfere at one of the six angles {15^0, 30^0, 45^0, 60^0, 75^0, 90^0 } of the model which implementing spatial covariance with interaural parameter for improving the performance of MESSL. The binaural spatial parameters, such as interaural phase difference IPD and interaural level difference ILD and spatial covariance are modeled in the short-time Fourier transform. The parameters of the model are updated with the expectation-maximization algorithm. The performance of the model is checked in term of Signal-to-distortion ratio (SDR) and the perceptual evaluation of speech quality (PESQ), and the results confirmed that the performance of this proposed model is improved in highly reverberant rooms.

***Keywords*—** MESSL, Spatial covariance, SDR, PESQ, MESSL+SC

I. INTRODUCTION

Humans are mostly interested to focus theiraattention only on the speech of a single speaker, while other speakers and back ground noise is exist [1]. This ability is greatly reduced when someone is listening with one ear, especially in reverberant environment. For example in automatic speech recognition the machine focus on a speaker of interest and separate speaker from background noise[2][3]. The process of automated separation of sources from measured mixtures without any prior information is known as blind source separation (BSS)[4]. In blind sources separation the sound sources are separated from the mixtures in which both the numbers of mixtures and sources are unknown and only mixtures signals are available. In some situations we want to recover all individual sources from the mixture signal, or at least a specific source[5]. In laboratory conditions most of the algorithms perform fine where number of sources present in mixtures, mixing methodology and acoustic condition are already known before the separation of sources from mixtures. But in real life scenario this problem is much more complex.

The solution for source separation problem is proposed by researchers belongs to different communities[6][7][8][9]. Examples included convolution in frequencyadomain, blindasource separation (BSS), beamforming and computational auditory scene analysis (CASA). Underdetermined problem in sound sources separation, where the numberaofasources is more than the number of sensors can be handle through TF approach[10]. In timeafrequency (TF) masking for source separation, by exploiting these cues, the mask for each source can be obtained and hoped that only a single sourceaisaactive at each TF unit. CASA is a "machine listening" system that separate mixtures of sound sources in the same way that human listeners do. It uses cues like pitch, onset/offsetatime, interauralaleveladifference (ILD) and interauralatime difference (ITD) to separate the sources[11].

In this thesis speech sources are separated by using a CASA approach which perform sources segregation on the base of interaural parameters such as interauralaleveladifference (ILD), interauralaphaseadifference (IPD) and spatial properties of the room. The modelaparameters are determined by using the EMaAlgorithm.

The observation and initial values of the parameters are used for estimation of probabilities in E-step. These parameters for the model are updatedabased on the measurementsaand the posterior probabilitiesafrom theaE-step. Theaproposed model produces probabilistic TFamasks for single sourcesathat are used for theirareconstruction.

II. MOTIVATION AND OBJECTIVE OF THE WORK

The main motivation in Blind Source Separation (BSS) is the cocktail party problem. Imagine someone is in a loud party, he is trying to convey his voice message to his friend. During this conversation the sounds received at the listener's ear is mixture of the voice of different musical instruments, people and so on. The signal arrived to the ear is in single wave form containing all sounds, and still listener is able to understand talk from this mixture[2][1].

For music students the sound source separation is an important tool by which he can separate a single instrument if many instruments are active simultiously. Similarlrly, the ability to remove an instrument from recording is also useful, as it would allow a student to play a different instrument in the orignal track in place of the removed instrument[12][13].

III. OVERVIEW OF PROPOSED ALGORITHM

Some statisticalaproperties of the sourcesaprovide a base for separation of sound sources[14]. The common statistical assumption made in our model are that the sources are

statisticallyaindependent, statisticallyaaorthogonala and nonstationarya[7][5][15].

The reverberant mixtures at left and right microphones are represented as $l(t) = \sum_{i=1}^{I} s_t\ (t) * h_{li}(t)$, and $r(t) = \sum_{i=1}^{I} s_t\ (t) * h_{ri}(t)$. In this expression s_i show sources, h_{li} represent the Roomaimpulsearesponse RIRafrom the source s_i to the left microphone, h_{ri} represent the Roomaimpulsearesponse RIRafrom the source s_i to the right sensor, and "*" show theaconvolution.

A. The ILD, IPD and spatial covariance models

The interauralaspectrogram, the ratio of leftaandaright mixture at every TF point is given by equation $\frac{L(t,\omega)}{R(t,\omega)} = 10^{\frac{\alpha(t,\omega)}{20}} e^{j\varphi(t,\omega)}$ where $\alpha(t,\omega)$ is the ILD measured in dB while $\varphi(t,\omega)$ isaIPD at frequencyaω and "t" is time. TheaIPD must exist in the range of $[-\pi,\pi)$ [16]. For ILD and IPD each source are frequency-dependent and is given as $\tau(\omega)$, and $\alpha(\omega)$ [15]. The model requires that Fourier transform window used, 1024 points (64ms) must be larger than maximum ITD of ≈0.75ms. We use a top down approach where ITD is measured in term of IPD. The phase residual error, which is defined as difference between IPD observed and IPD predicted is given by equation $\varphi(\omega,t,\tau) = \frac{L(t,\omega)}{R(t,\omega)} e^{-j\omega\tau}$. The phase residual error is modeled with a normaladistribution with frequency-dependentamean $\xi(\omega)$ andavariance $\sigma 2\ (\omega)$[17],

$$p\big(\varphi(\omega,t)\big|\tau\ (\omega),\sigma(\omega)\big) = N\ (\hat{\varphi}(\omega,t;\ \tau\ )|\xi(\omega),\sigma^{2}(\omega)).$$

The ILD is also modeled with a normal distribution with mean $\mu(\omega)$ and variance

$$\eta^{2}\ (\omega), p(\alpha(\omega,t)|\mu(\omega),\eta^{2}\ (\omega)) = N\ (\alpha(\omega,t)|\mu(\omega),\eta^{2}\ (\omega)).$$

Signal $x(\omega,t)$, is mixture signal consisting spatial images of I sources mixed with each other in present each channel. This can be easily modeled as a zero-mean Gaussian distribution with the covariance matrix [16].

$$R_x(\omega,t) = \sum_{i=1}^{I} v_i(\omega,t) R_i(\omega)$$

In above equation $v_i(\omega,t)$ show scalar variance while $R_i(\omega)$ is the covariance matrix containing the spatial properties of the source i. [18][19] The probability distribution of the proposed method is given by [16].

$$P(x(\omega,t)|\{vi(\omega,t), Ri(\omega), \forall\ i\}) = \frac{1}{det(\pi Rx(\omega,t))} exp(-x\ H(\omega,t) R^{-1}\ x\ (\omega,t) x(\omega,t))$$

Where $(\cdot)^{H}$ is the Hermitian transpose.

B. Expectation maximization and source separation

Theainterauralatimeadifference (ILD), interauralaphase difference (IPD) discussed in MESSL model [1] and the spatial covariance models, are combined [6] in this paper and new model has been formed which is combination of MESSL and Spatial covariance model. For determining the parameters of the models the expectationamaximization algorithm is used foraitsasolution. Their corresponding parametersaas

$$p(\alpha(\omega,t),\varphi(\omega,t),x(\omega,t)|\Theta) = N\ (\alpha(\omega,t)|\mu(\omega),\eta^{2}\ (\omega)) \cdot N\ (\hat{\varphi}(\omega,t)|\xi(\omega),\sigma^{2}\ (\omega)) \cdot N\ (x(\omega,t)|0,R_x(\omega,t))$$

All the model's parameter is given in a vector represented by Θ and its equation is given as

$$\Theta\ = \{\mu_i(\omega),\eta^{2}\ i\ (\omega),\xi_{\ i\tau}(\omega),\sigma^{2}{}_{i\tau}\ (\omega),v_i(\omega,t),\psi_{i\tau}\}$$

Where μ_i , $\xi_{i\tau}$, and $\eta_i{}^{2}$, $\sigma^{2}{}_{i\tau}$ are respectively the means and variancesaof theaILD, IPDamodels, and v_i is the scalaravariance related to the spatialacovariance model. The covarianceamatrix $R_i(\omega)$ required to calculate $R_x(\omega)$ is found by utilizing the posterior knowledge about the room properties. The notation i in subscript in all parameters represent that they belong toasource i, andaτ andaω show that these quantities depends on delay and frequency. Theaparameter $\psi_{i\tau}$ is the mixingaweight, is used for determining the probability of the point in TF spectrum belong to source i and the delay τ.[6]

The log likelihood function (L) given the observations can be written a

$$L(\Theta) = \sum_{t,w} \log p(\alpha(\omega,t),\varphi(\omega,t),x(\omega,t)|\Theta) = \sum_{t.w} log \sum_{i,t} [\ N\ (\alpha(\omega,t)|\mu_i(\omega),\eta^{2}{}_{i}\ (\omega)) \cdot N\ (\hat{\varphi}(\omega,t;\ \tau\ )|\xi_{i\tau}\ (\omega),\sigma^{2}{}_{i\tau}\ (\omega)) \cdot N\ (x(\omega,t)|0,R_x(\omega,t))\ \cdot\ \psi_{i\tau}\ ]$$

And the maximumalikelihood solution a vector of parameters moves the function toward maximum value. The EMaalgorithm is initializedawith the estimatedalocations of the speakers. In the expectationastep theaprobabilities are calculated given theaobservations and the estimatesaof the parameters as

$$\varepsilon_{i\tau}\ (\omega,t)\ =\ \psi_{i\tau}\ \cdot\ N\ (\alpha(\omega,t)|\mu_i(\omega),\eta^{2}{}_{i}\ (\omega)) \cdot N\ (\hat{\varphi}(\omega,t;\ \tau\ )|\xi_{i\tau}\ (\omega),\sigma^{2}{}_{i\tau}\ (\omega)) \cdot N\ (x(\omega,t)|0,R_x(\omega,t))),$$

Where $\varepsilon_{i\tau}\ (\omega,t)$ is a variable which provide the information of expectation belonging to a given source of hidden variable $m_{i\tau}\ (\omega,\ t)$, whose value is one in case time frequency point belong toasource i andadelay τ and zeroaotherwise. In M-step, the updating of parameters take place using previous observation of E-step. The IPDaandaILD parameters and $\psi_{i\tau}$ are re-estimated as in.

The contribution of spatialacovariance model are included from 2nd iteration, because in 1st iteration the $\varepsilon_{i\tau}$ is calculated only through ILDaandaIPD models, as it is dominate at the corresponding TF point for source i with delay τ. The time frequency mask for each source is given through equation.

$$M_i(\omega, t) = \sum_{\tau} \varepsilon_{i\tau}(\omega, \tau)$$

The masks are multiplied to mixtures and single source of interest are reconstructed. We next experimentally verify the efficiency of the proposed model through the MATLAB simulation.

IV. Experiments and Results

Experimentsawere performedaon mixtures of twoasources with varying levelsaofareverberation and differentaangels. The target source was present at 00 azimuth and interferer were present at one of the six angles {150, 300, 450, 600, 750, 900 }. Speechautterances were randomlyachosen from the TIMIT acousticaphonetic continuous speechacorpus to formamixtures with twoaspeech sources. The results obtained by performing simulation in different scenarios such as different angles, RIR and RT_{60}s for ten mixtures each containing two sources. The signalatoadistortion ratioa (SDR) and perceptualaevaluation of speechaquality (PESQ) were used to evaluateatheaperformance of the algorithms. The proposed model is named as (MESSL+ SC) in all graphs and figures.

The average SDR and average PESQ is calculated for ten different two-sources mixtures and for both proposed and MESSL models at different RT_{60}s i.e. 320ms, 470ms, 680ma and 890ms.

Figure. 1 and 2 show the average SDR and average PESQ comparisons of the proposed model with MESSL model at RT_{60}s of 320ms.

Text heads organize the topics on a relational, hierarchical

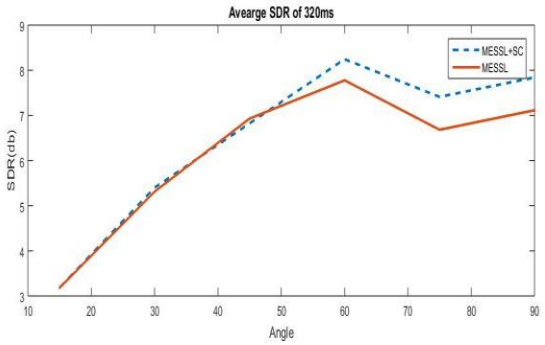

Figure.1 SDR Comparison of proposed model and MESSL Model at Reverberation of 320ms

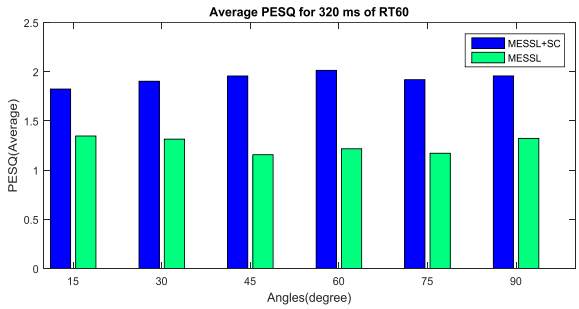

Figure. 2 PESQ Comparison of proposed model and MESSL Model at Reverberation of 320ms

The figure. 1 and 2 shows that at low reverberation the performance of proposed model and MESSL are same for average SDR and better than MESSL in the average PESQ.

Figure. 3 and 4 show the average SDR and average PESQ comparisons of the proposed model with MESSL model at RT_{60}s of 470ms.

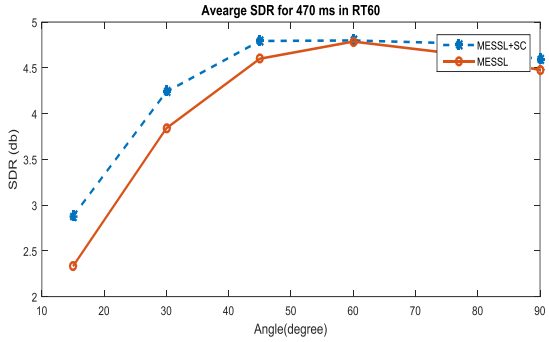

Figure. 3 SDR Comparison of proposed model and MESSL Model at Reverberation of 470ms

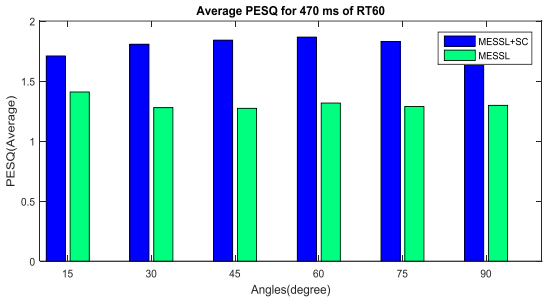

Figure. 4 PESQ Comparison of proposed model and MESSL Model at Reverberation of 470ms

The figure 3 and 4 clarifying that when reverberation increases from 320ms to 470ms the performance of the proposed model is enhanced from MESSL model in both evaluation matrix i.e. average SDR and average PESQ.

Figure.5 – 9 show the average SDR and average PESQ comparisons of the proposed model with MESSL model at RT_{60}s of 680ms and 890ms.

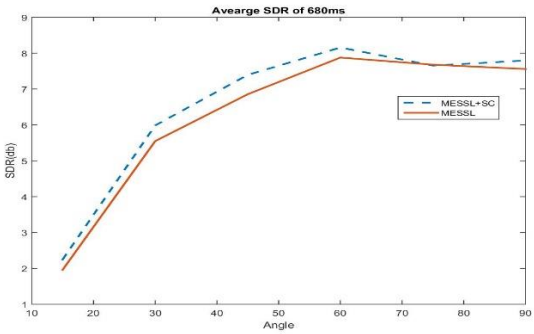

Figure.5 SDR Comparison of proposed model and MESSL Model at Reverberation of 680ms

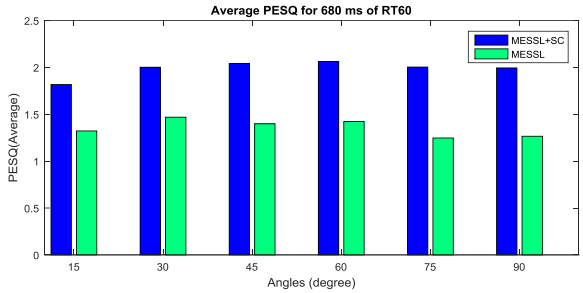

Figure.6 PESQ Comparison of proposed model and MESSL Model at Reverberation of 680ms

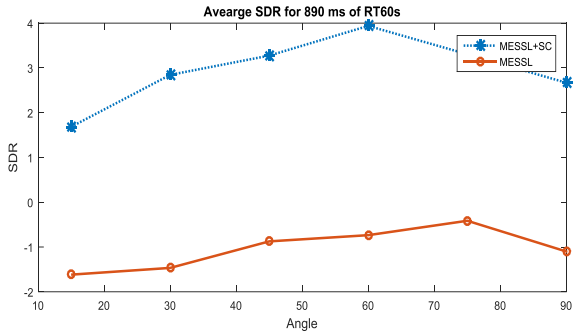

Figure.7 SDR Comparison of proposed model and MESSL Model at Reverberation of 890ms

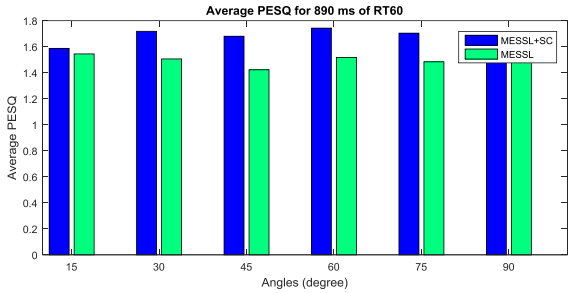

Figure.8 PESQ Comparison of proposed model and MESSL Model at Reverberation of 890ms

The above figures 5-8 clarify that whenever the reverberation time is further increased up to 680ms and 890ms the performance of the proposed model is improved more over the MESSL model. Figure. 8 shows that at the highest RT_{60}s(890ms) under consideration, the proposed model outperforms MESSL by and Average SDR of 3.05 db.

CONCUSLION

A speechaseparation algorithm in reverberant environment isaproposed, that integrate the models of the interauralaparameters and spatialacovariance. The observed mixtures wereamodeled and initialized using the EMaalgorithm giving improved source estimates. Results indicate that the performance of our proposed model is much better than the existing model MESSL.

FUTURE WORK

The algorithm proposedain the paper can be extendedain numberaofaways. In proposed model all the sources of sound are assume to be stationary sources. In practical scenarios this is not possible that all of the time the sources will be stationary they can change their positions. So case of movable sources is future challenges of this algorithm.

By applying the effect of covarianceamodel with models of IPD and ILD assumed that the reverberation time was known. Science the estimation for finding the reverberation time was not aims of this thesis. The future work in this thesis is that to include algorithm to detect the reverberation time i.e. by spectral subtraction. Finally, in futureawork researcher could focus on the algorithm complexity reduction.

REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[2] A. Ephrat *et al.*, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," vol. 37, no. 4, 2018.

[3] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[4] M. L. Thesis, "Blind Single Channel Sound Source Separation Mark Leddy B . Sc , M . Sc Dublin Institute of Technology Supervisors : Dan Barry , David Dorran , Eugene Coyle," 2010.

[5] N. Hassan and D. A. Ramli, "A Comparative study of Blind source separation for Bioacoustics sounds based on FastICA, PCA and NMF," *Procedia Comput. Sci.*, vol. 126, pp. 363–372, 2018.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] S. Rickard, "The DUET Blind Source Separation Algorithm," pp. 217–241, 2007.

[8] V. S. Narayanaswamy, S. Katoch, J. J. Thiagarajan, H. Song, and A. Spanias, "Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets," 2019.

[9] X. F. Gong, Q. H. Lin, F. Y. Cong, and L. De Lathauwer, "Double Coupled Canonical Polyadic Decomposition for Joint Blind Source Separation," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3475–3490, 2018.

[10] S. Rickard and O. Yilmaz, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[11] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source Localization in Reverberant Environments : Part I - Modeling," vol. 11, no. 6, pp. 1–22, 2003.

[12] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2, pp. 4409–4412, 2009.

[13] Z. Rafii, A. Liutkus, F. R. Stoter, S. I. Mimilakis, D. Fitzgerald, and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 8, pp. 1307–1335, 2018.

[14] M. Jia, J. Sun, C. Bao, and C. Ritz, "Separation of multiple speech sources by recovering sparse and non-sparse components from B-format microphone recordings," *Speech Commun.*, vol. 96, no. May 2017, pp. 184–196, 2018.

[15] S. Smita, S. Biswas, and S. S. Solanki, "Audio Signal Separation and Classification: A Review Paper," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 11, pp. 6960–6966, 2014.

[16] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 9–12, 2010.

[17] M. S. Khan, S. M. Naqvi, and J. Chambers, "Two-stage audio-visual speech dereverberation and separation based on models of the interaural spatial cues and spatial covariance," *2013 18th Int. Conf. Digit. Signal Process. DSP 2013*, 2013.

[18] B. Schuller, "【Metrics】Performance Measurement in Blind Audio Source Separation," vol. 14, no. 4, pp. 139–147, 2013.

[19] J. J. Thiagarajan, K. Natesan Ramamurthy, and A. Spanias, "Mixing matrix estimation using discriminative clustering for blind source separation," *Digit. Signal Process. A Rev. J.*, vol. 23, no. 1, pp. 9–18, 2013.

[20] K. Yatabe and D. Kitamura, "Determined Blind Source Separation via Proximal Splitting Algorithm," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, no. 4, pp. 776–780, 2018.

[21] A. Tsilfidis, E. Georganti, and J. Mourjopoulos, "Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 5, pp. 1737–1740, 2011.