Time difference of arrival (TDOA)-based acoustic source localization and signal extraction for intelligent audio classification

Mingsian R. Bai, *senior member*
Department of Power Mechanical Engineering
National Tsing Hua University
Hsinchu, Taiwan
msbai63@gmail.com

Shih-Syuan Lan
Department of Power Mechanical Engineering
National Tsing Hua University
Hsinchu, Taiwan
danny821020@gmail.com

Jong-Yi Huang
Department of Power Mechanical Engineering
National Tsing Hua University
Hsinchu, Taiwan
tea6329714@gmail.com

***Abstract*— An intelligent system is proposed to locate and classify audio source signals in large spaces. The system is composed of a sparsely distributed microphone array and an artificial intelligence (AI) system. A sparse array aimed for acoustic source localization and signal extraction is configured. The localization method is based on time difference of arrival (TDOA). This method begins with estimation of the TDOAs among at least 4 sensors, with the aid of a subspace-based time delay estimation algorithm. Next, a constrained least squares (CLS) algorithm is applied to locate the source in accord with the estimated TDOAs. Once the source is located, the source signal is extracted by using the minimum variance distortionless response (MVDR) beamformer and a postfilter. The extracted audio signals are further classified in light of machine learning. Convolutional Long Short-Term Memory (ConvLSTM) plays a central role in the AI-based classifier. Mel-Frequency Spectral Coefficients (MFSC) serves as the input layer in the ConvLSTM. The performance of the proposed system is quantified by using localization error, audio quality, and F_1 scores. Simulations are undertaken to validate the proposed TDOA-based localization and separation technique.**

Keywords—delay estimation, TDOA, localization, beamforming, neural networks

I. INTRODUCTION

The localization via sensor arrays has received considerable attention in numerous applications including radar, sonar, navigation, wireless communications and sensor networks [1-6]. There are many localization techniques, among which the methods based on time difference of arrival (TDOA) measurements are particularly suited for sparse array configuration as often seen in smart home scenarios. This work proposes a TDOA-based localization approach that is modified from So and Chan's method, the Constrained Least Square (CLS) algorithm [5] which was originally formulated in terms of time of arrival (TOA) [7]. TDOA is the time difference between two microphone signals. Time synchronization is required among all microphones. In a three-dimensional space, each TDOA corresponds to a hyperboloid. The location of one source position can be determined from the intersection of at least 3 TDOA measurement at 4 microphones.

Instead of conventional correlation-based delay estimation methods, a new subspace algorithm is developed for single-source scenarios in this work. The TDOA can be estimated by using array covariance matrix and the multiple signal classification (MUSIC) algorithm. With the estimated TDOAs, the source location can be determined by solving a nonlinear system of equations [8-10]. However, the nonlinear equations can also be linearized by introducing a constraint equation, which can be nicely solved using the technique of Lagrange multipliers [9]. Once the source is located, the source signal can be extracted by using the MVDR beamformer with a Bayesian minimum-mean-square-error (MMSE) postfilter [11].

In this paper, isolated vocabulary is employed as audio stimuli to be classified. The source signal extracted by the preceding microphone array is classified in light of machine learning. Traditional speech recognition methods based on frontend Gaussian mixture model (GMM) in conjunction with backend sequencing methods such as the Hidden Markov Model (HMM), the support vector machine (SVM) [12], and the k-nearest neighbors (k-NN) [13] have been well established. These conventional approaches rely on reliably extracted feature such as the Mel-frequency cepstrum coefficients (MFCC) [14]. By contrast, convolutional neural networks (CNNs) do not require hand-crafted high-level features as input data. The network model is capable of automatically "learn" the feature representations and network parameters in a congruent fashion. Recurrent neural networks (RNNs) are also incorporated as a backend classifier to enhance the classifier due to the time-evolving nature of audio signals. RNNs can be trained to map an input sequence of infinite length into a finite-dimensional vector representation.

II. DELAY ESTIMATION

Fig. 1 depicts the flowchart of audio classification. Once the microphones receive the signal, we can locate the source position based on the TDOA information. After the source is located, the source signal can be extracted by using the MVDR beamformer with a Bayesian MMSE postfilter. Finally, classify the source by the AI-based classifier.

978-1-5386-4752-3/18/$31.00 ©2018 IEEE

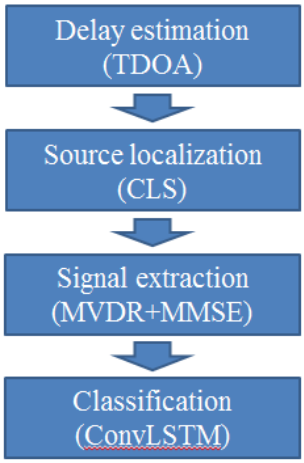

Fig. 1. The flowchart of audio classification.

TDOA is the difference in arrival times of the source signal received at a pair of sensors. Multiplying the TDOA by the sound speed gives the range difference between the source and two sensors. Many methods based on generalized cross-correlation can be used to estimate TDOAs. In this paper, we approach this problem with a MUSIC-inspired algorithm. This algorithm depends on the separation of two orthogonal subspaces, "signal subspace" and "noise subspace," pertaining to the spatial correlation matrix.

Assume a signal emanating from a remote source at two microphones. The impulse response between two microphones including the direct sound and reflections takes the form

$$h_{12}(t) = \sum_{d=1}^{D} s_d \delta(t - \tau_d), \tag{1}$$

where τ_d is the lth path delay (d=1 being the direct path and $d = 2,\ldots, D$ being reflections). Hence, the frequency response function is

$$H_{12}(j\omega) = \sum_{d=1}^{D} s_d e^{-j\omega\tau_d}, \tag{2}$$

Sampling the frequency response at K frequencies gives

$$H_{12}(j\omega_k) = \sum_{d=1}^{D} s_d e^{-j\omega_k\tau_d},\ k = 1,\ldots,K. \tag{3}$$

which can be rewritten in the matrix form

$$\mathbf{H} = \mathbf{G}\mathbf{s}, \tag{4}$$

where

$$\mathbf{H} = [H_{12}(j\omega_1) \quad \cdots \quad H_{12}(j\omega_K)]^T,\ \mathbf{s} = [s_1 \quad \cdots \quad s_D]^T,$$
$$\mathbf{G} = [\mathbf{a}_d \quad \cdots \quad \mathbf{a}_D],\ \text{and}\ \mathbf{a}_d = [e^{-j\omega_1\tau_d} \quad \cdots \quad e^{-j\omega_K\tau_d}]^T$$

It follows that the MUSIC pseudospectrum can be calculated

$$S_{MUSIC}(\tau) = \frac{1}{\mathbf{a}^H(\tau)(\mathbf{I} - \mathbf{U}_S\mathbf{U}_S^H)\mathbf{a}(\tau)}, \tag{5}$$

where the eigenvalue decomposition (EVD) [15] of the covariance matrix $\mathbf{R}_{xx} = E\{\mathbf{H}\mathbf{H}^H\} = \mathbf{U}\Lambda\mathbf{U}^H$, and $\mathbf{U} = [\mathbf{U}_S \quad \mathbf{U}_N]$. $E\{\bullet\}$ is the expectation operator, "H" denotes Hermitian transpose, $\mathbf{U}_S$ and $\mathbf{U}_N$ are associated with the signal and noise subspaces. $\mathbf{U}_S$ is one-dimensional for the single source scenario considered herein. The first peak of the MUSIC pseudospectrum gives the delay of the direct path.

III. Source Localization

Consider one source, M+1 microphones in which the 0th microphone is designated as the reference microphone. The difference in the distances of the source to the mth microphone and the source to the reference microphone can be written as

$$d_{m0} = \tau_{m0}c + n_m,\ \ m = 1,\ldots,M, \tag{6}$$

where τ_{mo} is the time delay, c is the sound speed, and n_m is Gaussian white measurement noise with variance $\sigma_m{}^2$. We assume that the source coordinate is (x, y, z), the mth microphone coordinate is $(x_m, y_m, z_m), m = 1,\ldots,M$, and the reference microphone is placed at the origin. Therefore, by omitting the measurement noise, (6) can be rewritten as

$$d_{m0} = \sqrt{(x - x_m)^2 + (y - y_m)^2 + (z - z_m)^2} - \sqrt{x^2 + y^2 + z^2}, \tag{7}$$

By defining $R = \sqrt{x^2 + y^2 + z^2}$, (8) can be converted to a system of linear equations. In matrix form,

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}, \tag{8}$$

where

$$\mathbf{A} = \begin{bmatrix} x_1 & y_1 & z_1 & -d_{10} \\ x_2 & y_2 & z_2 & -d_{20} \\ \vdots & \vdots & \vdots & \vdots \\ x_M & y_M & z_M & -d_{M0} \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} x \\ y \\ z \\ R \end{bmatrix},$$
$$\mathbf{b} = \frac{1}{2}\begin{bmatrix} x_1{}^2 + y_1{}^2 + z_1{}^2 - d_{10}{}^2 \\ x_2{}^2 + y_2{}^2 + z_2{}^2 - d_{20}{}^2 \\ \vdots \\ x_M{}^2 + y_M{}^2 + z_M{}^2 - d_{M0}{}^2 \end{bmatrix}.$$

By assuming high signal-to-noise ratio (SNR) of measurement, the squared distance can be written as

$$d_{m0}^2 = (\tau_{m0}c + n_m)^2 \approx (\tau_{m0}c)^2 + 2(\tau_{m0}c)n_m. \tag{9}$$

As a result, the deviation between the true and the measured squared distance is

$$\varepsilon_m = d_{m0}^2 - (\tau_{m0}c)^2 \approx 2(\tau_{m0}c)n_m. \tag{10}$$

Thus, the covariance matrix of the deviation is of the form

$$\boldsymbol{\psi} = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \mathbf{B}\mathbf{Q}\mathbf{B}, \tag{11}$$

with

$$\mathbf{B} = \text{diag}(2\tau_{10}c, \tau_{20}c, \ldots, \tau_{M0}c),\ \mathbf{Q} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2).$$

978-1-5386-4752-3/18/$31.00 ©2018 IEEE

Let the weighting matrix $\mathbf{W} = \boldsymbol{\psi}^{-1}$. The weighted least-square localization problem can be stated as a constrained optimization problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{A}\boldsymbol{\theta} - \mathbf{b})^T \mathbf{W} (\mathbf{A}\boldsymbol{\theta} - \mathbf{b}) . \quad (12)$$

subject to

$$x^2 + y^2 + z^2 - R^2 = \boldsymbol{\theta}^T \mathbf{P} \boldsymbol{\theta} = 0 \text{ with } \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}. \quad (13)$$

Minimizing the Lagrangian [9, 16],

$$L(\boldsymbol{\theta}, \lambda) = (\mathbf{A}\boldsymbol{\theta} - \mathbf{b})^T \boldsymbol{\psi}^{-1} (\mathbf{A}\boldsymbol{\theta} - \mathbf{b}) + \lambda (\boldsymbol{\theta}^T \mathbf{P} \boldsymbol{\theta}) , \quad (14)$$

leads to the solution

$$\hat{\boldsymbol{\theta}}_{cw}(\lambda) = (\mathbf{A}^T \boldsymbol{\psi}^{-1} \mathbf{A} + \lambda \mathbf{P})^{-1} (\mathbf{A}^T \boldsymbol{\psi}^{-1} \mathbf{b}) , \quad (15)$$

where the Lagrange multiplier λ can be determined from the constraint equation, $\hat{\boldsymbol{\theta}}_{cw}^T(\lambda) \mathbf{P} \hat{\boldsymbol{\theta}}_{cw}(\lambda) = 0$ by using the secant method. Once the source is located, the source signal can be extracted by using the MVDR beamformer cascaded with a Bayesian minimum mean-square-error (MMSE) postfilter [17].

IV. Classification Using Neural Networks

A. Audio preprocessing

Isolated word classification is adopted as an example problem in the paper. In the simulation, 10 English vocabulary, "yes, no, up, down, left, right, on, off, stop, go," are to be identified. The speech source signal emanated a assumed position and received at five microphones. Every speech clip for training and testing is one-second recording. Mel-Frequency Spectral Coefficient (MFSC) is used as the input layer in the network. The process of MFSC calculation illustrated in Fig. 2 is similar to MFCC calculated without the discrete cosine transform at the last stage. First, the speech signal sampled at 16 kHz is normalized with the maximum amplitude. Second, the speech signal is processed in 32-ms segments with 512-point FFT and 50% overlap. Third, we calculate the power spectrum for each frame, apply 40 Mel filters [14] to the power spectrum, and sum the energy in each band. Lastly, 40 MFSCs are obtained for each frame by taking the logarithm of all band energies.

Fig. 2. MFSC flowchart.

B. Setting of audio classification network

ConvLSTM networks are designed to classify the speech signals, as illustrated in Fig. 3. The input layer is composed of 63×40 MFSC matrices. Layer 1 contains one ConvLSTM layer with 128 hidden states, layer 2 contains one ConvLSTM layer with 64 hidden states, layer 3 contains one ConvLSTM layer with 32 hidden states, and layer 4 contains one fully connected layer with 10 neurons. All state-to-state and input-to-state kernel size are 3×3. The output layer is a 10-element probability vector obtained using Softmax functions. The Adaptive Moment Estimation (Adam) is utilized as the optimization algorithm for training the ConvLSTM networks, with a learning rate 0.001. The kernel parameters of ConvLSTM are initialized with the Glorot method [21]. The parameters of recurrent cell of ConvLSTM are initialized with orthogonal initialization [22]. The mini-batch size is set to be 100.

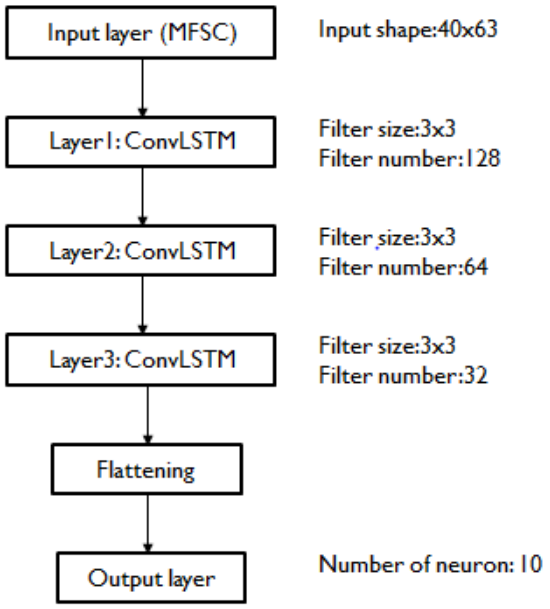

Fig. 3. Network setting.

C. ConvLSTM

In the following, a special ConvLSTM network that combines the CNN and the LSTM is presented. The ConvLSTM has convolutional structures in both input-to-state and state-to-state transitions, which is well suited to modeling the spatial and temporal relationships of data. The ConvLSTM is carried out in the following operations [19]:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi} * \mathbf{x}_t + \mathbf{W}_{hi} * \mathbf{h}_{t-1} + \mathbf{W}_{ci} \circ \mathbf{c}_{t-1} + \mathbf{b}_i) \quad (26)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf} * \mathbf{x}_t + \mathbf{W}_{hf} * \mathbf{h}_{t-1} + \mathbf{W}_{cf} \circ \mathbf{c}_{t-1} + \mathbf{b}_f) \quad (27)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo} * \mathbf{x}_t + \mathbf{W}_{ho} * \mathbf{h}_{t-1} + \mathbf{W}_{co} \circ \mathbf{c}_t + \mathbf{b}_o) \quad (28)$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(\mathbf{W}_{xc} * \mathbf{x}_t + \mathbf{W}_{hc} * \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (29)$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) , \quad (30)$$

where $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t, \mathbf{c}_t$ and $\mathbf{h}_t$ are the input gate, the forget gate, the output gate, the cell activation, and the hidden layer. $\sigma(.)$ is the logistic sigmoid function. $\mathbf{W}$'s are the weight matrices. The subscripts are self-explanatory, e.g., $\mathbf{W}_{xo}$ is the input-output gate matrix. $\mathbf{b}$'s are the bias vectors. ' $\circ$ ' denotes the Hadamard product and ' $*$ ' denotes the convolution operator.

978-1-5386-4752-3/18/$31.00 ©2018 IEEE

V. Simulation

A. Localization

Fig. 4 depicts the simulation setting. The sparse distributed array installed on room boundary. The reference microphone is located at (0, 0, 0)m, and another five microphones are located at (-4.714, 8.165, -3.333)m, (-4.714, -8.165, -3.333)m, (9.428, 0, -3.333)m, and (0, 0, 10)m. In the following Monte Carlo simulation, a point source broadcasting speech signals with 10 dB SNR is randomly positioned in a freefield. The root-mean-square (RMS) localization error is defined as

$$\Upsilon = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}\left|\hat{\mathbf{S}}_i - \mathbf{S}_i\right|^2}\quad ,\ \text{where}\ \ \hat{\mathbf{S}}_i,\ i=1,\ldots,N_s\ \text{being}$$

the estimated positions for each time, N_s being the number of iteration and $\mathbf{S}_i,\ i=1,\ldots,N_s$ are true positions for each time. Here, we assume N_s = 10000. The RMS errors in the localization results obtained using the MUSIC delay estimation method and the generalized cross-correlation with phase transformation (GCC-PHAT) [23-25] method are 2.3mm and 35mm, respectively. The former method is significantly superior to the latter.

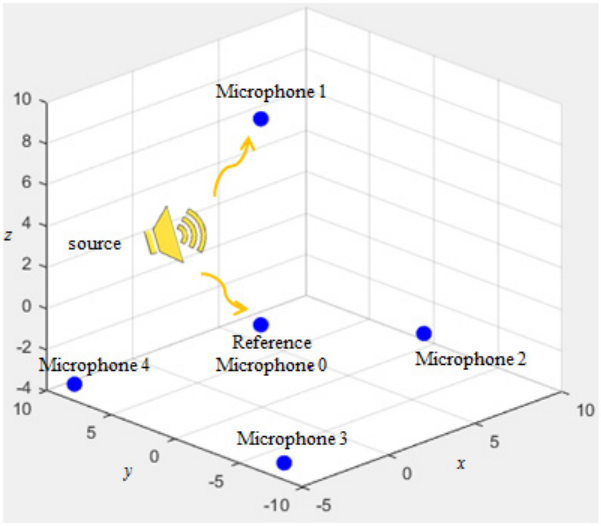

Fig. 4. The simulation setting.

B. Signal extraction

The source signals have been extracted by using the delay- and-sum (DAS) and MVDR beamformers, with results summarized in TABLE I. Perceptual Evaluation of Speech Quality (PESQ) [20] is employed as an objective metric to assess speech quality. The MVDR beamformer cascaded with the MMSE postfilter has performed the best.

TABLE I. Signal Extraction Performance

	DAS	MVDR	MVDR+MMSE
PESQ	1.0984	1.1903	1.5565

C. Performancn

The classification performances achieved using the NN by a single microphone and the microphone array are compared in TABLE II. Here, the averaged F_1 score [26] is used to quantify the classification performance. As can be seen in the result, the microphone array based on the TDOA localization has attained significantly higher F_1 score than the single microphone. However, combined use of the MVDR beamformer and the MMSE postfilter effectively suppress background noise at some cost of speech distortion.

TABLE II. Classification Performance

	Averaged F_1 score (%)
Single microphone	72.5
DAS	84.9
MVDR	84.9
MVDR+MMSE	83.4

VI. Acknowledgements

The work was supported by the Ministry of Science and Technology (MOST) in Taiwan, Republic of China, under the project number 105-2221-E-007-030-MY3. Thanks also go to Mr. Chin-Pu Tsai for his enthusiastic support of the Telecom Electroacoustics Audio laboratory (TEA lab).

References

[1] H. C. So , Y. T. Chan , and F. K. W. Chan , " Closed - form formulae for optimum time difference of arrival based localization , " IEEE Trans. Signal Process. , vol. 56 , no. 6 , pp. 2614 – 2620 , 2008 .

[2] K. W. Cheung, H. C. So , W. - K. Ma , and Y. T. Chan , " A constrained least squares approach to mobile positioning: algorithms and optimality , " EURASIP J. Adv. Signal Process. , vol. 2006, Article ID 20858, pp. 1 – 23, 2006

[3] K. W. K. Lui and H. C. So, "A study of two - dimensional sensor placement using time - difference -of - arrival measurements, " Digit. Signal Processing , vol. 19, no. 4 , pp. 650 – 659 , 2009 .

[4] H. C. So, "Source localization: Algorithms and analysis," Handbook of Position Location: Theory, Practice and Advances, Chapter 2, S. A. Zekavat and M. Buehrer, Eds., Wiley-IEEE Press, 2011

[5] K. W. Cheung, H. C. So, W.-K. Ma and Y. T. Chan, "Least squares algorithms for time-of-arrival based mobile location," IEEE Transactions on Signal Processing, vol.52, no.4, pp.1121-1128, April 2004

[6] H. C. So and L. Lin, "Linear least squares approach for accurate received signal strength based source localization," IEEE Transactions on Signal Processing, vol.59, no.8, pp.4035-4040, August 2011

[7] H. C. So and K. W. Chan, "A generalized subspace approach for mobile positioning with time-of-arrival measurements," IEEE Transactions on Signal Processing, vol.55, no.10, pp.5103-5107, October 2007

[8] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, pp. 1661–1669, Dec. 1987.

[9] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," IEEE Trans. Signal Processing, vol. 42, pp. 1905–1915, Aug. 1994.

[10] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," IEEE Trans. Speech, Audio Processing, vol. 9, pp. 943–956, Nov. 2001.

[11] P. Loizou Speech Enhancement: Theory and Practice FL Boca Raton:CRC Taylor & Francis 2007..

[12] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," IEEE Transactions on Neural Networks, vol. 14, no. 1, pp. 209–215, 2003.

[13] J. Dennis, H. D. Tran, and E. Si. Chng, "Image feature representation of the subband power distribution for robust sound event classification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 367–377, 2013.

[14] S. Davis, P. Mermelstein, "(1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4, pp. 357-366, 1980.

[15] C. Knapp, G. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 24(4), pp. 320-327, 1976..

[16] P. C. Chen, "A nonline-of-sight error mitigation algorithm in location estimation," in Proc. IEEE Wireless Commun. Networking Conf., vol. 1, New Orleans, LA, 1999, pp. 316–320.

[17] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on. pp. 496-499, 2001.

[18] D. E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning representations by back-propagating errors, Neurocomputing: foundations of research," MIT Press, Cambridge, MA, 1988

[19] X. Shi, Z. Chen, H. Wang and D. Yeung, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," arXiv: 1506.04214 ,2015

[20] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Geneva, 21pages (2001).

[21] X.Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proceedings of AISTATS 2010, vol. 9, pp. 249–256.

[22] A. M. Saxe, J. L. McClelland and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,"

[23] B. van den Broeck, A. Bertrand, P. Karsmakers, B. Vanrumste, H. van Hamme, M. Moonen, "Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays," IEEE Conference Publications, pp. 76-80, June 2013

[24] C. Knapp, "The generalized correlation method for estimation of time delay, " IEEE Journals & Magazines, pp.320-327, August 1976

[25] C. Knapp, G. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 24(4), pp. 320-327, 1976

[26] M. Hossin and M. N.Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, no.2, March 2015