TECHNICAL DOCUMENTATION

Neural Network based DOA Estimation as an Application to Two-Microphone Blind Speech Separation

Chandan K A Reddy, Gautam S Bhat, Nikhil Shankar, Issa Panahi, SSPRL UT Dallas

A typical approach to Blind Speech Separation (BSS) of convolutive mixtures is to find the Demixing matrix at each frequency bin to separate the sources. However, this approach results in the permutation problem. Independent Vector Analysis (IVA) is a BSS method that drew attention in the past decade, as it inherently solves the permutation problem and effectively separates the sources. Nevertheless, its iterative nature of obtaining the Demixing matrix makes it impractical to use in Real-Time audio applications that demand low computational latency. In this work, we propose a BSS method that can work in Real-Time and a Neural Network based detection criteria that indicates when to update the Demixing matrix, hence significantly reducing the average number of computations. The detection criterion is based on sensing significant changes in the transfer function between the speech source and the microphone.

Our studies show that, if the impulse response between the source and the microphone does not change significantly, then there is no significant change in the Demixing matrix, thereby insignificant performance improvement in source separation. This indicates that it would be inefficient to iteratively calculate the Demixing matrix for each incoming frame. We propose a Neural Network based detection criterion that tracks substantial changes in the source Direction of Arrival (DOA). The idea of using the DOA information is to update the Demixing matrix only if there is a significant change in the source direction. The Demixing matrix is updated only in the frame that satisfies the detection criterion. The detection criterion monitors at least a 30^0 change in the source direction. This approach substantially reduces the average number of computations making it feasible to run in Smartphone-Hearing Aid setup. Our objective and subjective experiments show significant improvements over conventional IVA and its variations.

Overview of Independent Vector Analysis (IVA)

The mixing process in the real-world acoustic environment includes delays, attenuations and reverberations, i.e., signals are convolutively mixed. For instance, if there are *P* sources and *Q* sensors, the signal captured by sensor *q* is given below,

$$x_q(n) = \sum_{p=1}^{P} a_{qp}(n) * s_p(n) \tag{1}$$

where (*Q*>=*P*), (*) is the convolution. $a_{qp}(n)$ is the finite duration impulse response mixing filter from source *p* to sensor *q*. On applying STFT (with STFT frame length sufficiently longer than the mixing filter length), the time domain expression in (1) can be converted to multiplication in the frequency domain given by,

$$x_q{}^{[f]}(m) = \sum_{p=1}^{P} a_{qp}{}^{[f]}\, s_p{}^{[f]}(m) \tag{2}$$

$$\boldsymbol{x}^{[f]}(m) = \boldsymbol{A}^{[f]}\boldsymbol{s}^{[f]}(m) \tag{3}$$

where $s_p{}^{[f]}(m)$, $x_q{}^{[f]}(m)$ and $a_{qp}{}^{[f]}$ are frequency domain versions of $s_p(n)$, $x_q(n)$ and $a_{qp}(n)$ respectively at frame index m. $\boldsymbol{x}^{[f]}(m) = [x_1{}^{[f]}(m), \dots\dots, x_Q{}^{[f]}(m)]$, $\boldsymbol{s}^{[f]}(m) =$

$[s_1{}^{[f]}(m), \dots \dots \dots, s_P{}^{[f]}(m)]$ and $\boldsymbol{A}^{[f]}$ is the mixing matrix for frequency bin *f*, with $a_{qp}{}^{[f]}$ as its entries for each frame *m*.

The goal of IVA is to find a demixing matrix $\boldsymbol{W}^{[f]}$ at each frequency bin f such that,

$$\boldsymbol{y}^{[f]}(m) = \boldsymbol{W}^{[f]}\boldsymbol{x}^{[f]}(m) \tag{4}$$

where $\boldsymbol{y}^{[f]}(m)$ is the close estimate of $\boldsymbol{s}^{[f]}(m)$.

The objective function of IVA is given below:

$$J_{IVA} = KL(p_{\boldsymbol{y}} || \prod_p p_{\boldsymbol{y}_p}) \tag{5}$$

The cost function is optimized using Natural Gradient method and the update equation is given by,

$$\boldsymbol{W}^{[f]} = \boldsymbol{W}^{[f]} + \eta \, \{\boldsymbol{I}\text{-E}\,[\Phi^{[f]}(\boldsymbol{y}^{[f]})(\boldsymbol{y}^{[f]})^{\mathrm{H}}]\}\, \boldsymbol{W}^{[f]} \tag{6}$$

Where

$$\Phi^{[f]}(\boldsymbol{y_p}) = \frac{y_p^{[f]}}{\sqrt{\sum_{f=1}^{F} \left|y_p^{[f]}\right|^2}} \tag{7}$$

The iterative approach to compute the Demixing matrix as in (6) is computationally very expensive, as there is uncertainty in the number of iterations required for convergence. Hence, computing demixing matrix for every frame is infeasible.

Overview of Neural Network based Source Localization

The proposed method is developed to estimate the DOA of a speech source in a two-microphone array scenario. Time Difference of Arrival (TDOA) is nonlinearly related to the estimate of DOA. In the conventional Generalized Cross Correlation (GCC) approach, the TDOA is estimated between the microphone pair by picking the lag with maximum correlation using the GCC pair. If the lag number is picked correctly, TDOA can be estimated accurately resulting in the accurate estimate of DOA. However, TDOA is unreliable when operating at lower SNR levels and in reverberant conditions. Therefore, using TDOA alone as a feature to estimate DOA will be incorrect. On the other hand, GCC vectors contain required patterns to estimate the DOA. Hence, GCC vector is chosen as the feature representation to train the Feedforward Neural Network model. Figure 1 shows the Block diagram of the Neural Network based DOA Estimation. Let $\boldsymbol{x}_1(n)$ and $\boldsymbol{x}_2(n)$ be the input vectors of frame length L from the 2 microphones at time index n. Let $\boldsymbol{r}(n) = \left[|r_{\boldsymbol{x_1x_2}}(-m)|, \dots, |r_{\boldsymbol{x_1x_2}}(0)|, \dots, |r_{\boldsymbol{x_1x_2}}(m)|\right]$ be a vector of absolute values of the cross-correlation coefficients between the input frames from the two microphones at time index n. We drop n for brevity. The input feature vector consisting of the normalized cross-correlation coefficients of $\boldsymbol{r}$ at the valid lags $-m$ to m is given by,

$$\boldsymbol{U} = \frac{\left[|r_{\boldsymbol{x_1x_2}}(-m)|, \dots, |r_{\boldsymbol{x_1x_2}}(0)|, \dots, |r_{\boldsymbol{x_1x_2}}(m)|\right]}{\max(\boldsymbol{r})} \tag{8}$$

In the case of 16 kHz sampling rate and 13 cm separation between the microphones, the value of m is 6.

In this work, the Feedforward Neural Network is considered to have only one hidden layer with 8 nodes, which is empirically decided. We consider "Rectified Linear Unit (ReLU)" as the activation function at the nodes of the hidden layer. Let $\boldsymbol{Z_1}$ denote the output of the hidden layer, which is given by,

$$\boldsymbol{Z}_1 = \max(0, \boldsymbol{W}_1\boldsymbol{U}) \tag{9}$$

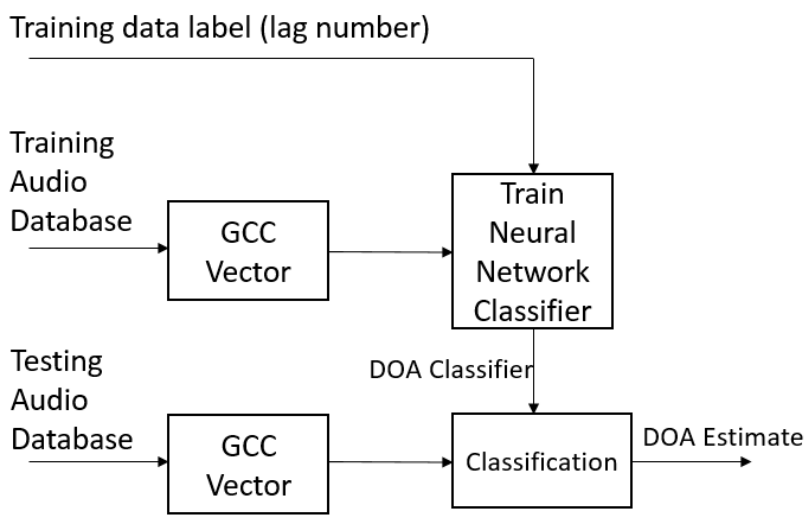

Figure 1. Block diagram of the Neural Network based DOA Estimation

$\boldsymbol{W}_1$ is the linear transformation weights from the input layer to the hidden layer. Max function is used to introduce non-linearity in the hidden layer. This helps in learning non-linear relationship between the input and the output. Let $\boldsymbol{Z}_2 = \boldsymbol{W}_2\boldsymbol{Z}_1$ be the linear transformation of $\boldsymbol{Z}_1$ to the output. $\boldsymbol{W}_2$ is the weights of the connections from the hidden layer to the output nodes. The weight vectors $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ is obtained using the first-order gradient-based optimization of stochastic objective function, which is called as 'Adam' [68].

The output layer of the Neural Network consists of the output classes, which are the angles of the DOA. In our work, 7 different angles between 0^0 and $180^0$are considered with a separation of 30^0. The reason for choosing only these 7 angles will be explained in the later section. Softmax function is used at the output nodes to give the probabilities of each class, which is given by,

$$p(\theta_n = c|\boldsymbol{U}(n)) = \frac{\exp(\boldsymbol{Z}_2(n))}{\sum_{k=1}^{C}\exp(\boldsymbol{Z}_2(n))}, c \in (0, C-1) \tag{10}$$

Each of the output class c will have a probability associated to it and the one with the highest probability will be the most probable class.

Experimental Results

The database is divided into 2 parts: 80% of the randomly selected data is used for training the Neural Network model and the remaining 20% is used for testing the accuracy of the model. The proposed method is compared with the traditional GCC based DOA estimation approach. Since we are dealing with a multi-class classification problem, the accuracy of the classifier is evaluated using a confusion matrix for different types of noises at different SNR conditions. We also cross validate the proposed method using K-Fold cross validation approach to check for robustness of the method with change in data. In the case of K-Fold cross validation, each unit in the database goes through both training and testing phase at least once.
The confusion matrix in the Figure 2 (a), (b) shows the overall classification performance of the proposed method and the GCC respectively for Machinery type of noise. The 10-Fold Cross Validation accuracy of proposed method is 86.2% with a Standard Deviation of 0.93%. The accuracy of GCC is around 44.1%.

	0	1	2	3	4	5	6
0	0.859	0.053	0.060	0.025	0.000	0.001	0.001
1	0.100	0.806	0.078	0.018	0.001	0.001	0.001
2	0.007	0.005	0.977	0.008	0.000	0.001	0.001
3	0.013	0.001	0.048	0.922	0.003	0.009	0.005
4	0.001	0.000	0.011	0.023	0.891	0.054	0.019
5	0.001	0.000	0.014	0.009	0.020	0.861	0.096
6	0.001	0.001	0.029	0.020	0.014	0.117	0.816

(a)

	0	1	2	3	4	5	6
0	0.000	0.594	0.068	0.331	0.007	0.000	0.000
1	0.000	0.505	0.152	0.337	0.006	0.000	0.000
2	0.000	0.000	0.476	0.524	0.000	0.000	0.000
3	0.000	0.000	0.000	1.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.355	0.645	0.000	0.000
5	0.000	0.000	0.001	0.308	0.109	0.112	0.470
6	0.000	0.000	0.003	0.333	0.081	0.054	0.530

(b)

Figure 2. Confusion matrix for the (a) Proposed method, and (b) GCC for Machinery noise

Proposed Solution to reduce the computational complexity of IVA

Figure 3 shows the block diagram of the proposed Real-Time implementation framework of IVA. This framework is based on the idea that the impulse response between the microphone and the speech source do not change significantly in short durations of time. The change in the impulse response is detected by tracking the change in the location of the source, which is achieved using the Neural Network based DOA

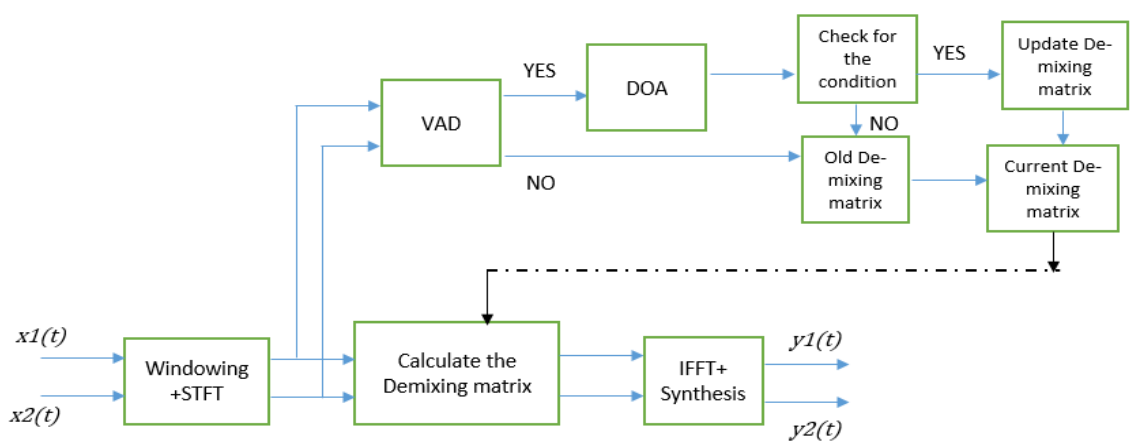

Figure 3 Block Diagram of the proposed method

estimation proposed in section 6.2. The convolutedly mixed noisy speech from the two microphones $x_1(n)$ and $x_2(n)$ are processed to estimate the DOA using the proposed Neural Network approach only at the speech portions.

Experimental Results

We compared our method with few other benchmark methods that are known to perform well in various noise conditions. The subjective results based on Mean Opinion Scores are shown in Figure 4. The results reflect the usefulness of the developed method in real-world noisy conditions.

We also compared our method objectively with few other benchmark methods that are known to perform well in various noise conditions. The objective results for Babble Noise based on PESQ (Perceptual Evaluation of Speech Quality), Signal to Distortion Ratio(SDR), Signal to Artifact Ratio(SAR), Signal to

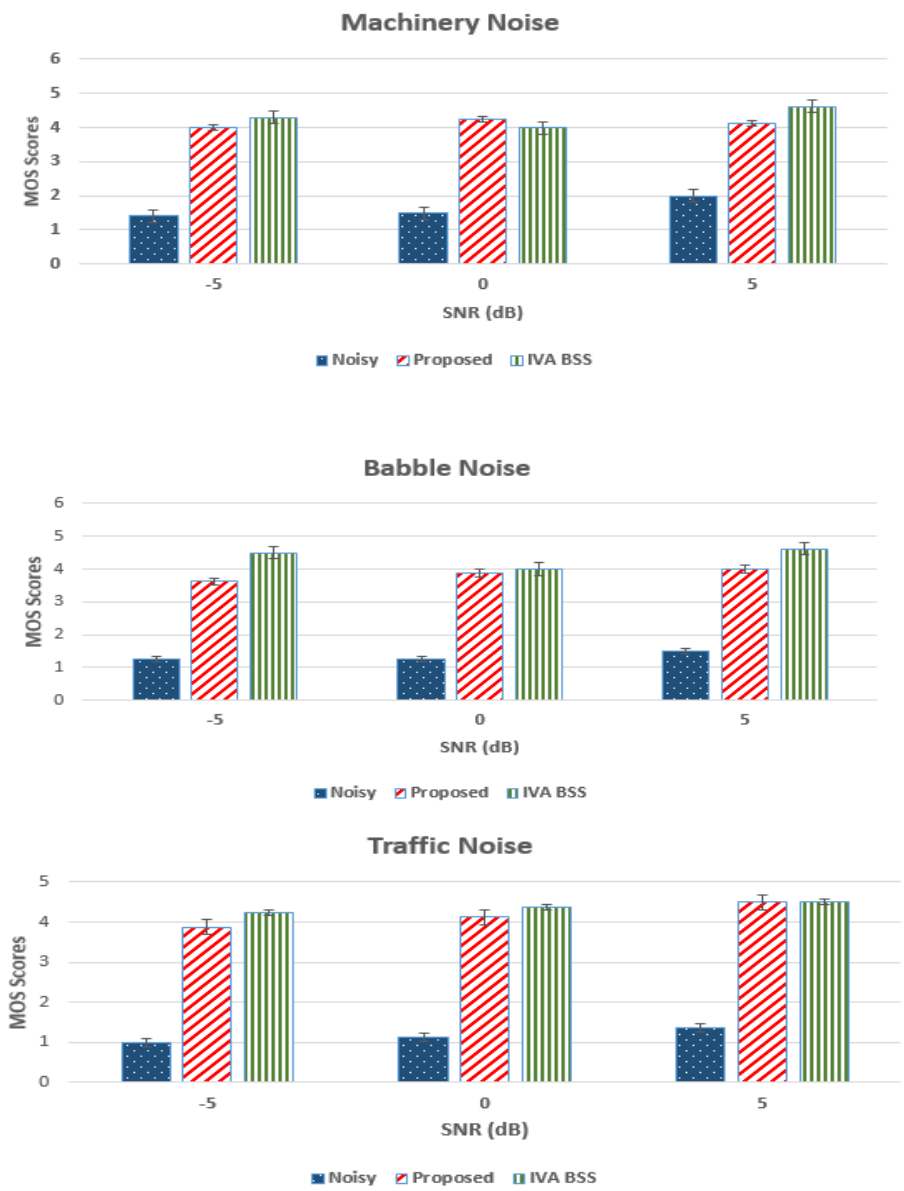

Figure 4 Subjective Test Results of IVA

Interference Ratio(SIR) are shown in Figure 5. The trends were the same for other noise types. The results reflect the effectiveness of the developed method in real-world noisy conditions.

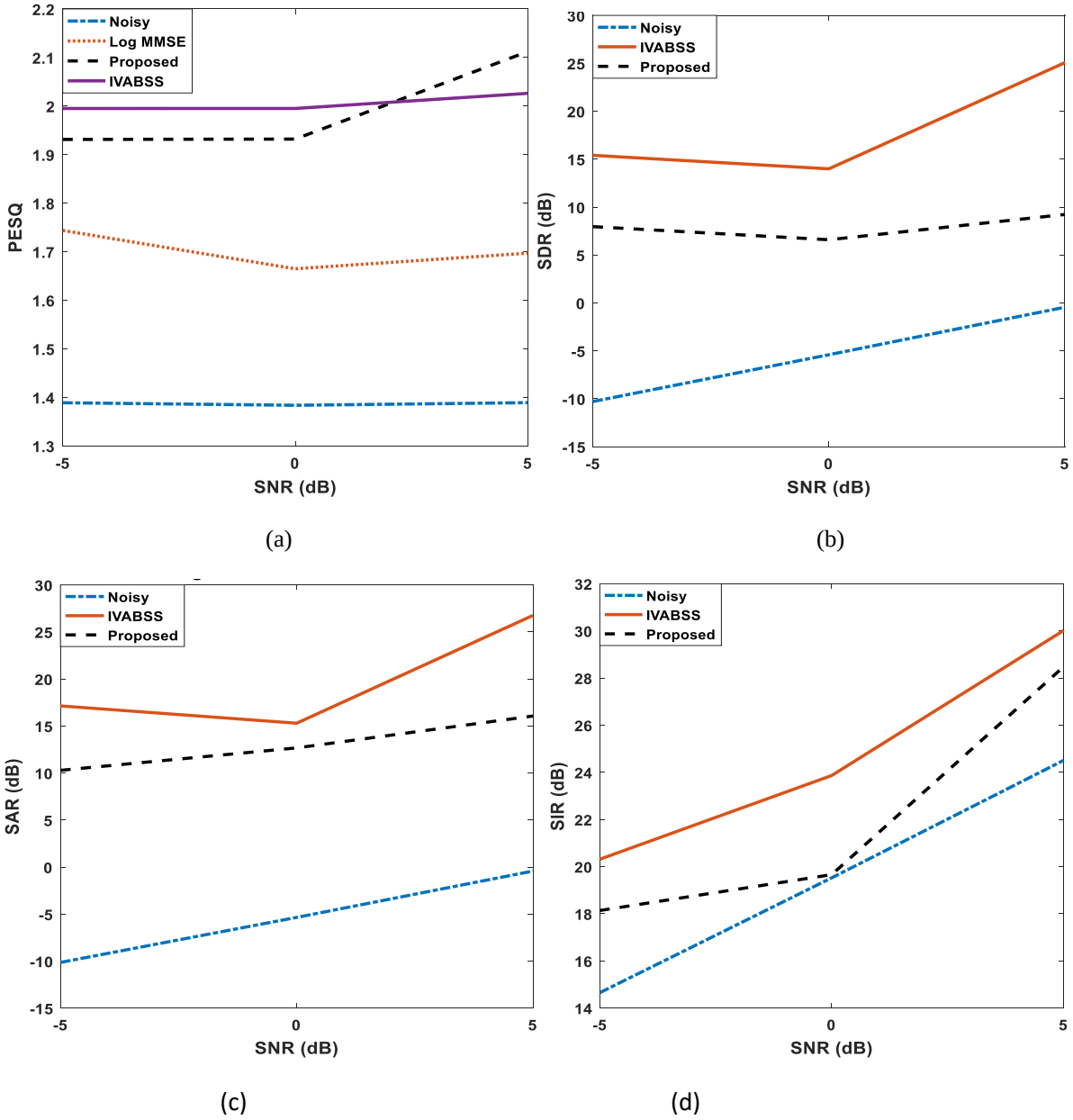

Figure 5 Performance evaluation for speech mixed with **Babble Noise** using (a) PESQ, (b) SDR (c) SAR and (d) SIR