Appeared in Proceedings of NCC, 2014

Improving Mobile Phone Based Query Recognition with a Microphone Array

Shrikant Venkataramani, Rajbabu Velmurugan and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India 400076
{vshrikant, rajbabu, prao}@ee.iitb.ac.in

***Abstract*—With mobile phone penetration high and growing rapidly, speech based access to information is an attractive proposition. However, automatic speech recognition(ASR) performance is seriously compromised in real-world scenarios where background acoustic noise is omnipresent. Speech enhancement methods can help to improve the signal quality presented to the automatic speech recognition at the receiving end. These methods typically exploit spectral diversity to achieve separation of speech from noise. While this works for most background noise, it fails for noise arising from speech sources such as interfering speakers in the vicinity of the caller. In this paper, we investigate the potential advantages of generating spatial cues via stereo microphones on the mobile phone handset to enhance speech. Such, enhancement of foreground speech can be done using blind source separation (BSS). This, when applied to the stereo mixtures before transmission is shown to achieve a significant improvement in ASR accuracy in the context of a mobile-phone based agricultural information access system.**

I. Introduction

Speech-based access to information is a powerful enabler given the widespread use of mobile phones across the country. The front-end of such an information retrieval system is an automatic speech recognition (ASR) system. While ASR technology is still far from perfect, practically useful systems are being deployed by exploiting domain-specific constraints in terms of language modeling and dialog design that embeds sufficient checks on decoded query accuracy. The main contributors to low recognition accuracy are signal degradation and speaker variability. The latter is addressed by the proper design of training datasets, and the incorporation of speaker normalization methods. Speech signal degradation is caused by environmental conditions such as noise and reverberation as well as channel degradation. Acoustic background noise could be caused by noise sources such as moving vehicles or by other speakers in the vicinity of the caller. While spectral diversity can be exploited to enhance speech in a background of non-speech sources, it is not applicable when the noise comprises interfering speech signals which typically share the same long-term spectrum.

With mobile phone hardware becoming increasingly sophisticated, better sound processing technology in the form of multiple microphones on the handset is realistic. For instance, iPhone5 and Samsung Galaxy S III both have a pair of microphones. The value of a microphone array in the enhancement of distant speech by means of beamforming has been recently established [1]. With the super-Gaussianity of speech signals providing the basis for blind adaptation in beamforming, improvements in ASR accuracy are reported on processed speech. Speech enhancement using spatial processing with the help of microphone array has gained interest recently [2].

In the present work, we explore the usefulness of spatial processing in the context of an available mobile-phone based information access system. The system provides information on agricultural commodities throughout Maharashtra based on a dialog in Marathi designed to elicit single-word or short-phrase queries from farmers seeking prices of specific commodities in particular district markets [3]. The database covers 32 districts, 279 markets and over 300 unique commodities. The system downloads, on a daily basis, prices of agricultural commodities from a website http://agmarknet.nic.in/ maintained by the Ministry of Agriculture, Government of India.

Accurate ASR would help improve both the user-friendliness and efficiency of the information access system. A prominent source of ASR errors is observed to be the degraded signal quality arising from the presence of acoustic background noise and interference at the farmer's end. The spatial separation of the caller and interfering sources can be exploited to achieve the suppression of interference via processing the output of a multi-microphone array. As mentioned earlier, interfering speakers pose the most challenging scenario to conventional single-channel speech enhancement techniques. We consider the situation of interfering speech coming from a speaker located at some distance from the caller. The stereo microphones assumed to be located on the mobile handset provide mixtures comprised of the caller and background speaker speech. As depicted in Fig. 1, BSS is expected to be carried out to obtain single-channel enhanced speech before the usual speech compression and transmission by the mobile handset. As the user responses are short phrases with brief pauses in between, the system requires fast operation of the BSS algorithm, but does not demand real-time response.

In the next section, we present a review of blind source separation methods including the signal-sparsity based method chosen for the present task. This is followed by a description of the present implementation on speech mixtures. The ASR system and datasets are described next. An experimental evaluation of the system without and with multi-channel processing

978-1-4799-2361-8/14/$31.00

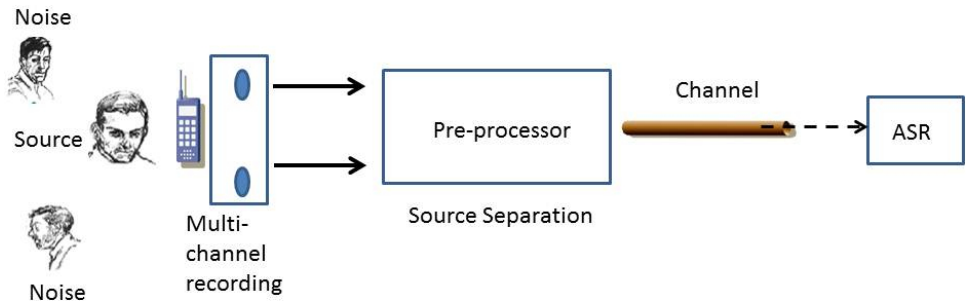

Fig. 1. A typical scenario where blind audio source separation can be applied to enhance the source or speaker signal amidst interfering signals or speakers. There are three sources ($S1$, $S2$, $S3$) and two microphones ($M1$, $M2$) in this scenario. The signals captured by the microphones represent the mixtures. The source locations are unknown and we assume the combined signals to represent an instantaneous mixture. From this limited information, the source signal needs to be extracted, transmitted and speech recognition task needs to be carried out.

is presented. The paper concludes with a discussion of the results and suggestions for further work.

II. Blind Source Separation (BSS) for Speech Enhancement

A simple BSS scenario is shown in Figure 1. In this scenario, there are three sources ($S1$, $S2$, $S3$) and two microphones ($M1$, $M2$) used to capture the utterances from these speakers. The microphone outputs will be an instantaneous mixture of the signals. Given the mixtures, our aim is to recover the individual sources ($S1$, $S2$ and $S3$). In the speech enhancement scenario considered here, one source signal is of interest (main caller using the information access system) and other two sources can be considered as the interferers. We will use source separation to separate and enhance the speech signal corresponding to the caller and use the processed signal in the query processing system.

The N mixture signals represented as $\mathbf{x}$ in a source separation problem is a product of transformation matrix $\mathbf{A}$ ($M \times N$) and the N source signals represented as $\mathbf{s}$. The transformation matrix referred to as the mixing matrix determines the contribution of each source in a mixture. Thus, a simple BSS scenario (in time domain) can be represented as,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{1}$$

The source vector s is of the form

$$\mathbf{s}(t) = \begin{bmatrix} s_1(t) & s_2(t) & \cdots & s_N(t) \end{bmatrix}^T. \tag{2}$$

The mixture vector x is of the form

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_M(t) \end{bmatrix}^T. \tag{3}$$

The mixing matrix A is a M x N matrix of the form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix}. \tag{4}$$

Here a_{ij} denotes the attenuation coefficient associated with the j-th source and the i-th mixture. They can be considered as the complex factor with which the signal gets attenuated as it travels from the source to the microphone. Each column of $\mathbf{A}$ relates to the spatial location of the corresponding source. In the system under consideration, $M = 2$ corresponding to a pair of microphones and $N \geq 2$. The mixing matrix elements are considered to be real, i.e., mixing is instantaneous wherein, no delay is assumed between the signals arriving at the pair of microphones. An extension to anechoic case is straight forward and is not considered here. Since we consider the queries to be usually made from open spaces or rooms, reverberation will not be an issue and need not be considered in this scenario. Accordingly, the BSS problem considered in this research is under- or evenly-determined and instantaneous in nature.

Audio signals are typically analysed in the time-frequency (TF) domain, as they are sparse in this domain. The source separation problem can be represented in the TF domain as

$$\mathbf{X}(t, f) = \mathbf{A}\mathbf{S}(t, f), \tag{5}$$

where t, f denote the TF point in a discrete TF framework. Most existing source separation problems make some assumptions on the source, mixing characteristics, and environment. The sources are usually modeled as a zero-mean Gaussian random variable in each of the TF point when analysed in the TF domain. Further these random variables corresponding to a single source will be assumed to be independent over time and frequency. The various sources in a mixture will also be assumed to be independent of each other. These assumptions lead to sparsity of speech mixtures in the TF domain and such sparsity has been exploited in many source separation methods. We consider two such methods in this paper. One being the degenerate un-mixing estimation technique (DUET) [4] and the other is the statistically sparse decomposition principle (SSDP) based method [5], [6]. They have a two step approach to the underdetermined BSS problem, by first estimating the mixing-matrix $\mathbf{A}$ and then recovering the sources using the estimated mixing-matrix. The DUET algorithm does both estimation of $\mathbf{A}$ and source separation. However, the SSDP methods assume availability of an estimated $\mathbf{A}$. In this work, we use the direction estimation of mixing matrix (DEMIX) algorithm [7], [8] to estimate $\mathbf{A}$ and use this in both the DUET and SSDP algorithms. Next, we provide a brief description of the DEMIX, DUET and SSDP algorithms.

A. DEMIX for mixing matrix estimation

We use the DEMIX algorithm proposed in [7], [8] to estimate the mixing matrix $\mathbf{A}$. It relies on the assumption that there are TF regions in the mixture where only one source contributes to the mixture. However, it uses a local TF region as opposed to the global TF region. Clustering using the global TF region as in DUET leads to inaccurate estimates since there may be more than one source active in a given TF point. The description here assumes the number of available mixtures (M) to be two. But the algorithm can handle more number of mixtures, anechoic and reverberant cases [8].

The DEMIX algorithm considers TF neighborhoods $\Omega_{t,f}$ around each TF point (t, f) in the discrete short-time Fourier Transform (STFT) $\mathbf{X}(t,f) = [X_1(t,f) \quad X_2(t,f)]^T$ of $\mathbf{x}(t) = [x_1(t) \quad x_2(t)]^T$. The STFT is computed using window length of L and $50\ \%$ overlap, with corresponding time indices $t = kL/2, k \in \mathbb{Z}$ and frequency indices $f = l/L, 0 \leq l \leq L/2$. One possible TF neighborhood for a (t, f) point is,

$$\Omega_{t,f} = \left\{ t + \frac{kL}{2}, f + \frac{k'}{L}, |k| \leq S_T, |k'| \leq S_F \right\}, \quad (6)$$

where S_T and S_F are chosen to define the number of points to be considered in the local neighborhood. The STFT points corresponding to a local neighborhood are represented as a matrix $\mathbf{X}(\Omega_{t,f})$, which is $2 \times |\Omega_{t,f}|$, where $|\Omega|$ represents cardinality of Ω. The data points in $\mathbf{X}(\Omega_{t,f})$ can also be viewed as a local scatter plot, from which an estimate of the principal direction or steering vector (SV) $\mathbf{u}(\Omega_{t,f}) \in \mathbb{R}^2$ and corresponding local confidence measure $\mathcal{T}(\Omega_{t,f})$ can be obtained. These can be estimated by performing a principal component analysis (PCA) on the covariance matrix $\mathbf{R}(\Omega_{t,f}) = \mathbf{X}(\Omega_{t,f})\mathbf{X}^H(\Omega_{t,f})$. Since we are considering the stereophonic case, the direction of the principal component obtained $\hat{\mathbf{u}}(\Omega_{t,f})$ can be translated into an angle $\hat{\theta}(\Omega_{t,f})$ corresponding to source direction. The confidence measure is defined as

$$\widehat{\mathcal{T}} := \frac{\hat{\lambda}_1(\Omega_{t,f})}{\hat{\lambda}_2(\Omega_{t,f})} \quad (7)$$

where $\hat{\lambda}_1(\Omega_{t,f}) \geq \hat{\lambda}_2(\Omega_{t,f})$ are the eigenvalues of $\mathbf{R}(\Omega_{t,f})$. The confidence measure distinguishes regions where one source is active from regions where more than one source is active. This also distinguishes regions where $\hat{\mathbf{u}}(\Omega_{t,f})$ points in the direction of a SV from regions where it might not be the case. Essentially, the SV and confidence measure estimates, together, indicate the direction of the most dominant source in a local neighborhood of the (t, f) point. The next step is to cluster these estimated $(\hat{\theta}, \widehat{\mathcal{T}})$ points. The clustering algorithm referred to as DEMIX and described in [8] was used in this work. We provide a brief description of this algorithm, which consists of three steps. The first step creates K clusters based on $\widehat{\mathcal{T}}$ and grouping closely spaced $\hat{\theta}$. The second step estimates the angle θ_k^c as centroid of clusters created in the first step. The variance of estimated angles is also computed and used in estimating the centroids. The third step eliminates spurious clusters from the K clusters formed in the earlier steps. This can be done using two approaches. One approach is to use the variance of the estimated centroids and thresholding based on these values to eliminate spurious clusters. The other approach is, when the number of sources N is known, to retain only the N clusters with smallest centroid variances. The estimated angles θ_k^c can be used to compute the mixing matrix A. In this paper, we are considering the three signal, two mixture, instantaneous case and the corresponding mixing matrix for source angles θ_1^c, θ_2^c and θ_3^c is

$$\widehat{\mathbf{A}} = \begin{bmatrix} \cos(\theta_1^c) & \cos(\theta_2^c) & \cos(\theta_3^c) \\ \sin(\theta_1^c) & \sin(\theta_2^c) & \sin(\theta_3^c) \end{bmatrix}. \quad (8)$$

This estimated $\hat{\mathbf{A}}$ will be used in both DUET and SSDP to extract the source signals from the mixture signals. The scatter plot obtained using the DEMIX algorithm for a mixture scenario with source signal at $50\,^\circ$ and interfering signals at $40\,^\circ$ and $60\,^\circ$ is shown in Fig. 2. The corresponding actual $\mathbf{A}$ and estimated $\widehat{\mathbf{A}}$ are

$$\mathbf{A} = \begin{bmatrix} 0.766 & 0.643 & 0.5 \\ 0.643 & 0.766 & 0.866 \end{bmatrix}, \widehat{\mathbf{A}} = \begin{bmatrix} 0.765 & 0.644 & 0.501 \\ 0.644 & 0.765 & 0.866 \end{bmatrix}$$

It should be noted that the DEMIX algorithm is applicable in cases where the number of sources N is more than three and can be applied even in case of anechoic mixtures. The superior performance of the DEMIX algorithm for the general, source separation problem has been established in [8]. Here we have provided results to justify its performance on the data set considered in our experiments.

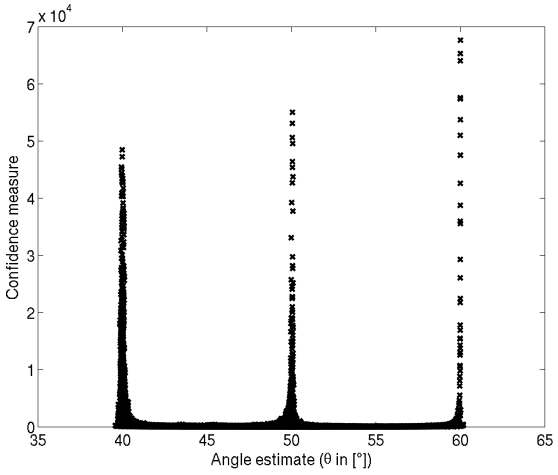

Fig. 2. Scatter plot of estimated source angles $\hat{\theta}$ and confidence measure $\widehat{\mathcal{T}}$ estimated using the DEMIX algorithm, for three sources at $40\,^\circ$, $50\,^\circ$, and $60\,^\circ$. The main speaker is positioned at $50\,^\circ$.

B. DUET for source separation

The principle behind DUET [4] is that it is possible to blindly separate an arbitrary number of sources given just two anechoic mixtures provided, the TF representations of the sources do not overlap much, which is true for speech. DUET assumes that the sources are sparse and hence already separated in TF domain.

The DUET algorithm can separate N sources from two mixtures. This method uses a two stage separation process, wherein it first estimates the mixing parameters corresponding to the sources, and then follows it up with a separation procedure. It assumes W-disjoint orthogonality (WDO) [9] of speech in the TF domain, i.e., speech signals occupy disjoint supports in the T-F domain shown as

$$S_i(t,f)S_j(t,f) = 0, \quad \forall i \neq j, \forall t, f. \tag{9}$$

where $S_i(t,f)$ and $S_j(t,f)$ represent TF domain representation of two source signals, obtained using a windowed STFT. Alternatively, every TF point has a contribution from only one source. Without loss of generality, if we include the mixing parameters of one of the mixtures into the definition of the sources, we have

$$\begin{bmatrix} X_1(t,f) \\ X_2(t,f) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_N \end{bmatrix} \begin{bmatrix} S_1[t,f] \\ \vdots \\ S_N[t,f] \end{bmatrix}. \tag{10}$$

At a TF point where the i-th source is active, we have

$$a_i = \left| \frac{X_2[t,f]}{X_1[t,f]} \right| \quad \delta_i = \angle \frac{X_2[t,f]}{X_1[t,f]},$$

as the attenuation and delay parameters. We do this for all (t,f) points. Clustering these values we obtain a two dimensional histogram as a function of a and δ. The number of dominant peaks would give an estimate of the number of sources and the peak centres would give an estimate of the corresponding a_i and δ_i. The sources can then be separated by applying a binary mask on the TF representation of any one of the mixtures, which retains only those TF points where the corresponding source is dominant, followed by an Inverse STFT operation.

The performance of DUET is often limited by the clustering technique used for mixing matrix estimation. If the sources are located very close to each other spatially, it is not possible to obtain accurate mixing parameters using this approach. Also, the clustering algorithm can estimate the delay parameters accurately only if they are within one sample time delay. This criterion restricts the separation between the microphones [4]. Hence, in this work we use the DEMIX algorithm discussed in Sec. II-A to estimate the mixing matrix and perform source separation using the binary mask as in the standard DUET. In the next section, we describe the SSDP technique ([5], [6]) where the number of active sources in a TF point can be equal to the number of mixtures or more.

C. SSDP algorithm for source separation

Consider the source separation problem in TF domain as shown in (5), and repeated here

$$\mathbf{X}(t,f) = \mathbf{A}\mathbf{S}(t,f),$$

where the objective is to recover the source STFT coefficients $\mathbf{S}$ from the mixture STFT $\mathbf{X}$. Here, we consider the instantaneous case and the mixing matrix $\mathbf{A}$ is independent of frequency and is real. The DUET algorithm discussed in Sec. II-B relies on WDO and hence assumes only one source to be active in a TF point. Further it is unable to separate sources that lie spatially close. These two limitations are overcome, by the time-frequency domain statistically sparse decomposition principle (SSDP-TF) algorithm presented in [6], which is an improvement over the time domain SSDP presented in [5]. Here, we provide a brief description of the SSDP-TF which is used for source separation in the case where two mixtures are available. We have used the Matlab implementation available at [10] and described in [6].

The SSDP-TF algorithm assumes that an estimate of mixing matrix $\mathbf{A}$ is available. The algorithm can recover a maximum number of active sources equal to the number of mixtures in each time-frequency point (i.e., two in the case considered here). An improved algorithm that relies on a local Gaussian model (SSDP-LGM) of source coefficients in the TF domain and can recover more sources is also presented in [6], but is not discussed here. The SSDP-TF method computes the covariance matrix of the mixture signal $\mathbf{X}(t,f)$ over a neighborhood of each (t,f) as

$$\hat{\mathbf{R}}_{XX}(t,f) = \frac{1}{\sum\limits_{t',f'} w(\Omega_{t',f'})} \sum_{t',f'} w(\Omega_{t',f'}) \mathbf{X}(t',f') \mathbf{X}(t',f')^H \tag{11}$$

where $w(\Omega_{t',f'})$ is a bi-dimensional window centered at (t,f). If we assume sources j_1 and j_2 are active in a TF point, the covariance matrix of the sources $\hat{\mathbf{R}}_{s_{j_1}s_{j_2}}$ and the mixture covariance matrix are related as

$$\widehat{\mathbf{R}}_{XX} = \mathbf{A}_{j_i j_2} \widehat{\mathbf{R}}_{s_{j_1}s_{j_2}} \mathbf{A}_{j1j2}^T, \tag{12}$$

where $\mathbf{A}_{j_1 j_2}$ is the 2×2 mixing matrix whose columns are j_1, j_2 columns of the initial mixing matrix $\mathbf{A}$. In the general case, where there are N sources, $\mathbf{A}$ will be a $2 \times N$ matrix. The objective in SSDP-TF is to find the two best columns of $\mathbf{A}$ that will diagonalize the source covariance matrix $\widehat{\mathbf{R}}_{s_{j_1}s_{j_2}}$. From (12), we have

$$\widehat{\mathbf{R}}_{s_{j1}s_{j2}} = \mathbf{A}_{j_1 j_2}^{-1} \widehat{\mathbf{R}}_{XX} (\mathbf{A}_{j_1 j_2}^{-1})^T = E \begin{bmatrix} s_{j_1}^2 & s_{j_1}s_{j_2} \\ s_{j_2}s_{j_1} & s_{j_2}^2 \end{bmatrix}. \tag{13}$$

The best pair of active sources can be obtained by finding the columns j_1, j_2 of $\mathbf{A}$ by [6],

$$[\widehat{j_1}, \widehat{j_2}] = \underset{j_1,j_2}{\operatorname{argmin}} \frac{\left| \hat{R}_{s_{j1}s_{j2}} \right|}{\sqrt{\hat{R}_{s_{j1}s_{j1}} \hat{R}_{s_{j2}s_{j2}}}} \tag{14}$$

where $\hat{R}_{s_{j_k}s_{j_l}}$ is the (k,l) element of $\widehat{\mathbf{R}}_{s_{j_1}s_{j_2}}$. This estimated j_1 and j_2 can be used to obtain the corresponding $\mathbf{A}_{j_1 j_2}$ and the source STFT coefficients can be estimated as,

$$\begin{cases} \widehat{\mathbf{S}}_{\widehat{j_1}\widehat{j_2}}(t,f) = \mathbf{A}_{j_1 j_2}^{-1} \mathbf{X}(t,f) \\ \widehat{\mathbf{S}}_j(t,f) = 0 \qquad\qquad \text{for all} j \notin \left\{ \widehat{j_1}, \widehat{j_2} \right\}. \end{cases} \tag{15}$$

The source signals $s_{j_1}(t)$ and $s_{j_2}(t)$ can be obtained from $\widehat{\mathbf{S}}_{\widehat{j_1}\widehat{j_2}}(t,f)$ by performing an inverse STFT. More details for

finding three or more sources (SSDP-LGM) and analysis of this can be found in [6]. Similar arguments for a time-domain SSDP approach can be found in [5].

D. Source speaker enhancement

The foreground speech or speaker enhancement for the scenario described in Fig. 1 was achieved using the algorithms described in this section. This enhancement was performed in two steps. The first step was to estimate the mixing matrix $\hat{\mathbf{A}}$ using the DEMIX algorithm described in Sec. II-A. The next step was to use either the DUET algorithm (Sec. II-B) or the SSDP algorithm (Sec. II-C). The output or processed signals from the source separation step was then fed to an ASR system to evaluate the enhancement. The experimental setup, quality of enhancement obtained for various signal-to-noise ratios (SNRs) and scenarios are presented next.

III. Experimental Results and Discussion

The testing dataset is created by simulating two-channel mixtures of commodity utterances by system users, as recorded at the interactive voice response (IVR) server of the speech-based access system, and independently collected continuous background speech at selected SNRs. The continuous speech comprises Marathi sentences recorded at the same IVR server by native callers reading out sentence prompts from text material provided to them. The mixtures thus simulate the situation of a system user querying the system for a commodity price in a background of independent speakers. The mixture parameters are chosen to depict various angular separations between the caller and the background speaker. Five commodity words uttered by each speaker are concatenated into a single utterance and input to the ASR system under various conditions of clean speech and noisy speech at various SNRs.

All the utterances were sampled at frequency of 8 kHz and normalized to an amplitude of $\pm$ 0.5 units so as to avoid any clipping, during the mixing process. One main speaker utterance (five words) and two noise speaker utterances (continuous speech) were mixed instantaneously, so as to distribute the speakers spatially. The main speaker was positioned at 10° and the two interfering speakers at other positions. The two positions of the interfering speakers were chosen from the five spatial positions of $-70, -50, -30, -10, 30$ degrees, respectively. These 10 mixing conditions require 50 words to be recognised. This scenario was repeated for 10 distinct, main speakers giving a total of 500 noisy words for speech recognition tests.

The use of continuous Marathi speech for noise sources ensured a maximum overlap of the main speaker and the noise speakers, in the time domain. All the aforementioned mixing conditions were simulated at SNRs of 0 dB and 2 dB by suitably scaling the noise sentences before instantaneous mixing. Note that, the system we propose in Fig. 1, performs a blind separation of the sources at the user end, before transmission. On the other hand, the simulations performed involve a BSS operation on speech signals obtained after transmission over the channel. The transmission cables are typically known to have a low-pass character, thus affecting the speaker and the noise utterances. These can be compensated for, although, such compensation has not been used in the current experimentation scheme.

A. Speech recognition system

The CMU Sphinx ASR system was used for testing. The system was trained on the speech of 1500 speakers across Maharashtra uttering *commodity*, *mandi*, and *district* names. The system uses MFCC features, 16 mixture diagonal covariance GMMs and 5-state HMMs for triphones.The ASR system uses 80 phones classified into 69 basic phones and 11 fillers respectively and a vocabulary of approximately 2500 words. A trigram language model trained on the transcriptions of the training data was employed. Out of vocabulary (OOV) words were not considered for experimentation.

B. Results and discussion

The BSS performance can be studied from the perception of separated speech in terms of mean opinion scores and PESQ scores. However, given the current application, we compare the speech recognition accuracies obtained without and with BSS to demonstrate the enhancement in speech obtained for the agricultural price information system. The location of the main speaker is assumed to be known. Thus, the angle estimates obtained from DEMIX can be used to identify the main speaker utterance. Speech recognition was performed on the separated main speaker utterance obtained from two BSS algorithms and compared with the recognition results for clean-speech (main speaker utterance) and unseparated noisy-speech. Recognition results for $N = 3$ and SNRs of 0 dB and 2 dB are provided in Tables I and II. Noisy speech obtains very poor recognition accuracy compared to clean speech, since the acoustic models used in the ASR were trained on clean speech, and the degraded speech is poorly matched to the training data. The ASR performance worsens with decreasing SNR. Pre-processing by the SSDP methods leads to a significant increase in recognition accuracy. Among the two algorithms, SSDP-LGM and SSDP-TF, as expected the SSDP-LGM performs slightly better than the SSDP-TF algorithm. We note that the accuracy is maintained at 60 % level across SNRs whereas the clean speech ASR accuracy is 84 %. This suggests that the source-separated speech achieves a uniform noise suppression at all SNRs based on exploitation of spatial diversity. The accuracy limitation arises from the artifacts / change in speech quality. This can probably be overcome if the acoustic models used in the ASR system are trained on source-separated speech. From the results shown in Sec. II-A, the mixing matrix estimates were reasonably accurate. This indicates that a better recognition can be achieved by improving the source separation algorithms.

In addition, we also performed similar experiments for the $N = 2$ case (i.e., one main speaker and one interfering speaker). The dataset used in the $N = 2$ case was similar, but had 20 distinct speakers and one interfering speaker. Here we provide results for one such setup (SNR = 2 dB) in Table III.

TABLE I
SPEECH RECOGNITION RESULTS WITHOUT AND WITH SOURCE SEPARATION FOR 0 DB SNR AND THREE SIGNALS ($N = 3$ FOR SOURCE SEPARATION).

Case	Substitution (%)	Deletions (%)	Insertions (%)	Error (%)	Correct (%)	Word Accuracy (%)
Clean-speech	12.0	0.0	4.0	16.0	88.0	84.0
Mixtures without separation	16.6	68.6	0.4	83.4	16.8	16.4
SSDP-LGM	17.0	17.8	3.4	38.2	65.3	61.8
SSDP-TF	13.3	25.5	2.4	41.2	61.2	58.8

TABLE II
SPEECH RECOGNITION RESULTS WITHOUT AND WITH SOURCE SEPARATION FOR 2 DB SNR AND THREE SIGNALS ($N = 3$ FOR SOURCE SEPARATION).

Case	Substitution (%)	Deletions (%)	Insertions (%)	Error (%)	Correct (%)	Word Accuracy (%)
Clean-speech	12.0	0.0	4.0	16.0	88.0	84.0
Mixtures without separation	13.6	65.2	0.8	79.6	21.2	20.4
SSDP-LGM	15.7	18.6	3.7	38.0	65.7	62.0
SSDP-TF	12.7	23.9	2.7	39.2	63.5	60.8

TABLE III
SPEECH RECOGNITION RESULTS WITHOUT AND WITH SOURCE SEPARATION FOR 2 DB SNR AND TWO SIGNALS ($N = 2$ FOR SOURCE SEPARATION).

Case	Substitution (%)	Deletions (%)	Insertions (%)	Error (%)	Correct (%)	Word Accuracy (%)
Clean-speech	13.0	0.0	5.0	18.0	87.0	82.0
Mixtures without separation	20.5	26.0	6.7	53.2	53.5	46.8
SSDP-LGM	10.5	3.2	5.1	18.8	86.3	81.2
SSDP-TF	10.4	3.2	5.1	18.7	86.4	81.3
DUET	13	12.4	2.6	28	74.6	72.0

It can be seen that the word accuracy obtained using the SSDP algorithms is at the 81 % level, which is close to clean speech ASR accuracy of 82 %. This is expected, given that the DEMIX algorithm provides a good estimate of the mixing matrix $\mathbf{A}$ of size 2×2 and is an evenly-determined source separation problem. However, the 72 % accuracy obtained using DUET is lower than that of the SSDP methods, as it relies on the assumption of a single active source per TF point.

IV. CONCLUSION

In this paper, we proposed a system that enhances foreground speech in a mobile phone setup for a speech-based agricultural price information system. This system uses blind source separation to exploit the difference in spatial locations of the main and interfering speakers via stereo microphones and enhance the relevant source signal using a pair of microphones. Our simulation results suggest that such an approach leads to improved speech recognition accuracy, with still further scope for improvement. As part of future research, we intend to consider characteristics of speech signals to improve source separation and speech recognition in challenging scenarios.

ACKNOWLEDGMENT

We would like to acknowledge Hitesh Tulsiani, DAP lab, IIT Bombay for helping with the ASR part.

REFERENCES

[1] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition," *IEEE Signal Processing Magazine*, pp. 127–140, Nov. 2012.

[2] S. Takada, T. Ogawa, K. Akagiri, and T. Kobayashi, "Speech enhancement using square microphone array for mobile devices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 313–316.

[3] T. Godambe and K. Samudravijaya, "Speech data acquisition for voice based agricultural information retrieval," in *Proc. Of 39th All India DLA Conference*, 2011.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[5] M. Xiao, S. Xie, and Y. Fu, "A statistically sparse decomposition principle for underdetermined blind source separation," in *Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2005, pp. 165–168.

[6] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, 2009, pp. 775–782.

[7] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. Intl. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006) LNCS 3889*, 2006, pp. 536–543.

[8] ——, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 121–133, Jan. 2010.

[9] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2002, pp. 529–532.

[10] E. Vincent and N. Q. K. Duong. (2009). [Online]. Available: http://www.loria.fr/~evincent/soft.html