Selective Tap Training of FIR filters for Blind Source Separation of Convolutive Speech Mixtures

Ali Khanagha, Vahid Khanagha

Selective Tap Training of FIR filters for Blind Source Separation of Convolutive Speech Mixtures

Ali Khanagha
Amirkabir University of Technology
Tehran, Iran
khanaghaa@ripi.ir

Vahid Khanagha
Iran University of Science and Technology
Tehran, Iran
vkhanagha@ee.iust.ac.ir

Abstract—**This Paper presents a novel low complexity time domain algorithm for blind separation of speech signal from their convolutive mixtures. We try to reduce intrinsic computational complexity of time domain algorithms by adapting only a small subset of taps from separating FIR filters which are expected to attain largest values. This selection is accomplished by recovering spatial dependencies using Linear Prediction (LP) analysis. Then we use Particle Swarm Optimization (PSO) in order to find best values for these selected taps. We employ the sparseness properties of speech signals in the Time-Frequency (TF) domain to define a low complexity and yet appropriate fitness function which numerically quantifies the amount of achieved separation by each one of the particles during PSO execution.**

I. Introduction

Blind Source Separation (BSS) has widely received attentions since it was the common problem for almost every multi sensory application. BSS consists in separation of N unknown, yet statistically independent sources, $\mathrm{s(n)} = [s_1(\mathrm{n}), \ldots, s_\mathrm{N}(\mathrm{n})]^\mathrm{T}$, from M observed mixtures of these sources, $\mathrm{x(n)} = [x_1(\mathrm{n}), \ldots, x_\mathrm{M}(\mathrm{n})]^\mathrm{T}$, where the mixture parameters are unknown. This would be a challenging task, especially in real world scenario when each sensor receives multiple copies of each source due to the presence of reverberation phenomena. This is often the case for speech signals in typical echoic acoustic environments. The relation between inputs and outputs of such convolutive mixing system can be written as

$$x_i(k) = \sum_{j=1}^{N} \sum_{l=-\infty}^{\infty} h_{ij}(k-l) s_j(l) \qquad (1)$$

Where $h_{ij}(k)$ is the fir impulse response between jth source to the ith sensor. Several methods have been proposed for solving this problem. A comprehensive review of these techniques can be found in [1]. Here we consider time domain BSS algorithms in which, Independent Component Analysis (ICA) is directly applied to mixtures and unmixing FIR filters are directly estimated. They have the advantage of possible convergence near the optimal point and inversion of mixing system but they suffer from computational complexity of adaptation of very long FIR filters[3]. Also time domain methods are prone to the so called whitening effect; since most of these algorithms were developed for i.i.d sources, they tend to reduce temporal dependencies as well as spatial dependencies[2]. This might be a problem for speech signals with their special temporal structure.

In this paper, we present a novel time domain algorithm which proceeds in two steps. First, we identify potentially important taps of unmixing FIR filters by identifying spatial dependencies using Linear Prediction (LP) analysis. Then we use Particle Swarm Optimization (PSO) technique in order to find best values for these selected taps. PSO is a stochastic, population-based evolutionary algorithm for problem solving in complex multidimensional parameter spaces. The main challenge of PSO is to find an appropriate fitness function with low computational complexity and yet with a global extremum at the true separation point. We define a novel cost function based on TF domain sparseness properties of speech signals.

II. Identification of Spatial Dependencies

Fig. 1. Shows a typical block diagram for mixing and unmixing systems, in which h_{ij} and w_{ij} are stand for the impulse response of corresponding FIR filters. It is easy to see that the prefect point of separation is met when following relations hold:

$$h_{11} * w_{21} = -h_{12} * w_{22}$$

$$w_{11} = \alpha_1 h_{22} \qquad w_{12} = -\alpha_1 h_{21} \qquad (2)$$

$$w_{21} = -\alpha_2 h_{12} \qquad w_{22} = \alpha_2 h_{11} \qquad (3)$$

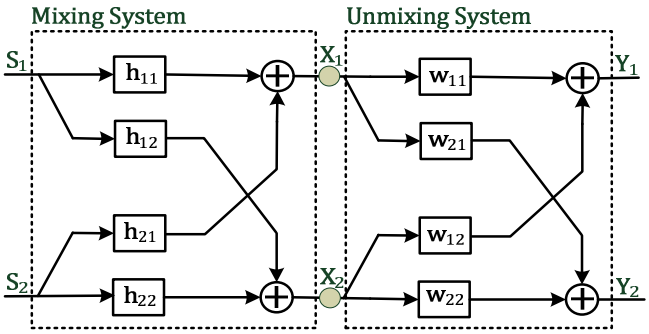

Fig. 1. Block diagram of mixing and unmixing systems.

Keeping these equations in mind, along with the fact that mixing FIR filters are the causes for spatial dependencies between observed mixtures, brings the possibility of attaining

a general knowledge about unmixing system by studying these spatial dependencies.

For example, if we observe that one mixture has a strong dependency with p lag delayed version of the other mixture, we may expect that corresponding mixing FIR filter, had a large value on its pth tap. For now, we leave the discussion for finding the value of this tap for section III and focus on how to extract the structure of these dependencies from speech mixtures.

Because of the temporal structure of speech signals, in which every sample of each speech signal has strong dependencies with its adjacent samples, it is useless to observe simple cross correlation between mixtures. For example, Fig. 2, shows the resulting covariance matrix for a very simple anechoic [1] mixture, calculated from raw observations. This covariance matrix has a special sylvester structure which corresponds with a special sylvester structure used in [3]. There exists four distinctive blocks corresponding to each one of four FIR filters of mixing system. In each block, parallel slant lines are representatives for taps of FIR filters (lower triangular section of each block must be studied for each filter). In Fig. 2, Although the spatial dependencies are observable as two thick slant lines on non diagonal blocks, but the exact position of this dependency is non distinctive. Specially, in more realistic mixing scenarios, where several reflections of sources to each sensor result in several slant lines, they are likely to overlap with each other and make this problem even harder.

Instead, we use LP analysis as a pre processing step to dismiss these temporal dependencies. We can generate a whitened version of mixtures using LP residuals; however, in order to preserve spatial dependencies, this prewhitening must be done in blocks of smaller length than 5 milliseconds [6]. Using these short blocks of prewhitened mixtures, we use the exact formulation of [3] to form the covariance matrix, which is now expected to solely represent spatial dependencies. Fig. 3, presents the resulting covariance matrix of prewhitened version of the same mixture. After applying some thresholding and 2D median filtering (in order to retrieve broken lines and suppress single meaningless points), the location of remaining slant lines gives us the exact location of strongest taps of mixing FIR filters.

Obviously, the real convolutive mixtures have larger orders of spatial dependencies, because they are formed by more complex FIR filters; However we can always select a small subset of strongest dependencies (strongest slant lines). Also we accept possible false detections, since we will run an optimization procedure, as will be explained in section III, which is able to find the best possible values for selected taps; in fact, we will train these selected taps to produce outputs that are as different (independent) as possible. In this way, the few falsely identified taps, would attain very small values near the separation point.

[1] each sensor receives only one copy of sources with different time delays.

III. Selective Tap Training Procedure

We use PSO as our optimization technique in order to find the most fitted values for selected taps of FIR filters.

A. PSO principles

Particle swarm optimization is a stochastic, population-based evolutionary algorithm for problem solving in complex multidimensional parameter spaces. It is a kind of swarm intelligence that is based on social-psychological principles.

A multidimensional optimization problem is given, along with an objective function to evaluate the fitness of each candidate point in parameter space; The swarm is typically modeled by particles in this multidimensional space that have a position and a velocity. After the definition of a random population of individuals (particles) as candidate solutions, they fly through hyperspace of parameters with the aid of two essential reasoning capabilities: their memory of their own best local position and knowledge of the global or their neighborhood's best [5].

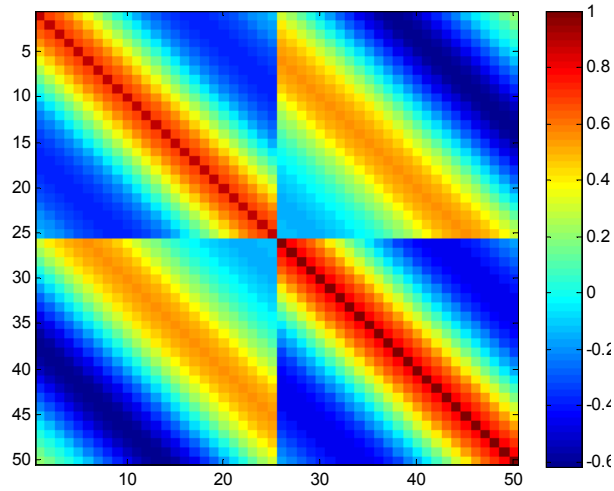

Fig. 2. Covariance matrix computed from raw mixtures of an anechoic mixture. The exact location of thick slant lines aren't distinctive enough to carry information about maximum valued taps of mixing system.

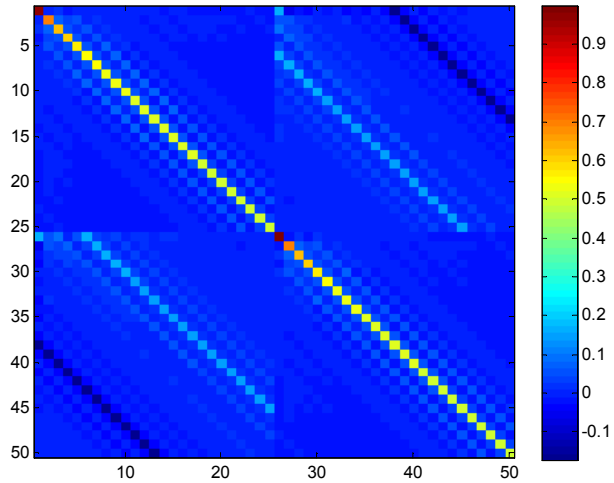

Fig. 2. Covariance matrix of prewhitened mixtures of an anechoic mixture. The exact locations of strong spatial dependencies are easy to extract.

PSO begins by initializing a random swarm of N_p particles $p_i = [p_{i1}, \dots, p_{iL}]$, each having L parameters. At each iteration, the fitness of each particle is evaluated according to selected fitness function. The most fit experienced position of each particle is stored and progressively replaced as $pbest_i, i = 1, \dots, N_p$ along with a single most globally fit particle ($gbest$) as fitter locations are encountered during algorithm iterations. The parameters of each particle in the swarm are updated at each iteration (n) according to a velocity vector as [6]:

$$\overline{vel_\iota}(n) = \omega \times \overline{vel_\iota(n-1)} + acc_1 \times diag[e_1, e_2, \dots, e_L] \times \left(gbest - p_i(n-1)\right) + acc_2 \times diag[e_1, e_2, \dots, e_L] \times \left(pbest_i - p_i(n-1)\right) \quad (4)$$

$$p_i(n) = p_i(n-1) + vel_i(n) \quad (5)$$

Here, $\overline{vel_\iota}(n)$ is the velocity vector of particle i, $e_r \in (0,1)$ is a random value, acc_1 *and* acc_2 are acceleration coefficients toward $gbest$ and $pbest_i$ and ω is the inertia weight.

In fact, the trajectory of each particle is determined by random superposition of its previous velocity with the location of local and global best particles found by far. As new *gbests* are encountered during the update process, all other particles begin to swarm toward this new $gbest$, continuing their random search along the way. The optimization is terminated when all of the particles have converged to $gbest$ or a sufficient condition of fitness function met.

The advantages of this optimization procedure for our case is its independence from the structure of underlying process. It doesn't need fine parameter selections and it doesn't require to consider source statistics in its update equations (in the form of probability distribution function (pdf) estimations).

B. Implementation of PSO for selective tap training

The most challenging task in implementing PSO for any optimization problem is to define a well behaved objective function which has a global extremum at the desired point. However, the computational complexity of such an objective function must be low enough since it must be calculated for large population of particles in every iteration of algorithm.

In our case, we need to define a cost function which quantifies the amount of achieved independence between two output signals, with current tap values. This is a challenging task for a wideband signal such as speech. It is apparent, that cross correlation metrics might result in self decorrelation instead of spatial decorrelation. Other measures of independence, such as those based on higher order statistics are too complicated to calculate for all of the particles. Also they are mostly developed for i.i.d sources and again they might result in temporal decorrelation. Here we define a novel, low complexity fitness function, which effectively measures the amount of achieved independence.

Our fitness function is based on sparsness properties of speech signals in Time-Frequency (TF) domain. A signal is said to be sparse in a given basis, when it is zero or nearly zero more than might be expected from its variance [7]. Speech signals are known to have such sparse representation in TF domain. This sparsness of speech signals, which is caused by different system properties of speakers and different sentences in their speech, causes the TF domain distribution of two different speech signals to be mostly disjoint (i.e. there is only one active source in most of given TF points) [7]. Hence, it seems reasonable to define a cost function that measures the amount of disjointness in TF domain and trying to maximize it.

Fig. 3, demonstrates the superimposed TF distribution of two pure speech signals, after applying an appropriate threshold. There, white color corresponds to TF overlaps and black and gray colors represent the points where only one source is active. It can be seen clearly that two sources rarely overlap with each other. In contrast, Fig. 4 shows the TF distributions of two observed measured mixtures, which almost completely overlap with each other.

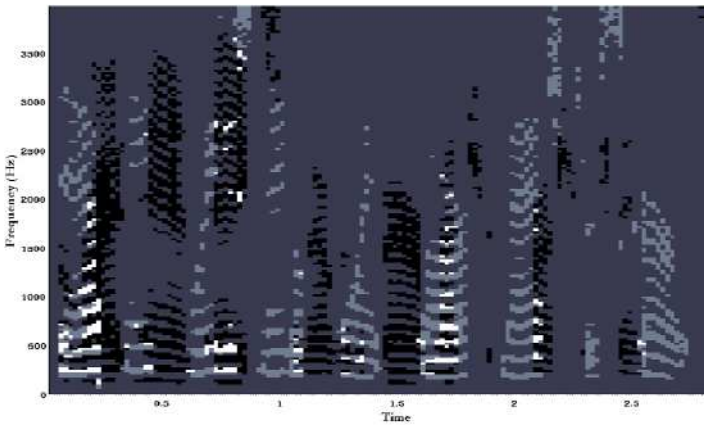

Fig. 3. TF distribution of two pure speech signals. The two sources are almost completely disjoint.

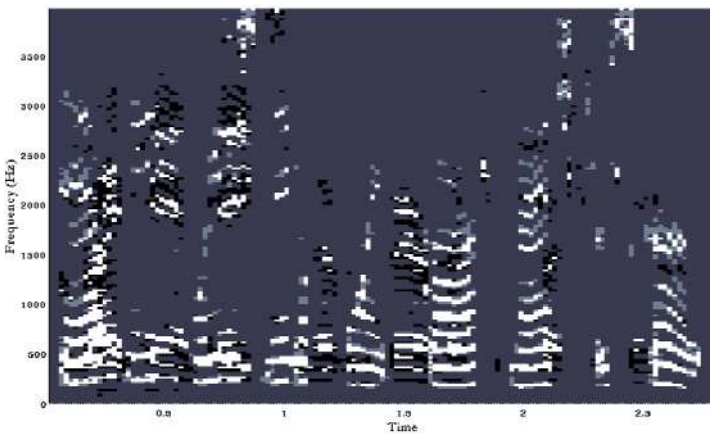

Fig. 4. TF distribution of two observed mixtures of speech. Two mixtures are almost completely non disjoint.

Clearly, the maximization of TF domain disjointness is equal to minimization of the overlap between high power TF points of two outputs. But we must be aware of the possible frequency selectivity of separating filters. In fact, when we

apply such measure to the whole spectrogram, it is very possible that we achieve two low pass and high pass signals (*i.e.* the TF distribution of outputs are non overlapping, but only in frequency domain). Although our incompletely defined fitness function is minimized and sources are disjoint in TF domain but they are not representatives of two separated sources. We cope this possible divergence of our fitness function by partitioning the spectrogram to several smaller windows and evaluating the fitness function in these smaller windows with the condition of the presence of both outputs in each window. In other word, if the amount of energy in each window is higher than a threshold for both outputs, we may calculate fitness of that window and sum it up with total fitness function. So we are sure that if we met disjointness in a given window, since both outputs were present there, no frequency selectivity has been occurred.

In this way, at the beginning, since the outputs are not separated yet, there exist many valid windows with large amount of TF overlap. Later, as the optimization process proceeds, the amount of overlaps are decreased and probably the number of valid windows degrades and the total fitness function becomes less and less. The selected window must be able to cover at least a vowel, since it must cover the simultaneous activities of two different speakers. So, it is reasonable to choose a time duration of 0.3 seconds for this window. The frequency width of window is chosen to be around 40-50 Hz; wide enough to cover at least one harmonic activity from each active source and narrow enough in order to avoid possible frequency selectivity inside the window (the frequency resolution of separating filters must be considered, which is directly proportional with the length of the FIR filter).

Finally, it should be noted that since we selected a small subset of taps of FIR filters for training, we have already abandoned many possible divergences at the first place. We have selected the taps that are likely to be most effective taps in separation task. It's the key feature of our proposed algorithm which guarantees fast convergence of particles of PSO around separation point. If we choose to train all of the taps of FIR filters, then our fitness function wouldn't be able to optimize them because of the strong possibility of divergence around temporal whitening point.

IV. Simulation results

In this section we present simulation results for proposed algorithm. First, we study the results for the proposed BSS method and then we compare the feasibility of using each feature set for resolving permutation ambiguity, on a large database of different speech sources.

we compare performance of the proposed BSS method with that of [6]. A 2×2 transfer function $\mathrm{H}(\boldsymbol{z})$ was chosen as

$$\begin{bmatrix} 1 - .7z^{-25} + .3z^{-30} - .1z^{-66} & .5z^{-23} - .3z^{-54} + .3z^{-66} - .1z^{-88} \\ .6z^{-18} - 0.4z^{-28} + .3z^{-66} & 1 - .6z^{-20} + .4z^{-88} + .1z^{-98} \end{bmatrix}$$

The measure of performance will be Signal-to-Interference-Ration (SIR) which is the ratio between the energy of desired signal and the cross-talk between channels, *i.e.*,

$$SIR = 10 \log\left(\frac{\|target\ source\|^2}{\|interfering\ source\|^2}\right) \quad (6)$$

Since the proposed method is based on the TF domain distribution of sources, which is strongly dependent to the speaker and the uttered speech, it is necessary to evaluate the suitability of our fitness function for a large number of speech signals. Therefore, we have used a database of 30 different speech signals (15 male speakers versus 15 female speakers) of length 3 seconds, to form 200 different combinations of sources. We use a population of 100 particles and choose 4 taps from each FIR filter for training. In other word, we have 12 parameters to be optimized during the PSO. The maximum number of iterations is set to 50 and the order of LP analysis filters is 15.

Table I, summarizes the average SIR results for this 200 different combinations of sources which demonstrates the slight improvement in separation quality compared to [6].

TABLE I
AVERAGE SIR IMPROVEMENT FOR 200 DIFFERENT COMBINATIONS OF SPEECH SIGNALS.

Algorithm	SIR 1	SIR 2	Average
Selective training	11.33	13.55	12.41
SOS [2]	10.21	11.35	10.78

Also, another experiment is done in order to test the feasibility of proposed method for identifying spatial dependencies. Thus, 50 different impulse responses, which are more complex than the one used in previous experiment are employed and the number of false detections in identifying the three most strongest taps of each of 4 sub channels are counted. From this $50 \times 12 = 600$ detections, only 67 false detections are observed, which proves the fidelity of proposed method in 88% of detections.

Beside the reduced computational complexity, the most important feature of the proposed method is its ability to train the whole four FIR filters shown in Fig. 1. Most of the gradient based time domain algorithms for speech signals, like the one in [6], are not able to train w_{11} and w_{22} because of the structure of their update equation. The cost function used in [6] is based on block diagonalization of Correlation matrix and concentrates on the cancellation of mutual spatial dependencies between two outputs. This method is incapable of reducing self dependencies of each output.

Instead, our proposed method is able to train all four FIR filters with its evolutional intelligence and based on equations (1) and (2), exhibits very good performance in channel identification. We use normalized mean squared error (NMSE) of channel identification as another performance index as:

$$NMSE = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{.5 \times \sum_{l=1}^{L}\left[\sum_{\tau=1}^{L}\left(h_{ij}^{l}(\tau) - w_{ij}(\tau)\right)^2\right]}{\sum_{\tau=1}^{L}\left(w_{ij}(\tau)\right)^2} \quad (7)$$

Table II, summarizes the NMSE for 200 different combinations of sources which demonstrates significant improvement in identification quality compared to [6].

TABLE III
AVERAGE NMSE IMPROVEMENT FOR 200 DIFFERENT COMBINATIONS OF SPEECH SIGNALS.

Algorithm	NMSE 1	NMSE 2	Average
Selective training	-5	-5.45	-5.2
SOS [2]	-0.2	-0.3	-0.25

Also, with the selection of this small subset of taps, which are truly responsible for spatial dependencies, we can successfully avoid the possible temporal whitening effect of regular time domain algorithms.The other strong aspect of the proposed method is that it is needless for fine parameter justifications (such as adaptive step parameter of update equation in[6]) or identification of statistical properties of sources (like the pdf estimation in information theoretic based methods).

V. CONCLUSION

We Proposed a novel method for the BSS of convolutive mixtures which is based on extracting spatial dependencies and direct concentration on removing them. We used PSO as optimization process, which doesn't require fine parameter justifications and pdf estimations. The most challenging aspect of implementing PSO was to define an appropriate cost function for its evaluations which was accomplished using TF domain sparsity properties of speech signals. Simulation results demonstrated the slight improvement in separation quality. Beside the reduced computational complexity, the ability to train all of the 4 unmixing FIR filters simultaneously and its reliability against temporal whitening effect was the main promises of the proposed method. It also demonstrates about 5 dB improvement in NMSE of channel identification.

VI. REFERENCES

[1] M. Syskind Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in Springer Handbook on Speech Processing and Speech Communication, 2006.

[2] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," IEEE Trans. Audio, Speech, Lang. Process., Vol. 14, No. 1, pp. 200-212, Jan. 2006.

[3] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Processing, 13(1):120–134, Jan. 2005.

[4] Jitendra K.Tugnait, "On Blind Separation Of Convolutive mixtures of Independent Linear Signals in Unknown Additive noise" IEEE TRAN. ON SIGNAL PROCESSING, VOL.46, NO.11, November 1998.

[5] D. J. Krusienski and W. K. Jenkins, "The Application of Particle Swarm Optimization To Adaptive IIR Phase Equalizer", ICASSP 2004.

[6] Swagatam Das and Ajith Abraham, "Synergy of Particle Swarm Optimization with Evolutionary Algorithms for Intelligent Search and Optimization", In IEEE International Congress on Evolutionary Computation, Vol. 1 (2006), 84-88.

[7] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," IJIST, vol. 15, pp. 18–33, 2005.