A SUBSPACE APPROACH TO SINGLE CHANNEL SIGNAL SEPARATION USING MAXIMUM LIKELIHOOD WEIGHTING FILTERS

Gil-Jin Jang[1], *Te-Won Lee*[2], *and Yung-Hwan Oh*[1]

[1]Spoken Language Laboratory, CS Division, KAIST, Daejon 305-701, South Korea

[2]Institute for Neural Computation, UCSD, La Jolla, CA, USA

{jangbal,yhoh}@speech.kaist.ac.kr, tewon@ucsd.edu
http://speech.kaist.ac.kr/~jangbal

ABSTRACT

The goal of this work is to extract multiple source signals when only a single channel observation is available. We propose a new signal separation algorithm based on a subspace decomposition. The observation is transformed into subspaces of interest with different sets of basis functions. A flexible model for density estimation allows an accurate modeling of the distributions of the source signals in the subspaces, and we develop a filtering technique using a maximum likelihood (ML) approach to match the observed single channel data with the decomposition. Our experimental results show good separation performance on simulated mixtures of two music signals as well as two voice signals.

1. INTRODUCTION

Extracting multiple source signals from a single channel mixture is a challenging research field with numerous applications. Various sophisticated methods have been proposed over the past few years in a research area called computational auditory scene analysis (CASA) [1]. Example proposals of CASA are auditory sound segregation models based on harmonic structures of the sounds [2], automatic tone modeling [3], and psycho-acoustic grouping rules [4]. Recently Roweis [5] presented a refiltering technique that estimates time-varying masking filters that localize sound streams in a spectro-temporal region. In his work, sources are supposedly disjoint in the spectrogram and a "mask" whose value is binary, 0 or 1, exclusively divides the mixed streams completely. This approach is, however, applicable only when these assumptions match well to the data.

Our work is motivated by this spectral masking, but free of the assumption that the spectrograms are disjoint. Its main novelty is that the masking filters can have any real value in $[0, 1]$. The algorithm recovers the original auditory streams by searching for the maximized log likelihood of the separated signals, computed by the pdfs (probability density functions) of the projections onto the subspaces. Empirical observations show that the projection histogram is extremely sparse, and the use of generalized Gaussian distributions [6] yields a good approximation. We use independent component analysis (ICA) to provide discriminative, statistically independent subspaces. The theoretical basis of this approach is **"sparse decomposition"** [7]. Sparsity in this case means that only a small number of instants in the representation differ significantly from zero. ICA maximizes the sparsity of the subspaces and hence reduces the overlap between the sources in the new coordinate.

2. METHOD

We first define the single channel separation problem, and derive the separation algorithm based on maximum likelihood (ML) estimation and the generalized Gaussian pdf modeling. Secondly, we explain how to obtain statistically independent subspaces. For simplicity we only consider the case of binary sources and 1-dimensional subspaces.

2.1. A Subspace Approach

Let us consider a monaural separation of a mixture of two signals observed in a single channel, such that the observation is given by

$$y^t = x_1^t + x_2^t\,, \quad \forall t \in [1, T]\,, \tag{1}$$

where x_i^t is the t^{th} observation of the i^{th} source. Note that superscripts indicate sample indices of time-varying signals and subscripts indicate the source identification. It is convenient to assume all the sources to have zero mean and unit variance. The goal is to recover all x_i^t given only a single sensor input y^t. The problem is too ill-conditioned to be mathematically tractable since the number of unknowns is $2 \times T$ given only T observations.

The proposed method decomposes the source signals into N disjoint subspace projections u_{ik}^t each filtered to contain only energy from a small portion of the whole space:

$$u_{ik}^t = \mathcal{P}(x_i^t; \mathbf{w}_k, d_k) = \sum_{n=1}^{N} w_{kn} x_i^{t-d_k+n}\,, \tag{2}$$

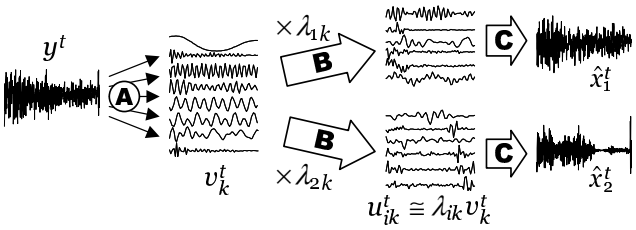


Fig. 1. Block diagram of subspace weighting. (**A**) input signal y^t is projected onto N subspace. (**B**) the projections v_i^t are modulated by weighting signals λ_{ik} (real valued between 0 and 1). (**C**) the separation process finally terminates with summing up the modulated signals.

where $\mathcal{P}$ is a projection operator, N is the number of subspaces, and w_{kn} is the n^{th} coefficient of the k^{th} coordinate vector $\mathbf{w}_k$ whose lag is d_k. In the same way we define the projection of the mixed signal as

$$v_k^t = \mathcal{P}(y^t; \mathbf{w}_k, d_k) = u_{1k}^t + u_{2k}^t \,. \tag{3}$$

Suppose the appropriate subparts of an audio signal may lie on a specific subspace over short times. The separation is then equivalent to searching for subspaces that are close to the individual source signals. More generally, u_{ik}^t are approximated by modulating the mixed projections v_k^t:

$$\begin{aligned} u_{1k}^t \cong \lambda_{1k} v_k^t, \;\; u_{2k}^t \cong \lambda_{2k} v_k^t \\ \lambda_{ik} \in [0,1], \;\; \lambda_{1k}{+}\lambda_{2k}{=}1 \,, \end{aligned} \tag{4}$$

where "latent variables" λ_{ik} are weights on the projections of subspace k, which is fixed over time. We can adapt the weights to bring projections in and out of the source as needed. The original sources x_i^t are then reconstructed by recombining $u_{i1}^t, u_{i2}^t, \ldots, u_{iN}^t$ and performing the inverse transform of the projection. Proper choices of the weights λ_{ik} enable the isolation of a single source from the input signal and the suppression of all other sources and background noises.

This approach, illustrated in fig. 1, forms the basis of many CASA approaches (e.g. [2, 4, 5]). The results of such an analysis are often displayed as a *spectrogram* that shows energy as a function of time and frequency. For example, it is possible to distinguish violin and cello sounds that are played simultaneously, based on the energy distribution in the spectrogram. The energy of a cello sound is usually concentrated on the lower bands of the spectrogram, and a violin sound is distributed in the higher bands. The projections, fig. 1-**A**, act like low- and high-pass filters in this case. Subspace weighting can also be thought of as Wiener filtering. If the original sources are known, an "optimal" filters can be computed. We might set λ_{ik} equal to the ratio of energy from one source in subspace k to the sum of energies from both sources in the same subspace.

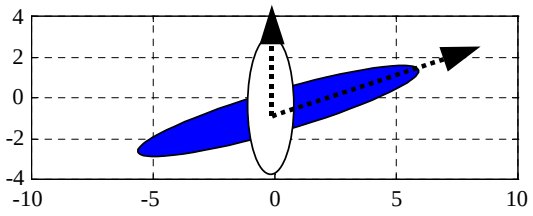


Fig. 2. Illustration of desired subspaces. The arrows show the directions of the maximum energy concentrations.

2.2. Estimation of Weighting Filters

The estimation of the weights λ_{ik} can be accomplished by simply finding the values that maximize the probability of the subspace projections. The histograms of natural sounds reveal that $p(u_{ik}^t)$ is highly super-Gaussian [8]. Therefore we use a generalized Gaussian prior [6] that provides an accurate estimate for symmetric non-Gaussian distributions by fitting the exponent q in its simplest form $p(s) \propto \exp(-|s|^q)$, by which we model $p(u_{ik}^t)$ as

$$\log p(u_{ik}^t) \propto -\left|u_{ik}^t\right|^{q_{ik}} \cong -\lambda_{ik}^{q_{ik}} |v_k^t|^{q_{ik}} \,. \tag{5}$$

We constrain that $\lambda_{ik} \in [0,1]$ and $\lambda_{1k}{+}\lambda_{2k}{=}1$. The value of λ_{2k} depends solely on λ_{1k}, so we need to consider λ_{1k} only. We define the object function Ψ_k^t of subspace k by the joint log probability density of u_{1k}^t and u_{2k}^t:

$$\begin{aligned} \Psi_k^t &\stackrel{\text{def}}{=} \log p(u_{1k}^t, u_{2k}^t) \\ &\cong \log p(u_{1k}^t) + \log p(u_{2k}^t) \\ &= -\lambda_{1k}^{q_{1k}} |v_k^t|^{q_{1k}} - (1-\lambda_{1k})^{q_{2k}} |v_k^t|^{q_{2k}} \,. \end{aligned} \tag{6}$$

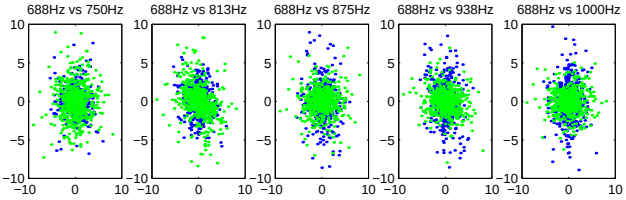
The maximum likelihood is achieved when $\partial\Psi_k^t / \partial\lambda_{1k} = 0$, calculated as

$$\frac{\partial \Psi_k^t}{\partial \lambda_{1k}} = \underbrace{\underbrace{-\overbrace{\lambda_{1k}^{q_{1k}-1}}^{\text{M.I.}} \cdot q_{1k} |v_k^t|^{q_{1k}}}_{\text{M.D.}} + \underbrace{\overbrace{(1{-}\lambda_{1k})^{q_{2k}-1}}^{\text{M.D.}} \cdot q_{2k} |v_k^t|^{q_{2k}}}_{\text{M.D.}}}_{\text{M.D.}}, \tag{7}$$

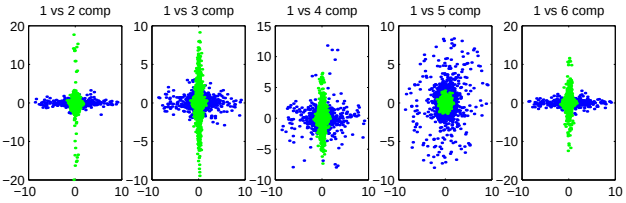
where M.D. and M.I. stand for 'monotonically decreasing [increasing]' w.r.t. λ_{1k}. Because eq. 7 is M.D. in the closed interval $[0,1]$, we can always find a unique solution by Newton's method [9].

2.3. Finding Independent Subspaces

A set of subspaces that well split target sources is essential in the success of the separation algorithm. Fig. 2 shows an example of desired subspaces. Two ellipses represent two different sources, whose energy concentrations are directed by the arrows. If we project the mixture onto the arrows (1-dim subspaces), the original sources can be recovered with the error minimized.



(a) Projected distributions of Fourier basis



(b) Projected distributions of learned ICA basis

Fig. 3. Example plots of subspace projections. Projections of music signals are dark points, and those of speech signals are bright points. (a) Fourier basis: (x, y) pair is the outputs of 6 non-overlapping bandpass filters that divide the range 625-1000Hz equally. The center frequencies of x and y axes are at the top of each 2D plot. (b) (x, y) pair is the projected values of 6 subspaces obtained by the ICA learning algorithm. The subspace numbers of x and y axes are displayed at the top of each 2D plot.

To obtain an optimal basis, we adopt ICA that estimates the inverse-translation-operator such that the resulting coordinate can be statistically as independent as possible [8]. ICA infers time-domain basis filters $\mathbf{w}_k$ generating the most probable outputs. The learned basis filters maximize the amount of information at the output, so that they constitute an efficient representation of the given sound source. Fig. 3 displays the distributions of subspace projections onto Fourier and learned ICA bases for the same data. Although the coordinates look almost uncorrelated (the distributions are spread equally in all directions), there are too much overlaps in Fourier subspaces between two signals for the weighting in eq. 4 to ever work. The ICA coordinates are not only uncorrelated but also non-overlapping; the distributions of music signals are broadened in x axis and shrunk in y axis, and those of speech signals act the other way round. At the forth column, the music distribution encloses the speech distribution, meaning that the energy of speech is very little concentrated on both subspaces in this case.

3. EVALUATION

3.1. Experiment Setup

We have tested the performance of the proposed method on single channel mixtures of four different sound types; monaural signals of rock and jazz music, male and female speech. We used different sets of speech signals for training the generalized Gaussian model parameters and for generating the mixtures. For the mixture generation, two sentences of the target speakers 'mcpm0' and 'fdaw0', one for each speaker, were selected from the TIMIT speech database. The training sets were designed to have 21 sentences for each gender, 3 each from 7 randomly chosen males and 7 randomly chosen females. Half of the music sound was used for training, half for generating mixtures. All signals were downsampled to 8kHz, from original 44.1kHz (music) and 16kHz (speech). Audio files for all the experiments are accessible at the website[1].

The proposed method deals with binary mixtures only. All the possible pairs are {(R,J), (R,M), (R,F), (J,M), (J,F), (M,F)}, where the symbols R, J, M, F stand for rock and jazz music, male and female speech. We provide three different types of bases as listed in Table 1. For the (R, J) and (M, F) mixtures, $\mathbf{W}_{\mathrm{m}}$ and $\mathbf{W}_{\mathrm{s}}$ are used. $\mathbf{W}_{\mathrm{ms}}$ is adopted in the other cases, because the input mixtures contain both music and speech signals. The weighting filters are computed block-wise, that is, we chop the input signals into blocks of fixed length and assign different weighting filters for the individual blocks. The computation of the weighting filter at each block is done independently on the other blocks, so the weighting becomes more accurate as the block length shrinks. However if the block length is too short, the computation becomes unreliable. We performed separation experiments with varying the block length to find the optimal length.

3.2. Experimental Results

We generated a synthesized mixture by selecting two sources out of the four and simply adding them. The proposed separation algorithm was applied to recover the original sources.

[1] `http://speech.kaist.ac.kr/~jangbal/rbss1/`

Table 1. ICA bases.

basis	description	training data
$\mathbf{W}_{\mathrm{m}}$	music basis	(R, J)
$\mathbf{W}_{\mathrm{s}}$	speech basis	(M, F)
$\mathbf{W}_{\mathrm{ms}}$	music and speech basis	(R, J, M, F)

Table 2. Calculated π values of the separation results. *'mix'* column lists the symbols of the sources that are mixed to the input. The other columns are the evaluated π values grouped by the block lengths (in miliseconds). The filters are computed at every block. The last row is the average π. Audio files for all the results are accessible at the website.

mix	basis	500ms	100ms	50ms	25ms	10ms
R+J	$\mathbf{W}_{\mathrm{m}}$	9.6	11.0	10.9	10.7	10.6
R+M	$\mathbf{W}_{\mathrm{ms}}$	0.9	4.7	6.1	5.6	5.4
R+F	$\mathbf{W}_{\mathrm{ms}}$	2.9	6.5	7.8	7.4	6.7
J+M	$\mathbf{W}_{\mathrm{ms}}$	7.1	7.4	7.7	7.9	8.0
J+F	$\mathbf{W}_{\mathrm{ms}}$	6.3	6.0	5.3	5.2	4.9
M+F	$\mathbf{W}_{\mathrm{s}}$	4.0	4.6	4.3	4.3	4.2
Average		5.11	6.69	**7.01**	6.84	6.61

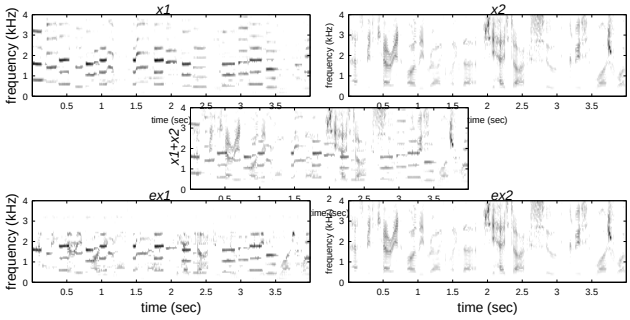


Fig. 4. Separation results of jazz music and male speech with block length 50ms. The graphs plot the spectrograms in logarithmic scale, time as x axis and frequency as y axis. In vertical order the graphs in (a) and (b) are for: original sources (x_1 and x_2), mixed signal ($x_1 + x_2$), and the recovered signals. Only the largest 20% of the spectral components (in terms of magnitude) are plotted.

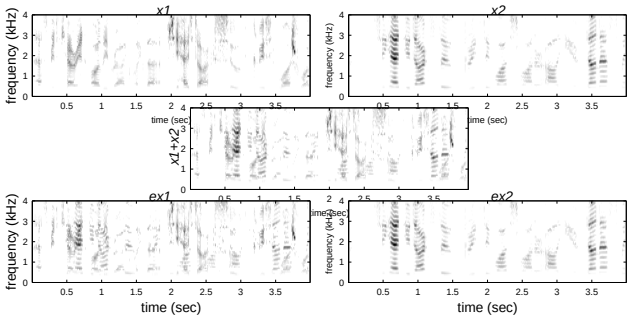


Fig. 5. Separation results of male and female speech with block length 50ms.

The similarity between the input and output signals is measured by signal-to-noise ratio (SNR), which is defined by

$$\mathrm{snr}(s, \hat{s})\ [\mathrm{dB}] = 10 \log_{10} \frac{\sum_t s^2}{\sum_t (s - \hat{s})^2}\,,$$

where s is the original source signal and $\hat{s}$ its estimate. To qualify a separation result we define a performance measurement function π as the sum of the increases in the SNR values of the two recovered source signals:

$$\pi(\hat{x}_1, \hat{x}_2; x_1, x_2) = \mathrm{snr}(x_1, \hat{x}_1) + \mathrm{snr}(x_2, \hat{x}_2)\,,$$

where $\hat{x}_i$ and x_i are the recovered sources and the original source signals. Table 2 reports the separation results. π values are grouped by block length, and the optimal block length was 50ms. Generally mixtures containing music were recovered more cleanly than male-female mixture.

Fig. 4 plots the spectrograms of the original sources and the recovered results for the mixture of jazz music and male speech. The recovered signals look very similar to the original sources.

4. CONCLUSIONS

We have presented a novel single channel signal separation algorithm based on subspace decomposition and maximum likelihood filtering. The original sources are recovered by projecting the input mixture onto the given subspaces, modulating the projections, and recombining the processed signals. The modulation filters are obtained by the ML estimation derived by the generalized Gaussian expansion of the projection pdf. The subspaces learned by the ICA algorithm achieve good separation performance. Experimental results showed successful separations of the simulated mixtures of rock and jazz music, and male and female speech signals. The proposed method has additional potential applications including suppression of environmental noise for communication systems and hearing aids, enhancing the quality of corrupted recordings, to preprocessing for speech recognition systems. On the theoretic end, We are currently working on alternative approaches to pdf modeling and object function, enhancing discriminacy of the subspaces, and optimizing the ML estimation towards real-time processing. On the practical end, we are interested in comparing our methods to other single channel denoising methods. Note that most de-noising methods are not suitable for comparison since they are not suited for non-stationary signals such as speech or music mixed into the single channel. We are also investigating the use of this approach in the AURORA 3 database evalution task.

5. REFERENCES

[1] A. S. Bregman, *Computational Auditory Scene Analysis*, MIT Press, Cambridge MA, 1994.

[2] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. on Neural Networks*, vol. 10, pp. 684–697, 1999.

[3] K. Kashino and T. Tanaka, "A sound source separation system with the ability of automatic tone modeling," in *Proc. of Int. Computer Music Conference*, 1993, pp. 248–255.

[4] D. P. W. Ellis, "A computer implementation of psychoacoustic grouping rules," in *Proc. 12th Int. Conf. on Pattern Recognition*, 1994.

[5] S. T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2001.

[6] M. S. Lewicki, "A flexible prior for independent component analysis," *to be published, Neural Computation*, 2000.

[7] Michael Zibulevsky and Barak A. Pearlmutter, "Blind source separation by sparse decomposition," *Neural Computations*, vol. 13, no. 4, 2001.

[8] A. J. Bell and T. J. Sejnowski, "Learning the higher-order structures of a natural sound," *Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 261–266, July 1996.

[9] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C: the art of scientific computing*, Cambridge University Press, second edition, 1992.