Single Channel Signal Separation Using MAP-based Subspace Decomposition

Gil-Jin Jang†, Te-Won Lee‡, and Yung-Hwan Oh†

[1]Spoken Language Laboratory, Department of Computer Science, KAIST
373-1 Gusong-dong, Usong-gu, Daejon 305-701, South Korea
Phone: +82-42-869-3556, Fax: +82-42-869-3510
Email: {jangbal,yhoh}@speech.kaist.ac.kr
‡Institute for Neural Computation, University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093-0523, USA
Email: tewon@ucsd.edu

An algorithm for single channel signal separation is presented. The algorithm projects the observed signal to given subspaces, and recovers the original sources by probabilistic weighting and recombining the subspace signals. The results of separating mixtures of two different natural sounds are reported.

Introduction: Extracting multiple source signals from a single channel mixture is a challenging research field with numerous applications. Conventional methods are mostly based on splitting mixtures observed as a single stream into different acoustic objects, by building an active scene analysis system for the acoustic events that occur simultaneously in the same spectro-temporal regions. Recently Roweis presented a refiltering technique to estimate time-varying masking filters that localize sound streams in a spectro-temporal region [1]. 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. Our work, while motivated by the concept of spectral masking, is free of the assumption that the spectrograms should be disjoint. The main novelty of the proposed method is that the masking filters can have any real value in $[0, 1]$, and that the filtering is done in the more discriminative, statistically independent subspaces obtained by independent component analysis (ICA). 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 [2] yields a good approximation.

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

$$y(t) = x_1(t) + x_2(t)\,, \;\; \forall t \in [1, T]\,, \tag{1}$$

where $x_i(t)$ is the $t\mathrm{th}$ observation of the $i\mathrm{th}$ source. 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 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. Our approach, illustrated in fig. 1, begins with decomposing the mixture signals into N disjoint subspace projections $v_k(t)$, each filtered to contain only energy from a small portion of the whole space:

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

where $\mathcal{P}$ is a projection operator, N is the number of subspaces, and w_{kn} is the $n\mathrm{th}$ coefficient of the $k\mathrm{th}$ coordinate vector $\mathbf{w}_k$ whose lag is d_k. Suppose the appropriate subparts of an audio signal 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, $u_{ik}(t)$ is approximated by modulating the mixed projections $v_k(t)$:

$$u_{1k}(t) \cong \lambda_k v_k(t), \;\; u_{2k}(t) \cong (1 - \lambda_k) v_k(t)\,, \tag{3}$$

where a "latent variable" λ_k is a weight on the projection 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 $\{u_{ik}(t)|k = 1, \ldots, N\}$ and performing the inverse transform of the projection. Proper choices of the weights λ_k enable the isolation of a single source from the input signal and the suppression of all other sources and background noises.

A set of subspaces that effectively split independent streams is essential in the success of the separation algorithm. Fig. 2 shows an example of desired subspaces. Two ellipses represent two different source distributions, 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 by the principle of orthogonality. To obtain an optimal basis, we adopt ICA, which estimates the inverse-translation-operator such that the resulting coordinate can be statistically as independent as possible [3].

Estimating Source Signals: The estimation of λ_k can be accomplished by simply finding the values that maximize the probability of the subspace projections. The success of the separation algorithm for our purpose depends highly on how closely the ICA density model captures the true source coefficient density. The histograms of natural sounds reveal that $p(u_{ik})$ is highly super-Gaussian [3]. Therefore we use a generalized Gaussian prior [2] that provides an accurate estimate for symmetric non-Gaussian distributions in modeling the underlying distribution of the source coefficients, with a varying degree of normality in the following general form: $p(u) \propto \exp\left(-|u|^q\right)$. We approximate the log probability density of the projections according to eq. 3:

$$\begin{aligned}\log p\left(u_{1k}\right) &\propto -\left|u_{1k}\right|^{q_{1k}} \cong -\lambda_k^{q_{1k}}\left|v_k\right|^{q_{1k}}\\ \log p\left(u_{2k}\right) &\propto -\left|u_{2k}\right|^{q_{2k}} \cong -(1-\lambda_k)^{q_{2k}}\left|v_k\right|^{q_{2k}}\ .\end{aligned} \tag{4}$$

We define the object function Ψ_k of subspace k by the sum of the joint log probability density of $u_{1k}(t)$ and $u_{2k}(t)$, over the time axis:

$$\begin{aligned}\Psi_k &\overset{\text{def}}{=} \textstyle\sum_t \log p\left(u_{1k}(t), u_{2k}(t)\right)\\ &\cong -\lambda_k^{q_{1k}} \textstyle\sum_t |v_k(t)|^{q_{1k}} - (1-\lambda_k)^{q_{2k}} \textstyle\sum_t |v_k(t)|^{q_{2k}}.\end{aligned} \tag{5}$$

The problem is equivalent to constrained maximization in the closed interval $[0, 1]$; we can find a unique value of λ_k at either boundaries (0 or 1), or local maximum by Newton's method.

Evaluation: 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, and male and female speech. Audio files for all the experiments are accessible at *http://speech.kaist.ac.kr/˜jangbal/rbss1el/*. We used different sets of sound signals for generating mixtures and for learning ICA subspaces ($\mathbf{w}_k$), and estimating generalized Gaussian parameters (q_{ik}, variances, etc.) to model the subspace component pdfs. While learning the subspaces, all the training data of source 1 and source 2 are used to reflect the statistical properties of both sound sources upon the resultant subspaces. The pdf parameters are estimated separately for each source. The weighting factors 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 of the other blocks; hence the weighting becomes more accurate as the block length shrinks. However if the block length is too short, the computation becomes unreliable. The optimal block length was 25ms in our experiments.

From the testing data set, two sources out of the four are selected and added sample-by-sample to generate a mixture signal. The proposed separation algorithm was applied to recover the original sources. Table 1 reports the separation results when Fourier and the learned ICA bases are used. The proposed method deals with binary mixtures only. The performances are measured by signal-to-noise ratio (SNR). With learned ICA subspaces, the performances were improved more than 1dB on the average, compared to Fourier basis. In terms of the sound source types, generally mixtures containing music were recovered more cleanly than the male-female mixture for both bases.

Conclusion: The algorithm can separate the single channel mixture signals of two different sound sources. The original source signals are recovered by projecting the input mixture onto the given subspaces, modulating the projections, and recombining the projected signals. 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, and preprocessing for speech recognition systems.

References

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

[2] T.-W. Lee and M. S. Lewicki, “The generalized Gaussian mixture model using ICA,” in *International Workshop on Independent Component Analysis (ICA’00)*, (Helsinki, Finland), pp. 239–244, June 2000.

[3] A. J. Bell and T. J. Sejnowski, “Learning the higher-order structures of a natural sound,” *Network: Computation in Neural Systems*, vol. 7, pp. 261–266, July 1996.

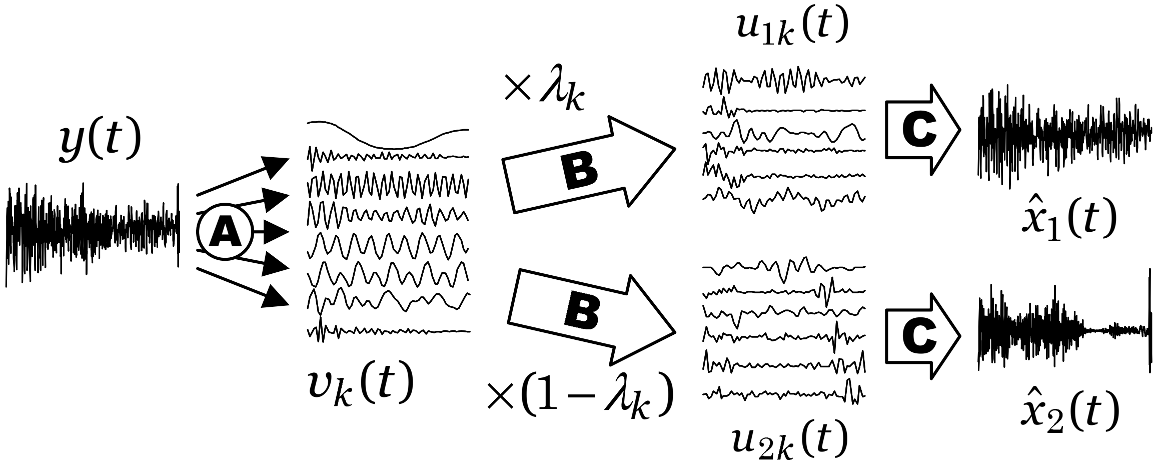

Figure 1: Block diagram of subspace weighting. (**A**) Input signal $y(t)$ is projected onto N, 1-dimension subspaces. (**B**) The projections $v_k(t)$ are modulated by weighting factors $\lambda_k, 1 - \lambda_k \in [0, 1]$. (**C**) The separation process finally terminates with summing up the N modulated signals.

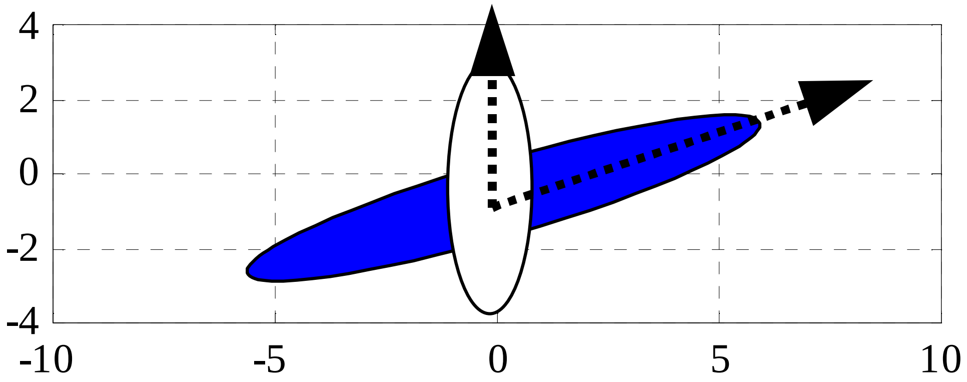

Figure 2: Illustration of desired subspaces. The ellipses represent the distributions of two different classes. The arrows are the 1-dimensional subspaces along the maximum energy concentrations of the classes. Projecting the mixtures onto each subspace provides minimum error separation.

Table 1: Computed SNRs of the separation results. The first row lists the two symbols of the sources that are mixed to the input. (R, J, M, F) stand for rock, jazz music, male, and female speech. The last column is the average SNR. Audio files for all the results are accessible at *http://speech.kaist.ac.kr/˜jangbal/rbss1el/*.

basis	RJ	RM	RF	JM	JF	MF	Avg.
Fourier	8.3	3.4	5.3	7.3	6.4	3.8	**5.74**
ICA	10.2	4.6	6.1	8.2	6.7	5.1	**6.82**