

NEW EM ALGORITHMS FOR SOURCE SEPARATION AND DECONVOLUTION WITH A MICROPHONE ARRAY

Hagai Attias

Microsoft Research
Redmond, WA 98052
hagaia@microsoft.com

ABSTRACT

This paper presents new algorithms for source separation with a microphone array. Key to our algorithms are exploiting detailed source models, using subband filtering ideas to model the reverberant environment, and employing explicit models for background and sensor noise. We demonstrate strong performance which is robust to noise and reverberations. Good scaling properties are obtained using machine learning techniques.

1. INTRODUCTION

This paper makes several contributions to the problem of real world source separation. In the problem as defined here, signals from K independent sources are received by each of $L \geq K$ sensors. The task is to obtain an *optimal estimate* of the sources from the sensor signals. One reason this task is difficult is that the received signals are distorted versions of the originals. There are two types of distortions. The first type arises from propagation through a medium, and is approximately linear but also history dependent. This type is termed reverberations. The second type arises from background noise and sensor noise, which are assumed additive. The task is also difficult for another reason, which is lack of advance knowledge of the properties of the sources and of the distortions.

Unfortunately, the intense activity this problem has attracted over the last several years (see, e.g., [8],[2],[6]) has not yet produced a satisfactory solution. In our opinion, the reason is that existing techniques fail to address three major factors. The first is noise robustness: algorithms typically ignore background and sensor noise, sometime assuming they may be treated as additional sources. It seems plausible that to produce a noise robust algorithm, noise signals and their properties must be modeled explicitly, and these models should be exploited to compute optimal source estimators. The second factor is filter modeling: algorithms typically seek, and directly optimize, a transformation that would unmix the sources. However, in many situations, the filters describing medium propagation are non-invertible, or have an unstable inverse, or have a stable inverse that is extremely long. It may hence be advantageous to estimate the mixing filters themselves, then use them to estimate the sources. The third factor is source properties: algorithms typically use a very simple source model (e.g., a one time point histogram). But in many cases one may easily obtain detailed models of the source signals, and incorporating them into the algorithms could potentially improve performance. This is particularly true for speech sources, where large datasets exist and much modeling expertise has developed over decades of

research. Separation of speakers is also one of the major potential commercial applications of source separation algorithms.

In the following, we present new separation algorithms that are the first to address all three factors. We work in the framework of statistical models. This framework allows us to construct models for sources and for noise, combine them with the reverberant mixing transformation in a principled manner, and compute parameter and source estimates from data which are Bayes optimal. We identify three technical ideas that are key to our approach: (1) a strong speech model, (2) subband filtering, and (3) variational EM.

2. SUBBAND FILTERING

We start with the concept of subband filtering. This is also a good point to define our notation. Let x_m denote a time domain signal, e.g., the value of a sound pressure waveform at time point $m = 0, 1, 2, \dots$. Let $X_n[k]$ denote the corresponding subband signal at time frame n and subband frequency k . The subband signals are obtained from the time domain signal by imposing an N -point window w_m , $m = 0 : N - 1$ on that signal at equally spaced points nJ , $n = 0, 1, 2, \dots$, and FFT-ing the windowed signal,

$$X_n[k] = \sum_{m=0}^{N-1} e^{-i\omega_k m} w_m x_{nJ+m}, \quad (1)$$

where $\omega_k = 2\pi k/N$ and $k = 0 : N - 1$. Notice the difference in time scale between the time frame index n in $X_n[k]$ and the time point index n in x_n .

The chosen value of the spacing J depends on the window length N . For $J \leq N$ the original signal x_m can be synthesized exactly from the subband signals (we omit the synthesis formula). Consider a filter h_m applied to x_m , and denote by y_m the filtered signal. In the simple case $h_m = h\delta_{m,0}$ (no filtering), the subband signals keep the same dependence as the time domain ones,

$$y_n = hx_n \quad \longrightarrow \quad Y_n[k] = hX_n[k]. \quad (2)$$

For an arbitrary filter h_m , we use the relation

$$y_n = \sum_m h_m x_{n-m} \quad \longrightarrow \quad Y_n[k] = \sum_m H_m[k] X_{n-m}[k], \quad (3)$$

with complex coefficients $H_m[k]$ for each k . This relation between the subband signals is termed subband filtering, and the $H_m[k]$ are termed subband filters.

Unlike the simple case (2), the relation (3) holds approximately, but quite accurately using an appropriate choice of J and

w_m . We have found using simulations that in the limiting case of a Gaussian i.i.d. signal x_m and filter h_m , with a linear regression estimate for $H_m[k]$, the relative mean squared error of $|Y_n[k] - \sum_m H_m[k]X_{n-m}[k]|^2$ was .034 using a Hamming window and $J = N/2$; all other filters provided in the Matlab signal processing toolbox performed worse. Furthermore, with the actual speech signals and the filters used in the experiments in this paper, the error decreases to .019. Throughout this paper, we will therefore assume that an arbitrary filter h_m can be modeled by the subband filters $H_m[k]$ to a sufficient accuracy for our purposes.

One advantage of subband filtering is that it replaces a long filter h_m by a set of short independent filters $H_m[k]$, one per frequency. This will turn out to decompose the source separation and deconvolution problem into a set of small (albeit coupled) problems, one per frequency. Another advantage is that this representation allows using a detailed speech model on the same footing with the filter model. This is because a speech model is defined on the time scale of a single frame, whereas the original filter h_m , in contrast with $H_m[k]$, is typically as long as 10 or more frames.

As a final point on notation, we define a Gaussian distribution over a complex number Z by

$$p(Z) = \mathcal{N}(Z | \mu, \nu) = \frac{\nu}{\pi} \exp(-\nu |Z - \mu|^2). \quad (4)$$

Notice that this is a joint distribution over the real and imaginary parts of Z . $\nu = (\langle |X|^2 \rangle - |\langle X \rangle|^2)^{-1}$ is termed the precision (defined as the inverse variance).

3. SPEECH MODEL

We assume independent sources, and model the distribution of source j by a mixture model over its subband signals X_{jn} ,

$$\begin{aligned} p(X_{jn} | S_{jn} = s) &= \prod_{k=1}^{N/2-1} \mathcal{N}(X_{jn}[k] | 0, A_{js}[k]) \\ p(S_{jn} = s) &= \pi_{js} \\ p(X, S) &= \prod_{jn} p(X_{jn} | S_{jn}) p(S_{jn}), \end{aligned} \quad (5)$$

where the components are labeled by S_{jn} . Component s of source j is a zero mean Gaussian with precision A_{js} . The mixing proportions of source j are π_{js} . A similar model was used in [1] for one microphone speech enhancement for recognition (see also [3]).

Here are several things to note about this model. (1) Each component s has a characteristic spectrum $\langle |X_{jn}|^2 \rangle = A_{js}^{-1}$, which may describe a particular part of a speech phoneme. (2) A zero mean model is appropriate since the mean of a sound pressure waveform is zero. (3) k runs from 1 to $N/2 - 1$, since for $k > N/2$, $X_{jn}[k] = X_{jn}[N - k]^*$; the subbands $k = 0, N/2$ are real and are omitted from the model, a common practice in speech recognition engines. (4) Perhaps most importantly, for each source *the subband signals are correlated* via the component label s , since

$$p(X_{jn}) = \sum_s p(X_{jn}, S_{jn} = s) \neq \prod_k p(X_{jn}[k]). \quad (6)$$

Hence, when the source separation problem decomposes into one problem per frequency, these problems are coupled (see below), and independent frequency permutations are avoided. (5) To increase model accuracy, a state transition matrix $p(S_{jn} = s | S_{j,n-1} = s')$ may be added for each source. The resulting HMM

models are straightforward to incorporate without increasing the algorithm complexity.

Our experiments focused on separating speakers, and we therefore used a speech model for each source. The model was speaker independent and was trained offline on a large dataset of clean speech signals, including 150 male and female speakers reading sentences from the Wall Street Journal (see [1] for details). The training algorithm used was standard EM (omitted) using 256 clusters, initialized by vector quantization.

4. NON-REVERBERANT MIXING

We now present a source separation algorithm for the case of non-reverberant (or instantaneous) mixing. Whereas many algorithms exist for this case, our contribution here is an algorithm that is significantly more robust to noise. Its robustness results, as indicated in the introduction, from three factors: (1) explicitly modeling the noise in the problem, (2) using a strong source model, in particular modeling the temporal statistics (over N time points) of the sources, rather than one time point statistics, and (3) extracting each source signal from data by a Bayes optimal estimator obtained from $p(X | Y)$. A more minor point is handling the case of less sources than sensors in a principled way.

The mixing situation is described by $y_{in} = \sum_j h_{ij} x_{jn} + u_{in}$, where x_{jn} is source signal j at time point n , y_{in} is sensor signal i , h_{ij} is the instantaneous mixing matrix, and u_{in} is the noise corrupting sensor i 's signal. The corresponding subband signals satisfy

$$Y_{in}[k] = \sum_j h_{ij} X_{jn}[k] + U_{in}[k]. \quad (7)$$

To turn (7) into a statistical model, we assume that noise i has precision (inverse spectrum) $B_i[k]$, and that noises at different sensors are independent (the latter assumption is often inaccurate but can be easily relaxed). This yields

$$\begin{aligned} p(Y_{in} | X) &= \prod_k \mathcal{N}(Y_{in}[k] | \sum_j h_{ij} X_{jn}[k], B_i[k]) \\ p(Y | X) &= \prod_{in} p(Y_{in} | X), \end{aligned} \quad (8)$$

which together with the speech model (5) forms a complete model $p(Y, X, S)$ for this problem.

The model parameters $\theta = \{h_{ij}, B_i[k], A_{js}[k], \pi_{js}\}$ are estimated from data by an EM algorithm. However, as the number of speech components M or the number of sources K increases, the E-step becomes computationally intractable, as it requires summing over all $\mathcal{O}(M^K)$ configurations of (S_{1n}, \dots, S_{Kn}) at each frame. Hence, an approximation must be made. The machine learning community has recently made some progress in developing powerful new techniques to tackle intractable statistical models. We have found one such set of methods, termed *variational techniques* (see [5] and appendix), to be particularly suitable to the present problem. Basically, we focus on the posterior distribution $p(X, S | Y)$ over the unobserved variables conditioned on the data, and compute an optimal tractable approximation for it, denoted $q(X, S | Y) \approx p(X, S | Y)$. Using q we update the sufficient statistics and the parameters. After convergence, the sources are obtained by the posterior mean (MMSE) estimator

$$\hat{X}_{jn}[k] = E(X_{jn}[k] | Y) = \int dX q(X | Y) X_{jn}[k], \quad (9)$$

from which their time domain waveforms \hat{x}_{jm} are synthesized.

M-step. The sufficient statistics (SS) required for the M-step are the posterior cross correlations

$$\begin{aligned}\lambda_{jj',m}[k] &= \sum_n E(X_{j,n+m}[k]X_{j'n}[k]^* | Y), \\ \eta_{ij,m}[k] &= \sum_n E(Y_{i,n+m}[k]X_{jn}[k]^* | Y),\end{aligned}\quad (10)$$

where E denotes averaging w.r.t. $q(X | Y)$. For the non-reverberant case only $m = 0$ is used. The update rule for the mixing matrix h_{ij} is obtained by solving the linear equation

$$\sum_k B_i[k]\eta_{ij,0}[k] = \sum_{j'} h_{ij'} \sum_k B_i[k]\lambda_{j',0}[k]. \quad (11)$$

The update rule for $B_i[k]$ is omitted due to space restrictions.

E-step. The posterior means of the sources (9) are obtained by solving

$$\hat{X}_{jn}[k] = \hat{\nu}_j[k]^{-1} \sum_i B_i[k]h_{ij} \left(Y_{in}[k] - \sum_{j' \neq j} h_{ij'} \hat{X}_{j'n}[k] \right) \quad (12)$$

for $\hat{X}_{jn}[k]$, which is a $K \times K$ linear system for each frequency k and frame n . Other SS are the state posterior means ρ_{jns} and state posterior precisions ν_{js} , given by

$$\begin{aligned}\rho_{jns}[k] &= \frac{\hat{\nu}_{jn}[k]}{\nu_{js}[k]} \hat{X}_{jn}[k], \\ \nu_{js}[k] &= \sum_i B_i[k]h_{ij}^2 + A_{js}[k], \\ \hat{\nu}_{jn}[k]^{-1} &= \sum_s \gamma_{jns} \nu_{js}[k]^{-1},\end{aligned}\quad (13)$$

where the ν_{js} turn out to be frame independent. Also $\hat{X}_{jn}[k] = \sum_s \gamma_{jns} \rho_{jns}[k]$. Finally, the last SS are the state responsibilities $\gamma_{jns} = q(S_{jn} = s | Y)$, obtained via

$$\gamma_{jns} = \frac{1}{z_{jn}} \prod_k \exp \left(\nu_{jns}[k] | \rho_{js}[k] |^2 + \log \frac{A_{js}[k]}{\nu_{js}[k]} \right) \cdot \pi_{js} \quad (14)$$

where z_{jn} is a normalization constant.

The SS $\lambda_{jj',m}$ and $\eta_{ij,m}$ can now be given directly in terms of all these quantities:

$$\begin{aligned}\lambda_{jj',m}[k] &= \sum_n \hat{X}_{j,n+m}[k] \hat{X}_{j'n}[k]^*, \\ \lambda_{jj,0}[k] &= \sum_n \left(\sum_s \gamma_{jns} | \rho_{jns}[k] |^2 + \hat{\nu}_{jn}[k]^{-1} \right), \\ \eta_{ij,m}[k] &= \sum_n Y_{i,n+m}[k] \hat{X}_{jn}[k]^*,\end{aligned}\quad (15)$$

where the first line holds except when both $j' = j$ and $m = 0$. Eqs. (12)–(15) constitute the variational E-step and are solved by iteration.

4.1. Results

We have performed three sets of experiments. In each set, K 20sec long speech signals at 16 kHz sampling rate from the Wall Street Journal dataset (see [1]) were mixed together by a random $L \times K$ mixing matrix sampled from a Gaussian distribution. L noise signals taken from a long noise sequence recorded in an office environment with A/C and two PCs were added to the mixtures. The signal to noise ratio (SNR) was 10dB. The experiment was performed for pairs $(L, K) = (3, 2), (5, 3), (5, 5)$, where for each pair 100 mixing matrices were sampled (not that some had zero or very low rank). The algorithm was applied to each of the resulting 300 datasets. The extracted sources were compared against the original ones and the SNR was computed. Averaged over mixing matrices, the SNR improvement for the 3 pairs over the noisy signals was 4.4dB, 4.2dB, 3.7dB, respectively.

The same datasets were analyzed using noiseless IFA [7], which resulted in SNR improvement of .9dB, .8dB, .6dB, respectively. Some of the datasets were analyzed with noisy IFA [7], obtaining just slightly better results than noiseless IFA: 1.2dB, 1.2dB, .9dB. We therefore conclude that our algorithm is significantly more robust to noise than IFA, which is competitive with state of the art instantaneous separation algorithms.

5. REVERBERANT MIXING

In this section we extend the algorithm to the case of reverberant mixing. In that case, due to signal propagation in the medium, each sensor signal at time n depends on the source signals not just at the same time but also at previous times. To describe this mathematically, the mixing matrix h_{ij} must become a matrix of filters $h_{ij,m}$: we have $y_{in} = \sum_{jm} h_{ij,m} x_{j,n-m} + u_{in}$. It may seem straightforward to extend the algorithm derived above to the present case. However, this appearance is misleading, because we have a time scale problem. Whereas the speech model $p(X, S)$ is frame based, the filters $h_{ij,m}$ are generally longer than the frame length N , typically 10 frames long and sometime longer.

This is where the idea of subband filtering becomes very useful. Using (3) we have

$$Y_{in}[k] = \sum_{jm} H_{ij,m}[k] X_{j,n-m}[k] + U_{in}[k], \quad (16)$$

which yields the statistical model

$$p(Y_{in} | X) = \prod_k \mathcal{N}(Y_{in}[k] | \sum_{jm} H_{ij,m}[k] X_{j,n-m}[k], B_i[k])$$

(compare to (8)). Hence, both X and Y are now frame based. Combining this equation with the speech model (5), we now have a complete model $p(Y, X, S)$ for the reverberant mixing problem.

The model parameters $\theta = \{H_{ij}[k], B_i[k], A_{js}[k], \pi_{js}\}$ are estimated from data by a variational EM algorithm, whose derivation generally follows the one outlined in the previous section and the appendix. Notice that the exact E-step here is even more intractable, due to the history dependence introduced by the filters.

M-step. The update rule for $H_{ij,m}$ is obtained by solving the Toeplitz system

$$\sum_{j'm'} H_{ij',m'}[k] \lambda_{j',j,m-m'}[k] = \eta_{ij,m}[k] \quad (17)$$

using the SS as defined in (10). The update rule for the $B_i[k]$ is omitted due to space restrictions.

E-step. The posterior means of the sources (9) are obtained by solving

$$\hat{X}_{jn}[k] = \hat{\nu}_{jn}[k]^{-1} \sum_{im} B_i[k] H_{ij,m-n}[k]^* \cdot \left(Y_{im}[k] - \sum_{j'm' \neq jm} H_{ij',m-m'}[k] \hat{X}_{j'm'}[k] \right) \quad (18)$$

for $\hat{X}_{jn}[k]$. Assuming K sources and P frames long filters $H_{ij,m}$, $m = 0 : P - 1$, this is a $KP \times KP$ linear system for each frequency k . All the other SS satisfy precisely the same equations as in the previous section, with the exception of the state posterior precisions that are given by

$$\nu_{js}[k] = \sum_{im} B_i[k] |H_{ij,m}[k]|^2 + A_{js}[k]. \quad (19)$$

Notice that, whereas in the M-step and in computing \hat{X}_{jn} we solve a separate problem for each k , these problems are coupled via the sum over s required to compute the SS in the E-step. As mentioned above, this eliminates the problem of independent frequency permutations.

5.1. Results

We have performed two sets of experiments. In each set, $K = 2$ 20sec long speech signals at 16 kHz sampling rate from the Wall Street Journal dataset (see [1]) were mixed together by a $L \times K$ matrix of filters with $L = 3$. The 6 filters were independently measured in an office environment and were 2000 taps long. L noise signals taken from a long noise sequence recorded in an office environment with A/C and two PCs were added to the mixtures. The signal to noise ratio (SNR) was 10dB. The algorithm was applied to the resulting data. The extracted sources were compared against the original ones and the SNR was computed. Averaged over mixing matrices, the SNR improvement for the 3 pairs over the noisy signals was 7.6dB, 7.1dB, 5.5dB, respectively. One thing listeners notice is the lack a whitening of the extracted speech signals. Whitening is well known to affect many source separation algorithms in the presence of reverberation. We largely avoid it due to the strong speech model, which in effect limits the spectral range of the resulting signals.

6. CONCLUSION

One extension of this work would make a connection with speech recognition engines. It would be interesting to see whether the improvement in SNR for speech sources demonstrated here would translate into improvement in word error rate, as was the case for the speech enhancement algorithm presented in [1]. In addition, we hope to form a connection between our framework and modern array processing methods (see, e.g., [4]).

7. APPENDIX

We give a brief outline of the derivation of the two algorithms in this paper. Variational techniques [5] come to our help when the exact E-step is computationally intractable, and our goal is to find

a tractable optimal approximation. One usually thinks of the E-step as computing sufficient statistics; more generally, the E-step computes the posterior distribution over the hidden variables, in our case X, S , conditioned on the data Y . Here we find a tractable optimal approximation $q(X, S | Y)$ to the exact posterior $p(X, S | Y)$.

We start with the cost function

$$\mathcal{F}[q] = \sum_S \int dX q(X, S | Y) \log \frac{p(Y, X, S)}{q(X, S | Y)} \leq \mathcal{L}, \quad (20)$$

which for any q is bounded from above by the data likelihood $\mathcal{L} = \log p(Y)$. The difference between \mathcal{F} and \mathcal{L} is the KL distance between q and the exact posterior. Our strategy is to specify a structure for q , and maximize \mathcal{F} w.r.t. that structure, thus obtaining an optimal q in the sense of minimal KL distance. We choose

$$q(X, S | Y) = \prod_{jn} q(X_{jn}, S_{jn} | Y), \quad (21)$$

where the hidden variables are factorized over the sources and the frames. This posterior maintains the dependence of X on S , and thus the correlations between different subbands $X_{jn}[k]$. A slightly more general form which allows inter-frame correlations by employing $q(S) = \prod_{jn} q(S_{jn} | S_{j,n-1})$ may also be used, without increasing complexity.

By optimizing \mathcal{F} w.r.t. q using variational calculus, we obtain

$$q(X_{jn}, S_{jn} = s) = \prod_k q(X_{jn}[k] | S_{jn} = s) q(S_{jn} = s), \quad (22)$$

where

$$q(X_{jn}[k] | S_{jn} = s) = \mathcal{N}(X_{jn}[k] | \rho_{jns}[k], \nu_{js}[k]) \quad (23)$$

and $q(S_{jn} = s) = \gamma_{jns}$. Both the factorization over k given s of $q(X_{jn} | S_{jn})$ and its Gaussian functional form fall out from the optimization under the structural restriction (21) and need not be specified in advance. The parameters appearing in (23) are computed in the body of the paper.

8. REFERENCES

- [1] H. Attias, L. Deng, A. Acero, J.C. Platt (2001). A new method for speech denoising using probabilistic models for clean speech and for noise. *Proc. Eurospeech 2001*.
- [2] S.Araki et al. (2001). Limitation of frequency domain blind source separation for convolutive mixture of speech. *Proc. IEEE HSC-01*, pp.91-94.
- [3] Ephraim, Y. (1992). Statistical model based speech enhancement systems. *Proc. IEEE* 80(10), 1526-1555.
- [4] S. Griebel, M. Brandstein (2001). Microphone array speech dereverberation using coarse channel modeling. *Proc. ICASSP 2001*.
- [5] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul (1999). An introduction to variational methods in graphical models. *Machine Learning* 37, 183-233.
- [6] T.-W. Lee et al. (2001) (Ed.). *Proc. ICA 2001*.
- [7] H. Attias (1999). Independent Factor Analysis. *Neural Computation* 11, 803-851.
- [8] A.J. Bell, T.J. Sejnowski (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation* 7, 1129-1159.