Fast Monaural Separation of Speech

Niels Henrik Pontoppidan and Mads Dyrholm

Technical University of Denmark, Informatics and Mathematical Modelling, 2800 Lyngby, Denmark

Correspondence should be addressed to Niels Henrik Pontoppidan (nhp@imm.dtu.dk)

ABSTRACT

We have investigated the possibility of separating signals from a single mixture of sources. This problem is termed the Monaural Separation Problem.
Lars Kai Hansen has argued that this problem is topological tougher than problems with multiple recordings. Roweis has shown that inference from a Factorial Hidden Markov Model, with non-stationary assumptions on the source autocorrelations modelled through the Factorial Hidden Markov Model, leads to separation in the monaural case.
By extending Hansens work we find that Roweis' assumptions are necessary for monaural speech separation. Furthermore we develop a Factorial hierarchical vector quantizer yielding a significant decrease in complexity of inference.

THE MONAURAL PROBLEM

The task of recovering multiple sources from a single mixture is the *monaural problem* – and humans do it all the time. Listening to a discussion through an open door, we are able to keep track of what the different people say – to great extend even when they speak simultaneously.

BAYESIAN APPROACH

We estimate the sources by measures of the posterior density. Others have investigated this problem using the *maximum a-posteriori* (MAP) estimator[1]. We build upon that work and use the *posterior mean*. Obtaining the posterior density involves formulating a generating model (likelihood) and assuming densities (priors) for the sources.

We assume that the signal x (a single number at a given time) is the result of instantaneous mixing of the two signals s_1 and s_2 with mixer coefficients a_1 and a_2.

$$x = \mathbf{as} \tag{1}$$

$$\mathbf{a} = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \tag{2}$$

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \tag{3}$$

Under the assumption that $\mathbf{a}$ is known, the likelihood is

$$p(x|\mathbf{a},\mathbf{s}) = \delta(x - \mathbf{as}) \tag{4}$$

With prior assumptions on the sources $p(\mathbf{s})$, we obtain the posterior through Bayes rule

$$p(\mathbf{s}|x,\mathbf{a}) = \frac{p(x|\mathbf{a},\mathbf{s})p(\mathbf{s})}{p(x|\mathbf{a})} \tag{5}$$

We assume that the sources are white noise signals following a Cauchy distribution, e.g. heavy tailed. Hansen has shown that the MAP estimate fails in separating the two sources[1] . Our experiments has shown that using the mean as an estimator does not solve this problem, even though that the mean square error is decreased.

$$MAP = \arg_{\mathbf{s}} \max p(\mathbf{s}|x,\mathbf{a}) \tag{6}$$

$$posterior\,mean = \int \mathbf{s}\, p(\mathbf{s}|x,\mathbf{a}) d\mathbf{s} \tag{7}$$

The MAP estimate grants all variance to one of the signals (implicitly decided by $\mathbf{a}$) and the mean estimate merely scales the observed signal according to $\mathbf{a}$. The problem is that the resulting estimates,

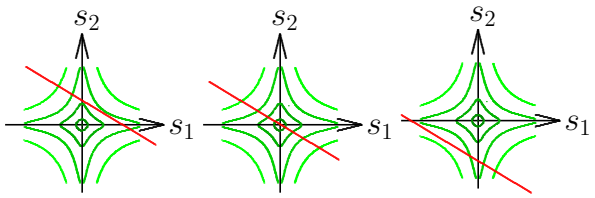

Fig. 1: Contours of a joint Cauchy-densities (the star shapes) with same **a** but different observation x. The valid source estimates s_1 and s_2 live on the straight lines.

which are functions of x (eq. (6) and eq. (7)), live on a curve, whereas the original sources live in the whole plane. We refer to this as the *collapse* of the estimated sources. Furthermore figure 1 show that the points on these curves have small probability in $p(s_1, s_2)$ thus the estimated sources are very unlikely to occur under the prior. However they are the best estimates that fulfill the likelihood ($x = a_1 s_1 + a_2 s_2$).

From this we can draw the conclusion that we are able to separate two white Cauchy sequences from one mixture.

Colored sources

We believe that problem is shortage of information – so we introduce additional information by assuming that the signals have different autocorrelation functions. In order to examine wether autocorrelation improves the estimation, we are faced with the task of formulating our new prior: the multivariate density for a correlated Cauchy sequence.

For the simplest case of correlation, namely the AR(1) process (autoregressive process of order 1), we derive the strucutre of the conditional density for the normal distribution and transfer this to the Cauchy distribution. The correlation is created by letting the previous value $s(t)$ move the mean value weighted by the correlation coefficients.

Given one signal $s(t)$ at time τ and $\tau - 1$, the Cauchy parameters α, β and the correlation coefficients $\gamma(0)$

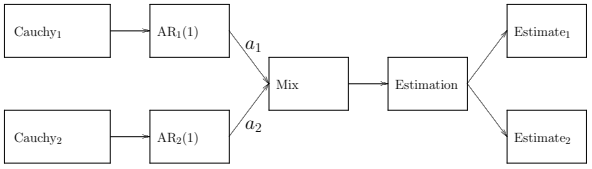

Fig. 2: Setup for experiments with colored Cauchy sequences.

and $\gamma(1)$ the conditional Cauchy density turns into

$$p\left(s(\tau)|s(\tau-1)\right) = \frac{\pi^{-1}\beta}{\beta^2 + \left(s(\tau) - \alpha - \frac{\gamma(1)}{\gamma(0)} s(\tau-1)\right)^2} \quad (8)$$

$$p(s(\tau), s(\tau-1)) = p(s(\tau), s(\tau-1))p(s(\tau-1)) \quad (9)$$

$$\mathbf{x} = \mathbf{aS} \quad (10)$$

$$\mathbf{S} = \begin{pmatrix} s_1(\tau) & s_1(\tau-1) \\ s_2(\tau) & s_2(\tau-1) \end{pmatrix} \quad (11)$$

$$p(\mathbf{S}) = \prod_{i=1}^{2} p(s_i(\tau), s_i(\tau-1)) \quad (12)$$

Estimating the sources

With the new two-dimensional densities with correlation (priors) and Bayes rule we are ready to re-estimate the sources. Our setup is as in figure 2, we mix two sources, both being the result of feeding an AR(1) process with white noise following a Cauchy distribution. We expect that the collapse of the estimated sources is decreased, i.e. that the estimates are likely estimates under both the prior and fulfilling the likelihood.

Figure 3 shows that the estimates have moved away from the line, thus the collapse is reduced. We note that this is also accompanied by a decrease of the mean squared error [2]. In some sense nothing has changed, with N observations we still need to estimate $2N$ source values, we just utilize the structure of the sounds.

We expect that if the sources have longer correlation, i.e. coming from higher order AR processes, then the collapse and the mean square error decreased

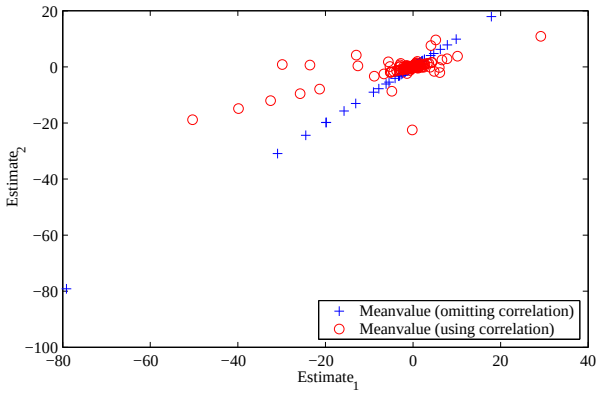

Fig. 3: The estimates obtained omitting respectively using autocorrelation. The sources where colored sequences following Cauchy distributions. The ones estimated using the autocorrelation does not live on a line

further. The reason is that the longer the correlation the more observations go into estimating each source value – and we know this is good for the estimation process.

This toy example sketches how the knowledge of autocorrelation yields better estimates of the sources. If we want to use this for separating the voices of two speakers, we'll also need a system that is able to learn and handle their individual but multiple autocorrelation functions.

AN ALGORITHM: HIERARCHICAL RE-FILTERING

The fact that the human is capable of focusing on one particular stream of sound has delivered many CASA systems with built-in knowledge about individual sources (see e.g. [3, 4]). In our method we represent any signal by the logarithm of an element-wise squared spectogram (log-power spectogram) — mimicing the human cochlear time-/frequency decomposition. Roweis has pointed out that the log-power spectogram of a mixture of two speakers is "nearly" the element-wise max of the individual log-power spectograms [5]. In context, note that a log-power spectrum is the Fourier spectrum of an autocorrelation function.

Let K denote the number of frequency bands constituting one log-power spectrum, and let $b_n(t)$ be the n'th subband signal component of the mixture

$$s_1(t) + s_2(t) \approx \sum_{n=1}^{K} b_n(t). \tag{13}$$

To extract the contribution from one speaker over the other we make use of spectral properties of both speakers. *Re-filtering* is then applied yielding an estimate of one speakers post-cochlear contribution:

$$\hat{s}_1(t) = \sum_{n=1}^{K} \alpha_n(t) b_n(t), \tag{14}$$

and now the key to succesful separation is finding "good" *masking signals* $\alpha_n(t)$. We restrict masking signals to be binary and piecewise constant with the spectogram time resolution.

By finding that pair $(\mathbf{S}^{(1)}, \mathbf{S}^{(2)})$ of log-power spectograms that best approximates the measured spectogram $\mathbf{S}$, in the 2-norm i.e.

$$(\hat{\mathbf{S}}^{(1)}, \hat{\mathbf{S}}^{(2)}) = \arg \min_{\mathbf{S}^{(1)}, \mathbf{S}^{(2)}} \sum_{i,j} [S_{ij} - \max(S_{ij}^{(1)}, S_{ij}^{(2)})]^2, \tag{15}$$

we find masking signals for extracting speaker-1 by

$$\alpha_n(t) = \begin{cases} 1 & , \hat{S}_{nt}^{(1)} > \hat{S}_{nt}^{(2)} \\ 0 & , \text{otherwise} \end{cases} . \tag{16}$$

In this presentation we make no use of structure across time intervals[1] so solving eq. (15) simplifies to solving the problem for each time-step individually. We use two lookup tables—one for each speaker—with typical log-power spectra which are found by clustering spoken examples hierachically using a very simple algorithm:

1. Extract the mean from the data
2. Find direction of largest variance using the first singular vector of the Singular Value Decomposition.
3. Project data onto that direction, and split data in two clusters using the sign of the projection to discriminate.

[1] We have a generalized algorithm which includes inter-window time structure—indeed giving rise to other complications which is currently subject to future work.

4. For each of these two clusters: go recursively to step-1 unless satisfactory level of precision (small within-cluster varinace) has been achieved.

By storing the mean from step-1 of this algorithm into an appropriate data structure we later utilize this to solving eq. (15) very efficiently:

1. Given $\mathbf{s}$ as one column vector from the mixture's log-power spectogram we start at the top hierarchical level.
2. On this level, pick a combination of cluster means from the two speaker-dependent clustered data structures. The clustering algorithm doubles the number of clusters for every hierarchical descend, so on this hierarchical level we only have to measure $2 \times 2 = 4$ combinations, and pick the one that satisfy (15).
3. Descend one hierarchical level in the direction of the picked combination.
4. If not on bottom hierarchical level go to step-2.

...this strategi yields *four times the number of levels* comparisons — in contrast, exhaustive searching among combinations of all typical spectra would yield a number of comparisons growing exponentially with the number of typical spectra.

This substantial reduction in calculations has enabled us to do "fast" separation of speech. Figure 4 and 5 show two speech signals prior to their mixing which is shown in Figure 6. As promised, we calculate the log-power spectogram of the measured signal and Figure 7 shows the result.

Based on the spectogram from Figure 7 we use the proposed hierarchical algorithm to find estimates of individual spectograms (shown in Figure 8 and Figure 9).

Equation (16) gives gives the masking signals shown in Figure 10 and Figure 11, and using those to refilter the mixed signal the procedure ends up in finding the separated signals shown in Figure 12 and Figure 13.

This example of separation was performed on a 800MHz Intel machine using Matlab. The complete separation process took less than 10 seconds making us witnessing *a real-time algorithm for separation of speech*.

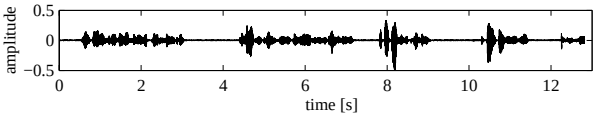

Fig. 4: Female speech signal.

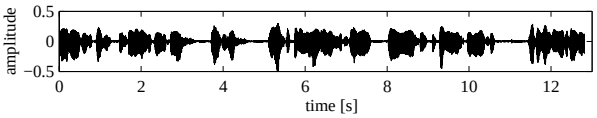

Fig. 5: Male speech signal.

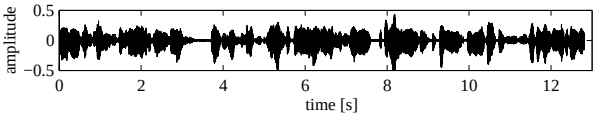

Fig. 6: Mixture of female and male speech.

WRAP UP

We have shown that knowledge of source autocorrelation is necessary for monaural separation. By training a pattern recognition system on the speaker spectrograms we obtain a model capable of handling the multiple local autocorrelation functions for each speaker. In order to make the training feasible we have developed the Hierarchical Vector Quantization making real-time monaural separation possible.

ACKNOWLEDGEMENTS

This paper refer work done by the authors as their Masters Thesis at Technical University of Denmark. It was made under excellent and appreciated supervision of Prof. Lars K. Hansen.

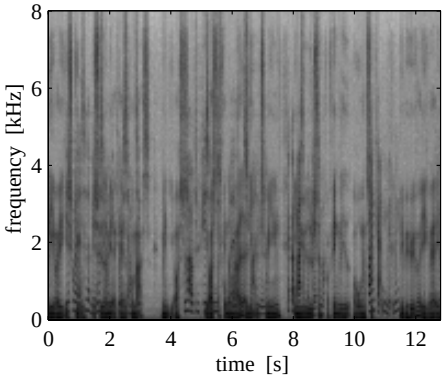

Fig. 7: Log-power spectogram of mixed speech.

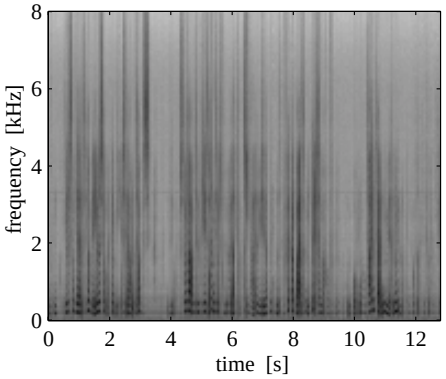

Fig. 8: Result of hierarchical matching typical female spectra.

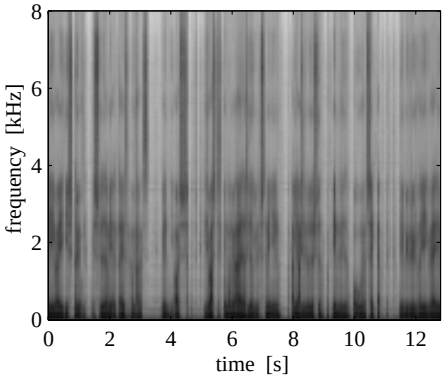

Fig. 9: Result of hierarchical matching typical male spectra.

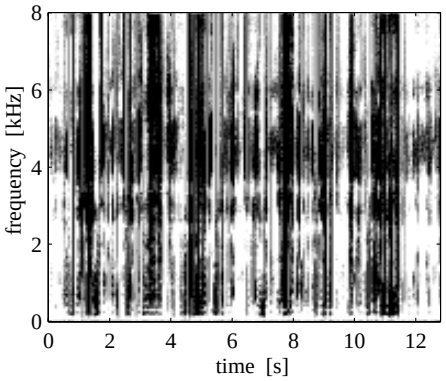

Fig. 10: Found female masking signals.

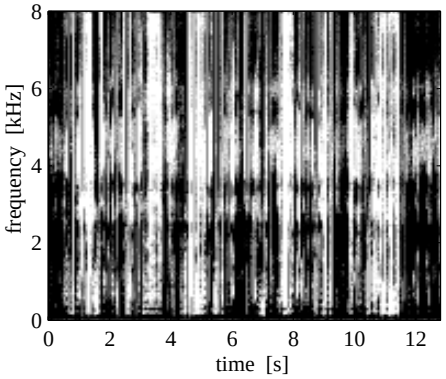

Fig. 11: Found male masking signals.

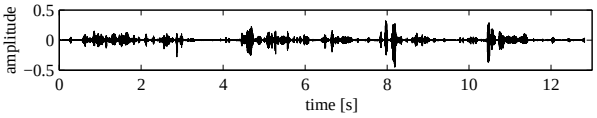

Fig. 12: Female speech separated by refiltering the mixture.

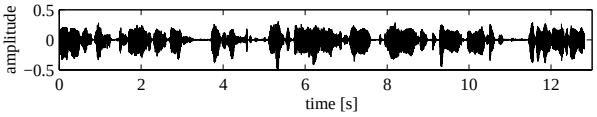

Fig. 13: Male speech separated by refiltering the mixture.

REFERENCES

[1] L. K. Hansen and K. B. Petersen. Monaural ica of white noise mixtures is hard. In *Proceedings of ICA'2003 Fourth Int. Symp.. on Independent Component Analysis and Blind Signal Separation, Nara Japan, April 4,*, 2003.

[2] Mads Dyrholm and Niels Henrik Pontoppidan. Blind signalseparation. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, January 2002.

[3] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. ICASSP '97*, pages 1331–1334, Munich, Germany, 1997.

[4] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE-NN*, 10(3):684, 1999.

[5] Sam Roweis. One Microphone Source Separation. In *Neural Information Processing Systems 13 (NIPS'00)*, 2000.