UNDERDETERMINED NOISY BLIND SEPARATION USING DUAL MATCHING PURSUITS

Paul Sugden and Nishan Canagarajah

Department of Electrical & Electronic Engineering
University of Bristol
Merchant Venturer's Building
Woodland Road, Bristol, U.K.

ABSTRACT

Underdetermined blind source separation is a key application in audio where it is desirable to extract multiple sources from a stereo recording. A new variant on the stereo matching pursuit, the dual matching pursuit, is presented whereby independent matching pursuits are run on both channels of a stereo mixture of greater than two sources. By identifying correlating atoms from each decomposition, a histogram plot is applied to identify the position of each source in the stereo image and the atoms grouped to recover the original signals. To improve the atomic correlation between channels, a fixed overcomplete representation for each of the signal types present in the mixtures is obtained by applying a learning algorithm to existing sources of that type and reducing the redundancy in the resulting basis set via a correlation-based algorithm. The resulting dictionaries are then used as a time-frequency basis for the independent matching pursuits. The results show improved separation quality compared to the dual matching pursuit with mathematical time-frequency dictionaries. The noise immunity of this method due to the use of overcomplete representations is also demonstrated showing that the system can withstand mixture signal-to-noise ratios down to 30dB.

1. INTRODUCTION

Underdetermined blind separation has been achieved using overcomplete methods, both mathematical [1] and learned [2]. Separation quality is always significantly lower than the determined case where methods such as independent component analysis can be used. Here, a system using a combination of the two previously described methods is described, allowing good quality separation of several sources from stereo mixtures.

We can model noisy stereo audio as a mixture of I sources as follows

$$y_c(t) = \sum_{i=1}^{I} x_i(t) + n_c(t), \qquad c = l, r \tag{1}$$

where we assume that the mixing environment is stationary and linear. A panpot parameter Θ_i is used to describe the mixing balance between channels; $\Theta_i = 0$ represents a signal i solely in the left channel and $\Theta_i = \pi/2$ a source solely in the right channel. The stereo mixture $y(t) = (y_l(t), y_r(t))$ can therefore be modelled as

$$y(t) = \sum_{i=1}^{I} \lambda_i(\cos\Theta_i\ x_i(t), \sin\Theta_i\ x_i(t)) + n(t) \tag{2}$$

where λ_i is a gain parameter.

The majority of audio sources can be described as vectors in a Hilbert space $\mathcal{H}$. Signal x_i can have a sparse decomposition in a basis function dictionary $\mathcal{D}$ such that $x_i = \sum_k a_{i,k} g_k$ where $g_k \in \mathcal{D}$ and $\{a_{i,k}\}$ has a fast decay as $k \to \infty$. Hence, these signals can be decomposed effectively using an iterative algorithm like the matching pursuit [3].

2. DUAL MATCHING PURSUIT

The stereo matching pursuit described in [1] demonstrated a novel means of exploiting the correlations between the channels of stereo recordings in order to achieve underdetermined separation. One disadvantage of this system is that each atom calculated in the pursuit is the result of a minimization of the residual across both channels. This results in a fully correlated decomposition of both channels (as all resulting atoms have a coefficient in both y_l and y_r) but possibly compromises the effectiveness of the original algorithm in relation to the matching pursuit.

The dual matching pursuit uses independent matching pursuits on each channel, eliminating this problem. The cost of this approach that full atomic correlation between channels is highly unlikely. However, the logarithmic decay of the residual during the matching pursuit means that most of the energy within the signal is removed during the initial iterations. These high-energy atoms are the most likely to provide a match between the two decompositions as they represent fundamental parts of the mixed signals. Hence,

0-7803-8484-9/04/$20.00 ©2004 IEEE

the unmatched atoms are likely to be of lesser amplitude and their absence will not adversely effect the separated signal.

With dictionary $\mathcal{D}$, we decompose each mixture $y(t)$ using the standard matching pursuit ([3]). For a decomposition of $y(t)$ into $M-1 \geq 0$ atoms, an M-atom matching pursuit is obtained the following way:

1. Compute $|\langle R^{M-1}, g\rangle|$ for all $g \in \mathcal{D}$
2. Select the best atom of the dictionary using

$$g_m := \arg\max_{g \in \mathcal{D}} |\langle R^{M-1}, g\rangle| \tag{3}$$

3. Compute the new residual

$$R^M(t) := R^{M-1}(t) - \alpha_m g_m(t) \tag{4}$$

with $\alpha_m := \langle R^{M-1} g_M \rangle$.

At this point, we make the assumption that each individual atom calculated from the matching pursuit is strongly associated with a particular source i. This is reasonable, as uncorrelated sources have a low probability of sharing elements localized in both frequency and time. The implication of this is that an atom appearing in both channel decompositions should have coefficients that reflect the panpot parameter θ of the original mixture. Therefore, if atom $g_{m,L}$, obtained from iteration m of the left channel matching pursuit, also appears at iteration n from the matching pursuit of the right channel then the coefficient values can be used to calculate $\theta_{m,n}$ using

$$\theta_{m,n} = \tan^{-1}\left(\frac{\alpha_{m,L}}{\alpha_{n,R}}\right) \tag{5}$$

where $\alpha_{m,L}$ and $\alpha_{n,R}$ are the left and right coefficients respectively. When all atoms are correlated between channels, the resulting values of θ_n are analyzed using a histogram plot. The resulting peaks of such a plot should correspond to the individual panpot parameters Θ.

We can obtain a demixing matrix from the values of Θ obtained from the histogram. However, linear demixing cannot completely separate the sources in the overcomplete case ($M > 2$). In order to recover the sources, the histogram is divided into $\hat{I}$ clusters $K_i := \{m \ : m \in K_i\}$. Summing the components of each cluster then gives us a nonlinear estimate of the source (including gain)

$$\widehat{\lambda_i x_i} := \sum_{m \in K_i} \alpha_m g_m \tag{6}$$

This method is derived from the stereo matching pursuit algorithm of [1]. Here, a mathematical time-frequency basis, the multiscale Gabor dictionary, is used to decompose signals using the matching pursuit. Cosine packets [4] can also be used to a similar effect.

Improvement to the separation quality can be obtained by increasing the correlation of atoms between channels. This can be achieved by selecting a dictionary that can span the original source signals in a sparse manner more efficiently than traditional time-frequency sets. The learned overcomplete representation [5], [6] is a means of determining an overcomplete basis ϕ_i for a signal $s_i(t)$ using a maximum likelihood and gaussian around the posterior fit. In [7] this system was used to create large signal dictionaries $\Phi = [\phi_1 \ldots \phi_i]$ by learning representations for i sources of a certain type and then reduced in size to create $\mathcal{D}_r$ using a correlation based algorithm with negligible effect representation quality. It is possible to use the instances of $\mathcal{D}_r$ for underdetermined blind source separation with a linear program solver as shown in [7]. Here, the instances of $\mathcal{D}_r$ are used as a basis for the matching pursuit.

Using the learning algorithm in equation 7, an overcomplete signal dictionary of size q with element length n can be learned from an individual source.

$$\Delta\mathbf{A} = \mathbf{A}\mathbf{A}^T \frac{\delta}{\delta\mathbf{A}} \log P(\mathbf{x}|\mathbf{A}) \approx -\mathbf{A}(\phi\hat{\mathbf{s}}\hat{\mathbf{s}}^T + \mathbf{I}) \tag{7}$$

where $\mathbf{A}$ is a $q \times n$ matrix where the the resulting elements are stored and $\phi(\hat{s}_i) = \delta \log P(s_k)/\delta\hat{s}_i$ and is the cost function.

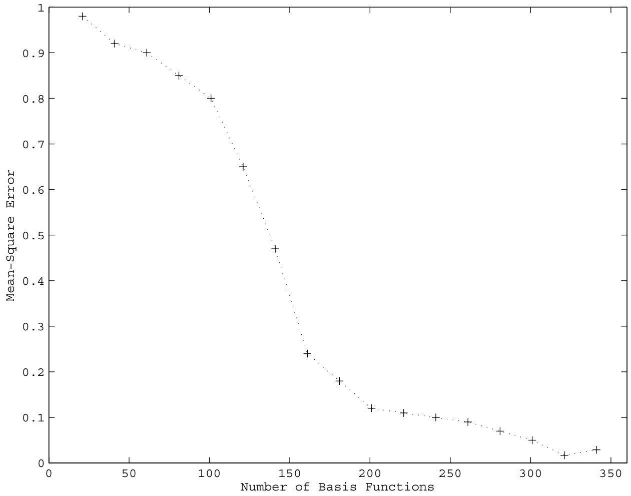

Fig. 1. An example of the tradeoff between number of elements in an overcomplete source dictionary and mean-square error after decomposition of a source of the same type. A good compromise point is at around 100 elements.

A selection of p speech signals from the same or similar source were analyzed using [5] and a basis function set of q was obtained from each. These sets were combined to produce a set of $z = p \times q$ functions $\mathbf{\Phi} = (\phi_1 \ldots \phi_z)$. A correlation table $\mathbf{\Lambda}$ was built, whereby individual functions were compared with each other function and the l most similar found as seen in equation 8.

$$\mathbf{\Lambda} = \max_{1\ldots l}(\mathrm{corr}(\mathbf{\Phi}^T, \mathbf{1}\cdot\mathbf{\Phi})) \qquad (8)$$

Using $\mathbf{\Lambda}$, the most similar elements of the combined library can be removed. The results in [7] show that a 40-60% of elements could be removed whilst retaining the ability of the library to represent signals from a similar source. An example of this tradeoff between number of elements and effectiveness of representation (in terms of mean-square error) can be seen in Figure 1. The reduced library can be used in conjunction with the dual matching pursuit as it can be considered to span the source variables effectively.

3. RESULTS

For the simulations, four different source types sampled at 8kHz and $n = 2^{14}$ samples long (approx. 2 seconds) were used. These were selected to represent a wide variety of frequencies and sound textures. Sources one to four were an acoustic guitar, a synthesizer, a vocal harmony and an electric bass respectively.

From ten samples of each source type, 64 basis functions were learned making a total of $64 \times 10 \times 4 = 2560$ elements in the learned library. Using the technique from [7], the error trade-off allowed a reduction to a total of 1260 elements. This basis set was then used for the dual matching pursuit.

Stereo mixtures were made as in equation 2. The four sources were evenly spaced between 0 and $\pi/2$, i.e. $\Theta = [\pi/10 \quad \pi/5 \quad 3\pi/10 \quad 2\pi/5]$ and gain was uniform i.e. $\lambda = [1 \quad 1 \quad 1 \quad 1]$. These parameters were chosen for clarity and can be varied freely. For each matching pursuit, a total $M = 1200$ iterations was used. If atoms were repeated during the matching pursuit, the coefficients were summed before the θ calculation in equation 5. For comparison, a matching pursuit using cosine packets was also performed.

3.1. Qualitative Results

Figure 2 shows the histogram plot for the dual matching pursuit using both cosine packets and the learned basis. The learned basis provides a considerably more defined set of peaks, especially for the fourth source. The channel correlation is higher using the learned basis with 85% of iterations matching compared to only 53% using cosine packets. Recovering the sources using the clustering method and equation 6 obtained signal-to-noise ratio (SNR) results of between 26dB and 30dB. Waveform plots can be seen in Figure 3. They clearly show that the learned dictionary provides a superior fit to the signal in comparison to the cosine packet method.

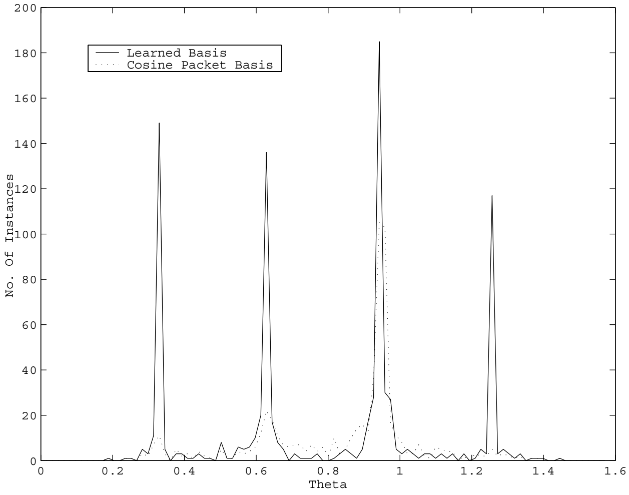

Fig. 2. Histogram plot of the values of θ obtained from the dual matching pursuit using both a cosine packet and a learned reduced basis. The range 0 to $\pi/2$ is divided into 100 bins.

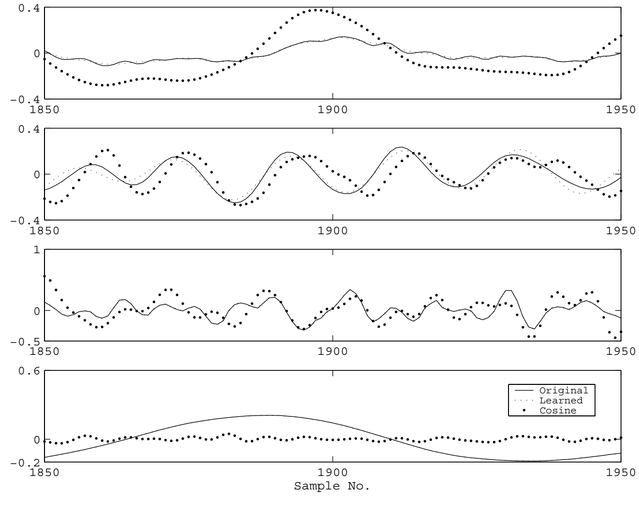

Fig. 3. Waveform plots of 100 samples from sources 1 to 4 (top to bottom) (solid line) and inferred sources using the dual matching pursuit using the cosine packet basis (●) and the learned reduced basis ($\cdots$).

One advantage of using sparse decomposition is the provision of a certain amount of noise immunity. When using traditional Independent Component Analysis (ICA) for source separation, noise is not generally considered as it makes the pdf estimation integral intractable [8]. To test the noise immunity of the dual matching pursuit using a learned basis, additive white Gaussian noise (AWGN) was added to each stereo mixture, creating SNRs of 10dB to 70dB at 10dB intervals. Figure 4 shows the SNRs for the recovered sources at each noise level. At very low SNRs, the separated results are poorer than the original mixtures due to low atom correlation between channels. Above 30dB, SNR of the recovered sources stays relatively constant, suggesting noise immunity above this level. The small glitch at 40dB is most likely due to the particular suitability of the learned dictionary elements at this specific noise level.

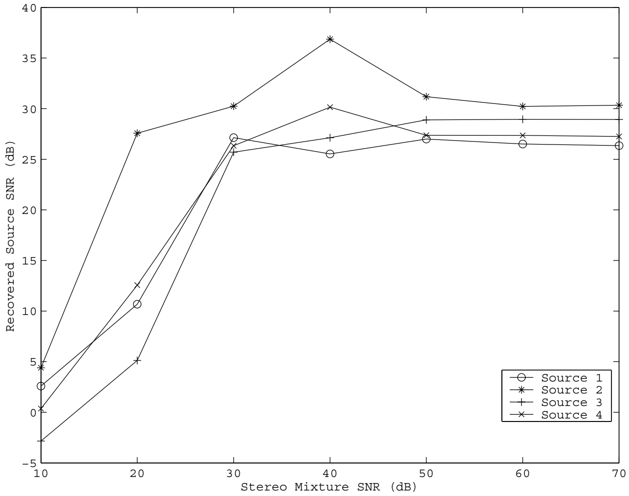

Fig. 4. Plot showing recovered source SNR against varying levels of AWGN added to the stereo mixture.

3.2. Subjective Quality

In terms of the perceived quality of separation, due to the fact that some of the information is lost due to atoms not correlating between channels, perfect reconstruction will not be possible using this method and there is an obvious degradation of quality in all sources. Despite this, separation is of good quality in sources where there are few transients, as in the case of the vocal harmony, or where the sound is structured as in the synthesizer sample. The least successful is the acoustic guitar, where the picking transients are not well represented.

In comparison to the separation achieved with cosine packets, all sources were superior with the exception of the synthesizer, which particularly suited to the sinusoidal nature of the cosine packets. In this case, the separation quality was about equal between the methods.

4. CONCLUSION

It can be seen that the dual matching pursuit is an effective means of performing underdetermined blind source separation. It is also demonstrated that using a learned basis function dictionary as a decomposition set for the dual matching pursuit has produced superior separation quality compared to the use of traditional time-frequency dictionaries. Finally, experiments with varying degrees of gaussian noise in the source mixtures show that the dual matching pursuit method has a reasonable noise immunity and that separation quality is relatively constant above SNRs in excess of 30dB.

5. REFERENCES

[1] R. Gribonval, “Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture,” *Proc. Int. Conf. Acoust. Speech Signal Process (ICASSP 2002)*, vol. 3, pp. 3057–3060, 2002.

[2] Te-Won Lee et al, “Blind source separation of more sources than mixtures using overcomplete representations,” *IEEE Transactions on Signal Processing*, vol. 11, no. 2, pp. 417–441, 1999.

[3] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[4] Ronald Raphael Coifman, Yves Meyer, Stephen R. Quake, and Mladen Victor Wickerhauser, “Signal processing and compression with wavelet packets,” pp. 77–93, 1993.

[5] Michael S. Lewicki and Terence J. Sejnowski, “Learning overcomplete representations,” *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[6] Michael Zibulevsky and Barak A. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[7] Paul Sugden and Nishan Cangarajah, “Underdetermined blind separation using learned basis function sets,” *IEE Electronic Letters*, vol. 39, no. 1, pp. 158–160, 2003.

[8] Aapo Hyvärinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.