OPTIMIZING COHERENT DEMODULATION FOR IMPROVED SEPARATION OF OVERLAPPING SOURCES

Gregory Sell

Institute for Systems Research, Electrical and Computer Engineering Department
University of Maryland, College Park, MD, USA
gsell@umd.edu

ABSTRACT

The complex modulators of coherent demodulation make the algorithm a natural fit for source separation, but the overlapping bands typically found in real audio mixtures present interference problems for the algorithm. This paper proposes reframing coherent demodulation as an optimization problem that distributes the energy in overlapping bands according to an optimally low-frequency strategy. The extension is shown to improve separation for sinewave mixtures and for mixtures of speech and music.

Index Terms— Modulation, Optimization methods, Source separation

1. INTRODUCTION

Several new algorithms have been developed in the past several years for deconstructing a signal into a slowly-varying modulator and the corresponding fine-temporal carrier. Coherent demodulation [1] is unique among these algorithms in that it utilizes a complex modulator, while other methods require a real modulator (and, in most cases, a non-negative real modulator). The preservation of complex phase within the modulator is useful for many reasons, most especially that it preserves linearity in mixtures of multiple sources.

Modulation has been used as a source separation criterion in past research [2, 3, 4, 5], and it has also been shown that the use of complex phase can improve source separation for other non-negative real representations [6], so coherent demodulation would seem to be a natural fit for source separation. And, indeed the method has been shown to perform reasonably well for a flute and castanet mixture [7].

However, overlapping components create an important problem for separation via coherent demodulation — interference between bands [8]. In source separation, this interference manifests as sources bleeding into each other and is especially problematic in signals with a rich harmonic profile.

This paper will present an extension to the coherent demodulation algorithm that finds the optimal low-frequency solution for overlapping bands that preserves additivity across components. It will be shown that this extension improves the performance of the algorithm for recovering sources from mixtures of sinusoids, music, or speech.

2. COHERENT DEMODULATION

The coherent demodulation algorithm considers a harmonic signal within a sum-of-products model:

$$s[n] = \sum_{k=0}^{K-1} s_k[n] = \sum_{k=0}^{K-1} m_k[n] \cdot c_k[n]. \quad (1)$$

The carriers c_k are restricted to unit-norm complex and are usually the harmonics of the original signal (as measured by a pitch esimator). As a result, the modulators m_k can be estimated by multiplying the signal by the complex conjugate of the carrier (thus canceling out the carrier) and low-pass filtering. So, the k^{th} modulator is given by

$$m_k[n] = h[n] * (s[n] \cdot c_k[n]^*) \quad (2)$$

where $h[n]$ is a low-pass filter and * denotes complex conjugation. This formulation ensures the modulator is low-frequency and also allows it to include complex phase.

This algorithm is elegant and simple, and it is linear, time-invariant (LTI) as long as the carrier is re-estimated after the time shift.

2.1. The Overlapping Subband Problem

However, a problem arises with overlapping bands. If the estimated carriers are separated by less than the bandwidth of the low-pass filter ($h[n]$ in Eq. (2)), the bands will interfere with each other in the modulator estimation. This issue is avoidable in the single-source case, because the carriers are separated by the pitch, and so the filter bandwidth can be selected to ensure it is less than the lowest pitch. But, in the case of multiple sources, the problem is more difficult to avoid.

If bands do overlap, their modulators will include their own band's energy as well as energy from all other overlapping bands. This will not only corrupt the individual modulator estimates, but also will amplify the energy in the over-

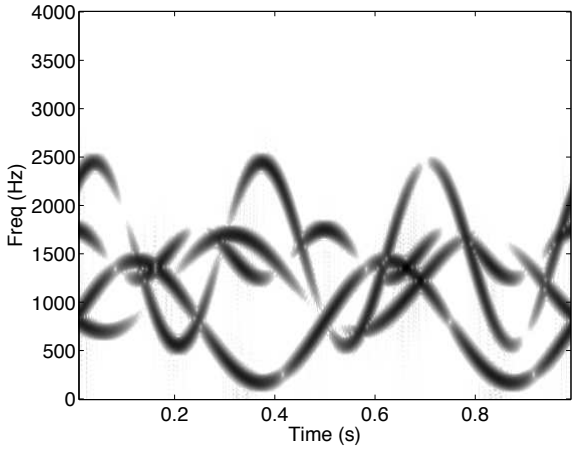


Fig. 1. The spectrogram of a synthetic signal of overlapping components modulated by the sinusoidal signals in Fig. 2(a).

lapping regions for the summation in Eq. (1), because the interfering energy has spread to multiple bands and is therefore added multiple times.

A simple demonstration of this concept is to demodulate a series of overlapping, modulated sinusoids, such as the mixture shown in Fig. 1. This signal is made of four intersecting carriers modulated by the slowly-varying envelopes shown in Fig. 2(a). Coherent demodulation of this signal yields the modulators shown in Fig. 2(b). These modulators are visibly distorted by the interference between overlapping bands.

It is possible to reduce this interference by reframing coherent demodulation as an optimization problem, presented in the next section.

3. AUGMENTING COHERENT DEMODULATION WITH OPTIMIZATION CRITERIA

In order to improve performance with overlapping carriers, we first must require that Eq. (1) hold in all cases, which will force the energy in overlapping bands to somehow be divided between those bands. However, there is an infinite set of strategies for dividing the overlapping energy, such as greedily giving all energy to the nearest carrier, or splitting it equally between all proximate bands. So, the second criterion required is to define the strategy by which the energy should be divided. Because demodulation is fundamentally based on low-frequency modulators, a sensible choice for the strategy is to create the optimally low-frequency set of modulators.

Combining these principles into an optimization problem, the new algorithm can be posed as

$$\text{minimize} \quad \sum_{k=0}^{K-1} ||W\mathcal{F}\{m_k\}||^2 + \lambda||s - \sum_{k=0}^{K-1} m_k \cdot c_k||^2 \quad (3)$$

where $\mathcal{F}$ denotes the Fourier transform and W is a diagonal matrix of frequency-dependent weights.

In this problem, W can be designed to penalize high-frequency content in the modulators m_k (similarly to in [9]), and the second term in the cost function penalizes any deviation from the sum-of-products model in Eq. (1). The regularization parameter λ allows for prioritizing the two costs.

The problem can be further simplified by recognizing that a properly designed W will make modulator frequencies above some cutoff too expensive to include in the modulators. So, the dimensionality of the optimization can be reduced by instead optimizing for the DFT coefficients only up to that cutoff frequency, which for slowly-varying modulators will be quite low (50 Hz in the examples to follow). With this in mind, the optimization in Eq. (3) can be redesigned to solve for the DFT coefficients x_k instead of the modulators m_k.

$$\text{minimize} \quad \sum_{k=0}^{K-1} ||BWx_k||^2 + \lambda||s - \sum_{k=0}^{K-1} (Bx_k) \cdot c_k||^2 \quad (4)$$

The columns of the new matrix B are the DFT sinusoids at the appropriate frequencies for the corresponding coefficients in x_k. As a result, $Bx_k = m_k$.

One may wonder why the sum-of-products cost was not included instead as a linear equality constraint, which would ensure the sum-of-products model holds, instead of simply encouraging it to hold, as the optimization in Eq. (4) does. The selected approach is preferred for several reasons.

The first reason to include the summation in the cost function rather than as a constraint is that, in some cases, such as noisy signals, bandlimited modulators will not be able to perfectly replicate the entire signal (specifically, the regions between widely spaced harmonics). Including the summation in the cost function allows some leniency in these cases.

Second, by posing the problem as it is with no constraints and only l_2-norms, the problem can be solved with simple least-squares minimization rather than requiring slower and more cumbersome gradient descent.

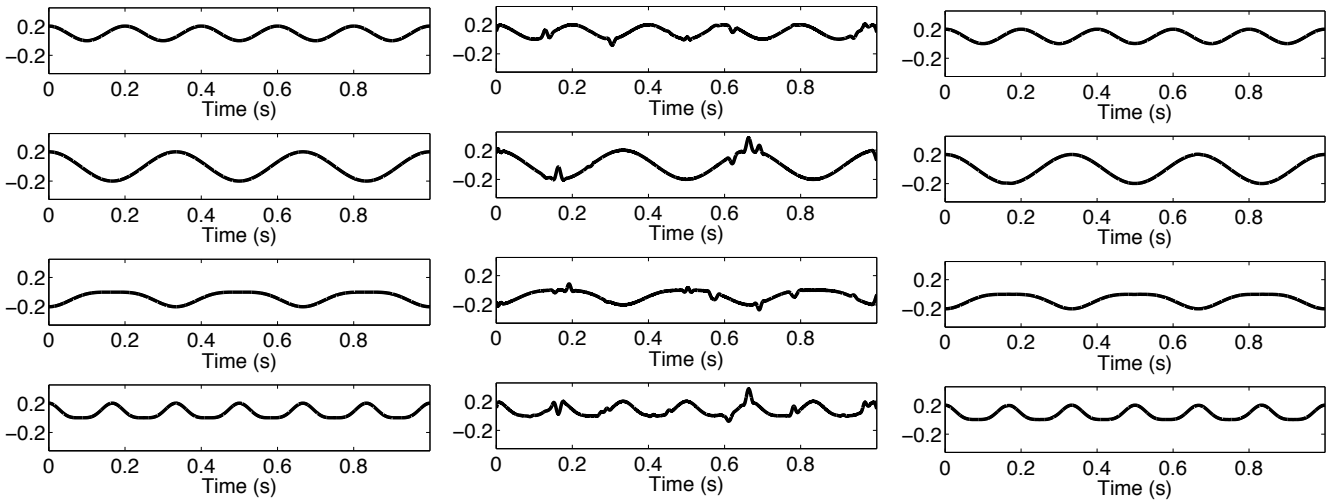
$$\mathbf{x} = (\mathbf{B}_W^T\mathbf{B}_W + \lambda\mathbf{B}_C^T\mathbf{B}_C)^{-1}(\lambda\mathbf{B}_C)^T\tilde{s} \quad (5)$$

In this equation, $\mathbf{x}$ is a vector concatenation of the DFT coefficients for all modulators. $\mathbf{B}_W$ is a block diagonal matrix with the weighted DFT sinusoids BW repeating K times. The matrix $\mathbf{B}_C$ is built from horizontally concatenated repetitions of the matrix B multiplied by each successive carrier c_k. Finally, $\tilde{s}$ is the analytic signal of s. Conceptually, it is only important to understand that Eq. (5) is the solution to Eq. (4).

It is worth noting that in this extension to coherent demodulation, much of the original process remains the same. Essentially, the only change is that the low-pass filter $h[n]$ in Eq. (2) is replaced by a time varying filter $h_t[n]$ whose frequency response at any given time is determined by the optimization.

4. EXAMPLES

In the following examples, the frequency-dependent weighting was set to a linearly increasing value $W(f) = \frac{f}{50}$ and



(a) Original (b) Coherent Estimate (c) Optimized Coherent Estimate

Fig. 2. The modulators applied to each component from the original signal in Fig. 1 are shown in (a). The extracted modulators using the coherent method (b) and the optimized coherent method (c) are also shown. Interference is seen in the coherent modulators while the optimized coherent method suppresses the interference.

the regularization variable was set to $\lambda = 1$. Experimentation with these numbers showed general robustness to small variations, though it is important to increase the weight with frequency if a low-frequency solution is sought.

4.1. Sinusoidal Example

A synthetic example (Fig. 1) was utilized previously to demonstrate the interference between overlapping bands in coherent demodulation (seen in Fig. 2(b)). We can now examine the modulators from the optimized coherent approach, shown in Fig. 2(c). It is clear that the new approach has reduced the interference and resulted in modulators that more closely match the desired modulators. The improvement is also quantifiable, with error reduced by 3 orders of magnitude.

Also note that the second modulator has regions of both positive and negative values, a situation that cannot be modeled with non-negative demodulation algorithms, such as the Hilbert envelope or half-wave rectification.

In the case of added white noise (with the same synthetic base signal), both methods decline smoothly and similarly, with optimized coherent consistently outperforming the original algorithm (error rates shown in Fig. 3).

4.2. Music and Speech Mixtures

It is not surprising that the optimized coherent approach outperforms the original coherent algorithm in the synthetic case above, because it was previously known that, for that example, the optimally low-frequency modulators are the ideal solution. However, it is not as certain or obvious that the optimally low-frequency solution will improve separation for mixtures of real audio sources.

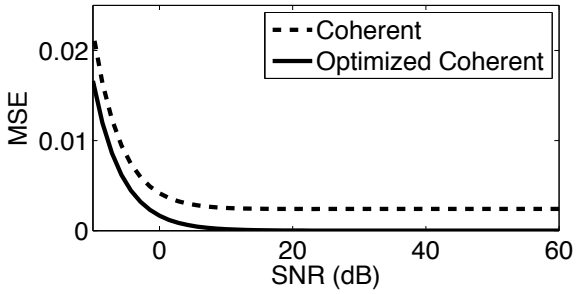


Fig. 3. Mean-squared error for both algorithms applied to the sinusoidal example in Fig. 1 with added white noise.

To test the new algorithm with real audio, 8 signals of 3-5 seconds in length were selected, 4 musical (cello, clarinet, flute, piano) and 4 speech (2 by the same male speaker, 2 by the same female speaker). All signals were single channel and resampled to 16kHz. Every pairwise mixture of these signals was then created (28 total combinations) and separated using both coherent demodulation and the optimized coherent method introduced here. The mixtures were processed in 0.5 second, windowed blocks with a step of 0.25 seconds. It is important to note, though, that pitch estimates were used from the original signals rather than requiring multi-pitch estimation for the experiment.

A comparison of separations by the two algorithms can be seen in Fig. 4, which plots the SNR of the separated signals for each algorithm, with noise defined as the error from the

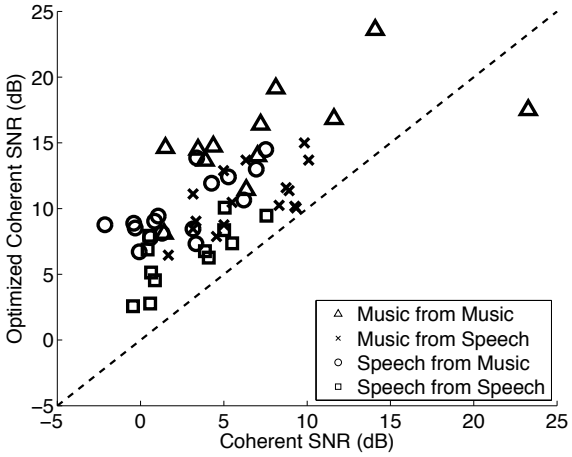


Fig. 4. Power ratios of the component signal to residual noise after separation from music/music, music/speech, and speech/speech mixtures for coherent demodulation (x-axis) and optimized coherent demodulation (y-axis). Data points located above the diagonal line were improved by the optimization extension.

clean source representation for each respective algorithm.

It is clear in the plot that the optimized coherent algorithm improves on standard coherent demodulation in all but one case, with an average improvement of 5.78 dB.

Interestingly, the one example where standard coherent separates better than optimized coherent (though both perform very well) is a flute separated from a piano, which is the same source (flute) that was successfully separated in [7].

It is also worth noting that both algorithms perform best with music sources separated from music sources, and worst with speech signals (especially when separated from another speech signal by the same speaker), though improvement from the optimization extension is similar in all cases.

5. CONCLUSION

This paper introduced a new framing of coherent demodulation that optimally determines the ideal low-frequency set of modulators that simultaneously minimize error in the signal representation. This not only allows the algorithm to preserve the sum-of-products model even in the case of overlapping bands, but it was also shown to improve the separation of real audio sources from mixtures.

The results presented demonstrate a noteworthy improvement over the original algorithm, but they also make an interesting statement about low-frequency modulation decompositions. The algorithm presented finds the optimally low-frequency solution, a solution that improves on previous results. Conceptually, this is an important finding in support of modulation decompositions. In real audio, the foundational principles do indeed translate to improved results.

However, that doesn't necessarily mean that the low-frequency solution is the best approach. While the separations are improved, they are not perfect, and it is quite possible that an alternative strategy would improve performance even further. Fortunately, by shifting the algorithm into an optimization framework, incorporating new strategies such as alternative modulator criteria, cross-band correlation, or source-specific profiles is straightforward. It is hoped that future research will explore these other possibilities.

6. REFERENCES

[1] Steven Schimmel and Les Atlas, "Coherent Envelope Detection for Modulation Filtering of Speech," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2005, pp. 221–224.

[2] John Woodruff and Bryan Pardo, "Using pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[3] Guoning Hu and DeLiang Wang, "Auditory Segmentation Based on Event Detection," in *Workshop on Statistical and Perceptual Audio Processing*, 2004.

[4] Yipeng Li, John Woodruff, and DeLiang Wang, "Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–71, September 2009.

[5] Tuomas Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[6] Brian J. King and Les Atlas, "Single-Channel Source Separation Using Complex Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2591–7, November 2011.

[7] Steven M. Schimmel, Kelly R. Fitz, and Les E. Atlas, "Frequency Reassignment for Coherent Modulation Filtering," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2006.

[8] Pascal Clark and Les Atlas, "A sum-of-products model for effective coherence modulation filtering," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2009.

[9] Gregory Sell and Malcolm Slaney, "Solving Demodulation as an Optimization Problem," *IEEE Transactions on Audio, Speech, and Language Processing*, November 2010.