AUTOMATIC CARRIER PITCH ESTIMATION FOR COHERENT DEMODULATION

Gregory Sell

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA
gsell@jhu.edu

ABSTRACT

This paper proposes a method for estimating carrier frequencies for coherent demodulation using low-frequency amplitude modulation criteria and a frequency-smoothing regularizer. This process combines with coherent modulator estimation to create an iterative approach for simultaneously determining the optimally low-frequency modulator and corresponding carrier for an arbitrary signal. The method is demonstrated with unsupervised semi-blind source separation of speech mixed with several types of tonal interference.

Index Terms— Amplitude modulation, source separation, pitch estimation

1. INTRODUCTION

Coherent demodulation [1] of an acoustic signal defines each harmonic component of a signal as the product of a complex sinusoidal carrier and a complex low-frequency modulator. The use of complex components is advantageous for several tasks, such as source separation [2, 3], but, in order to estimate these modulators, an estimate of carrier frequency is required (unlike other techniques like Hilbert demodulation).

Providing an accurate carrier frequency estimate is a challenge, especially in the presence of interference. Early versions of coherent demodulation used fixed-frequency carriers based on spectral center-of-gravity estimates [1], but those are problematic in the likely case that carriers cross subband boundaries. Instantaneous frequencies were later estimated from conditional mean frequencies in the STFT [4]. Most recently, instantaneous pitch estimates determined carrier frequencies [5]. However, estimating carriers in the presence of interference or multiple sources still remains a challenge.

In this paper, an approach for carrier frequency estimation is proposed that seeks carriers corresponding to low-frequency modulators. This combines with modulator estimation to automatically determine both the optimally low-frequency modulator and corresponding carrier in an unsupervised fashion.

Though this is the first approach to jointly optimize the complex modulator/carrier pairs in coherent demodulation, other algorithms have performed similar estimations. Probablistic amplitude demodulation [6], for example, solves for a frequency-constrained real modulator according to desired carrier statistics. Residual Interfering Signal Cancelers [7] and dynamic tracking filters [8] have also been used to automatically track harmonic or formant components.

This paper will first provide additional background on coherent demodulation, and then the proposed carrier estimation technique will be introduced, followed by several examples.

2. BACKGROUND

2.1. Coherent Demodulation

Coherent demodulation models a signal as the sum of modulated complex sinusoids, called the sum-of-products [9]:

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

As discussed above, the carriers are estimated first. The modulators m_k are determined by multiplying the signal with the complex conjugate of the carrier and low-pass filtering. So, the k^{th} modulator is given by

$$m_k[n] = h_{LP}[n] * (s[n] \cdot c_k[n]^*) \quad (2)$$

where $h_{LP}[n]$ is a low-pass filter.

2.2. Optimization for Overlapping Components

Estimating clean modulators with Eq. (2) requires that the carriers are sufficiently spaced in the frequency domain. Otherwise, the low-pass filter will pick up any interfering energy from neighboring bands.

A solution for this problem was recently proposed [10], which finds the optimally low-frequency set of modulators that satisfies the sum-of-products model in Eq. (1). This approach improved modulation-based source separation for audio mixtures, but it also required *a priori* knowledge of the pitch of each source in the mixture, a significant assumption.

In the section to follow, an approach is proposed for estimating the carrier frequency automatically from the modulators. In [10], it was determined that seeking optimally

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

low-frequency modulators is an effective strategy for source separation, so here we apply the same strategy to carrier estimation. This approach seeks a time-varying version of the center-of-gravity estimate with frequency modulation constraints. A smoothed, time-varying center-of-gravity estimate was also discussed in [11], though in that case the smoothing constraint was controlled via STFT window length.

3. AUTOMATIC CARRIER PITCH DETECTION

In its simplest form, the proposed approach is to find the carrier pitch that results in a low-frequency modulator while remaining smooth. This requires defining constraints on the spectra of both the amplitude and frequency modulation of each component in Eq. (1). The amplitude modulation constraints will encourage low-frequency modulators, and the frequency modulation constraints will encourage a smooth pitch trajectory for the carrier.

To formalize this, assume that we have some initial carrier pitch estimate $p^{(i)}$, which is a vector of R pitch estimates corresponding to R time frames of length K samples each. We can drive a carrier at these frequencies and estimate the associated modulator $m^{(i)}$ via Eq. (2).

$$m^{(i)}[n] = h_{LP}[n] * (s[n] \cdot e^{j2\pi p^{(i)}[n]nT})$$

We want to find some set of carrier frequency offsets $\hat{x}^{(i)}$ so that the modulator associated with a carrier driven at $p^{(i)} + \hat{x}^{(i)}$ is lower frequency than $m^{(i)}$ (while still being sufficiently smooth in the pitch domain). If an optimal $\hat{x}^{(i)}$ is found, the pitch estimate is updated by simply adding the offset, $p^{(i+1)} = p^{(i)} + \hat{x}^{(i)}$.

One way to define $\hat{x}^{(i)}$ is as the solution to a least-squares problem with two cost functions. First, we need a penalty for high frequencies in the amplitude modulation.

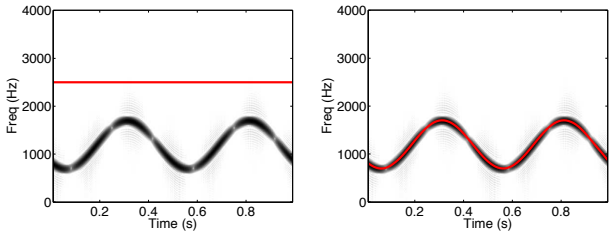
$$C_{AM}(x^{(i)}) = \sum_r |M_r^{(i)}(f - x_r^{(i)})|^2$$

$M_r^{(i)}$ is a matrix whose diagonal is the length K magnitude spectrum of the r^{th} frame of $m^{(i)}$, and f is a vector of modulation frequencies corresponding to the bins in $M_r^{(i)}$. This function weights the spectrum of the modulator by the spectral frequencies (adjusted by the offset $x^{(i)}$), which penalizes high energy at high frequencies. Minimizing this cost function solves for the set of carrier offsets that give a modulator with a spectral center-of-gravity of zero at all frames (as in [11]). However, there are no assurances about the smoothness of the carrier.

Second, the frequency modulation must also be constrained to smooth the updated pitch estimate $p^{(i)} + x^{(i)}$.

$$C_{FM}(x^{(i)}) \quad = \quad ||h_{HP} * (p^{(i)} + x^{(i)})||^2$$

h_{HP} is a high-pass filter whose cut-off frequency defines the degree of frequency modulation allowed in the carrier pitch.



(a) Initialization (b) After convergence

Fig. 1. A demonstration of the proposed algorithm (a) at initialization and (b) after iteration converges, with the carrier estimate in red. The carrier frequency is successfully tracked.

These two cost functions can simply be summed and minimized to solve for the optimal offset $\hat{x}^{(i)}$ that balances low-frequency amplitude modulation and smooth pitch.

$$\hat{x}^{(i)} = \underset{x^{(i)}}{\text{argmin}}\ C_{AM}(x^{(i)}) + \lambda C_{FM}(x^{(i)}) \tag{3}$$

Because this problem is in least-squares form, $\hat{x}^{(i)}$ is the point where the cost function gradient is zero.

$$\hat{x}^{(i)} = A^{(i)-1} b^{(i)} \tag{4}$$

where

$$\begin{aligned} A^{(i)} &= I_K^T M^{(i)T} M^{(i)} I_K + \lambda H^T H \\ b^{(i)} &= I_K^T M^{(i)T} M^{(i)} f_v - \lambda H^T H p^{(i)} \end{aligned}$$

H is a high-pass filter convolution matrix, $M^{(i)}$ is a matrix whose diagonal is the concatenated magnitude spectra of each frame of $m^{(i)}$, f_v is a vector of modulation frequencies corresponding to the entries in $M^{(i)}$, and I_K is a binary matrix of size RK x R with ones in the first column of the first K rows, ones in the second column of the next K rows, and so on.

By iterating between modulator estimation and carrier pitch update estimation, the optimal set of low-frequency modulators and corresponding carriers can be automatically calculated.

It is important to note that an initial estimate for the carrier pitch $p^{(0)}$ is still required to estimate the first modulator $m^{(0)}$. The requirements for this initial estimate are closely tied to the modulator bandwidth, because the amplitude modulation cost function C_{AM} will only consider frequencies that are within the modulator bandwidth (typically less than 100Hz). This means that the initial estimate must be reasonably close to the actual carrier frequency, or the modulator bandwidth must be expanded for the carrier estimation process. Removing this requirement is a goal in future work.

As an initial demonstration of the proposed algorithm, the carrier estimates for a single amplitude- and frequency-modulated sinusoid are shown for initial conditions in Fig. 1(a) and after convergence in Fig. 1(b), where the carrier frequency is accurately estimated.

4. EXAMPLE: SPEECH PLUS INTERFERENCE

We will next examine separation of speech from a mixture using the proposed algorithm. However, due to the potential susceptibility to poor initialization, we require the assumption that the interference is of a higher fundamental frequency than the speech. This implies that the lowest-frequency carrier in the mixture is the speech fundamental f_0.

Based on this assumption, the separation process is as follows:

1. Initialize a carrier at 100Hz; iterate to estimate f_0.
2. Set speech harmonics as multiples of f_0, extract modulators, and subtract components from mixture.
3. Initialize interference carrier with frequency of maximal spectral energy in residual; iterate.
4. Repeat 2-3 until exit criteria met (such as a maximum number of interfering components).
5. Pool all carrier estimates (speech and interference) and update until convergence.
6. Separate speech from interference according to [10].

To visualize this process, Fig. 2(a) shows the spectrogram for a speech signal from the TIMIT corpus (female uttering "The pipe began to rust while new") with sinusoidal interference. In this example, as well as those that follow, the high-pass matrix H is set as a toeplitz matrix of the symmetric, non-causal impulse response $[-\frac{1}{2}, 1, -\frac{1}{2}]$, and the parameter λ is set to 10^7 (the regularizer is large to offset the squaring of the frequency vector f in C_{AM}).

In accordance with step 1, the speech fundamental is estimated and used to estimate the harmonic frequencies for speech (red lines in Fig. 2(b)). The subsequent modulator/carrier pairs are removed from the mixture (step 2), and the carrier frequency of the interferer is estimated in the residual with the same proposed method (step 3), shown with a blue line in Fig. 2(b). These frequency estimates are all simultaneously fine tuned (step 5) and then used to separate the speech (step 6), resulting in the separated spectrogram in Fig. 2(c). The metrics in Table 1 also show the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and the signal-to-artifacts ratio (SAR) for the separation, calculated with the BSS_EVAL toolbox [12]. All metrics show good separation, though SAR is a bit lower than SIR.

This same process was repeated for several tonal interferers (added with equal power to the same speech sample from above): a flute (Fig. 3(a)), a European emergency siren (Fig. 3(b)), and an American emergency siren (Fig. 3(c)). The SDR, SIR, and SAR measurements for the separated speech are shown in Table 1 and the resulting separated speech spectrograms are shown in Fig. 3(d), Fig. 3(e), and Fig. 3(f), respectively. In each case, five interfering carriers were estimated for step 4.

Interferer	SDR	SIR	SAR
Sinewave	17.0 dB	25.2 dB	17.1 dB
Flute	6.8 dB	16.2 dB	7.4 dB
European Siren	7.5 dB	15.9 dB	8.4 dB
American Siren	9.1 dB	17.3 dB	9.9 dB

Table 1. SDR, SIR, and SAR for speech separated from each interfering signal.

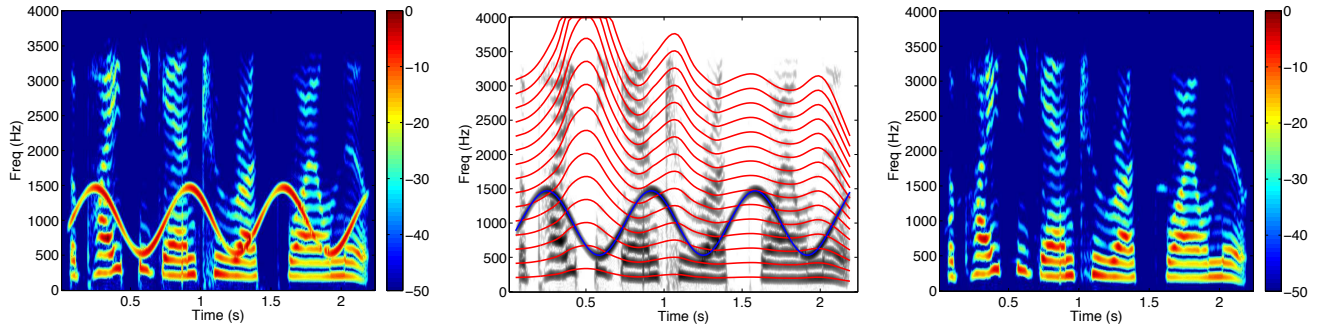
In all cases, the interference is visually reduced in the spectrograms, and the separation metrics also indicate good separation. As was the case with the sinewave separation, artifacts contribute more to the distortion than residual interference. The interference that does remain also gives some insight into areas for future improvement.

In some cases, tonal components in the interference are missed by the residual iterations (such as the harmonic around 900Hz at 2.5 seconds in the speech/flute mixture). Similarly, some of the interference estimates converge to regions of the voice and strip away speech content, reducing SAR. These are both problems that could be fixed with better pitch estimate initializations or a more effectively constrained update.

Non-tonal elements of interference also remain in the reconstructed speech (though these elements are not evident in the spectrograms). In the case of the flute, for example, most of the remaining interference is the flautist blowing on the mouthpiece. Similar non-tonal components persist for the European siren. Eliminating these interferences is a more challenging problem, as the coherent demodulation theory is based on sinusoidal carriers. Removing these elements would require expanding the theory to include non-tonal carriers or developing a modulation speech model to exclude non-speech-like elements.

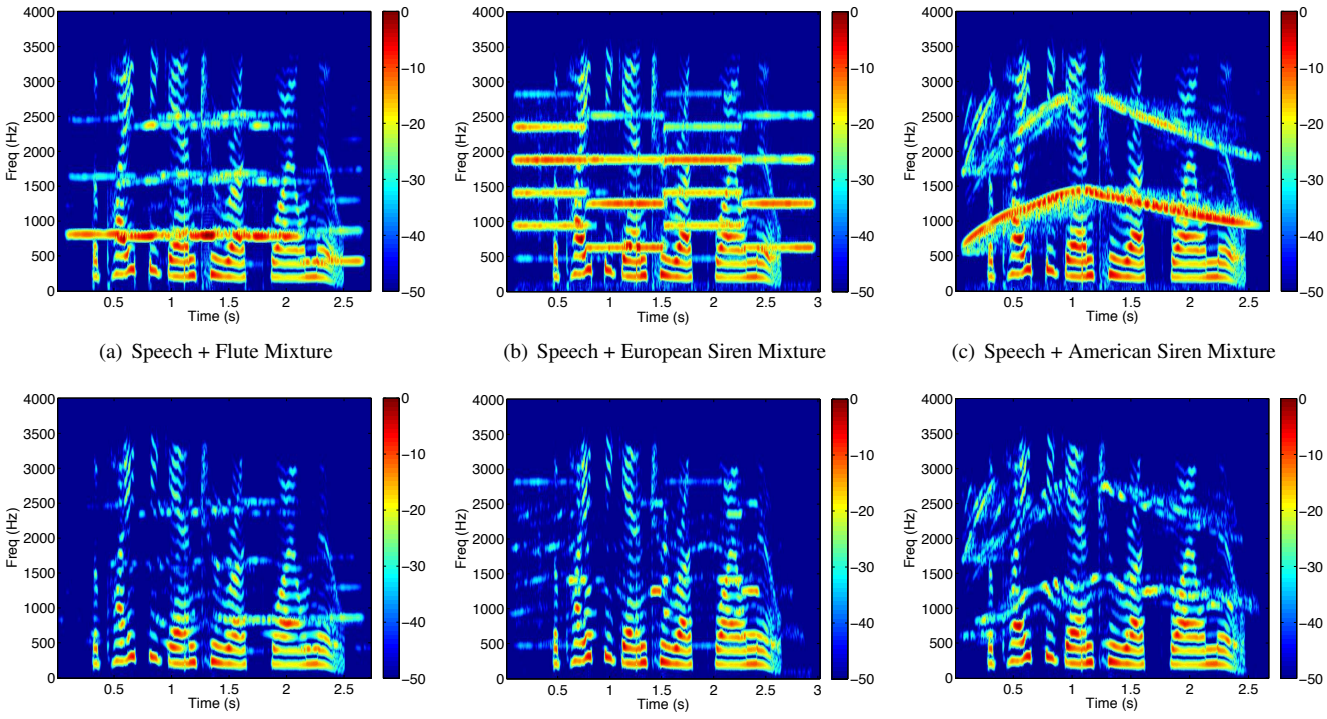
5. CONCLUSION

A new approach for coherent demodulation carrier estimation was proposed, and it was shown how this process jointly integrates with modulator estimation. Several examples were shown in which the approach separates speech from several types of interfering signals in an unsupervised fashion with only an assumption of relative pitch. However, these examples also demonstrated the need for improvement, most especially a more accurate initialization strategy prior to carrier frequency updates. But, even in its current form, the algorithm is an effective means for speech enhancement or separation with minimal available information.



(a) Speech+Sine (b) Frequency Estimates (c) Separated from Sine (15.1 dB)

Fig. 2. Visualizations of several stages of the separation process for speech with an interfering sinusoid: (a) the initial mixture; (b) carrier estimates derived with the proposed method for speech (red) and interference (blue); and (c) separated speech.



(a) Speech + Flute Mixture (b) Speech + European Siren Mixture (c) Speech + American Siren Mixture

(d) Speech Separated from Flute (e) Speech Separated from European Siren (f) Speech Separated from American Siren

Fig. 3. Spectrograms of mixtures and separated speech for several interfering signals: a flute ((a) and (d)); a European siren ((b) and (e)); and an American siren ((c) and (f)).

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] Les Atlas and Christiaan Janssen, "Coherent Modulation Spectral Filtering for Single-Channel Music Source Separation," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2005.

[3] 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.

[4] Qin Li and Les Atlas, "Coherent Modulation Filtering For Speech," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2008.

[5] Xing Li, Naibao Nie, Les Atlas, and Jay Rubinstein, "Harmonic Coherent Demodulation For Improving Sound Coding in Cochlear Implants," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2010.

[6] Richard E. Turner and Maneesh Sahani, "Probabilistic Amplitude Demodulation," in *Independent Component Analysis and Signal Separation*, 2007, pp. 544–551.

[7] C. S. Ramalingam and R. Kumaresan, "Voiced-Speech Analysis Based on the Residual Interfering Signal Canceler (RISC) Algorithm," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 1994, vol. 1, pp. 473–6.

[8] Ashwin Rao and Ramdas Kumaresan, "Dynamic Tracking Filters for Decomposing Nonstationary Sinusoidal Signals," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 1995, vol. 2, pp. 917–20.

[9] 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.

[10] Gregory Sell, "Optimizing Coherent Demodulation for Improved Separation of Overlapping Sources," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[11] Pascal Clark and Les Atlas, "Time-Frequency Coherent Modulation Filtering of Nonstationary Signals," *IEEE TSP*, vol. 57, no. 11, pp. 4323–4332, Nov. 2009.

[12] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–9, July 2006.