AM-FM DEMODULATION OF SPECTROGRAMS USING LOCALIZED 2D MAX-GABOR ANALYSIS

Tony Ezzat, Jake Bouvrie, Tomaso Poggio

Center for Biological and Computational Learning,
McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA
tonebone@mit.edu, jvb@mit.edu, tp@ai.mit.edu

ABSTRACT

We present a method that de-modulates a narrowband magnitude spectrogram $S(f,t)$ into a frequency modulation term $cos(\phi(f,t))$ which represents the underlying harmonic carrier, and an amplitude modulation term $A(f,t)$ which represents the spectral envelope. Our method operates by performing a two-dimensional local patch analysis of the spectrogram, in which each patch is factored into a local carrier term and a local amplitude envelope term using a Max-Gabor analysis. We demonstrate the technique over a wide variety of speakers, and show how the spectrograms in each case may be adequately reconstructed as $S(f,t) = A(f,t)cos(\phi(f,t))$.

Index Terms— speech analysis, spectral analysis, time-frequency analysis, modulation

1. INTRODUCTION

A particularly useful model of a narrowband magnitude spectrogram $S(f,t)$ is

$$S(f,t) = A(f,t)cos(\phi(f,t)) \tag{1}$$

where $cos(\phi(f,t))$ is a 2D spectro-temporal modulation term representing the underlying harmonic carrier, and $A(f,t)$ is a 2D amplitude modulation term representing the overall spectral envelope. In keeping with similar AM-FM approaches used to model 1-D speech signals, we call this model a 2-D AM-FM model of narrowband speech spectrograms.

It is important in many speech applications to be able to *de-modulate*, or separate, the spectrogram into separate AM and FM components, and we present such a method in this work. Our method operates by performing a 2-D local patch analysis of the spectrogram, in which small spectro-temporal patches $P(f,t)$ from the spectrogram are themselves individually de-modulated into local patch carriers $cos(\phi(f,t))$ and local amplitude envelopes $A(f,t)$.

Our algorithm works in two steps: In the first step, the local carrier $cos(\phi(f,t))$ within a patch is estimated. The basic assumption made here is that the underlying carrier belongs to a parameterized family of 2-D spectro-temporal Gabor filters. Our algorithm thus finds the "best-fit" 2-D Gabor filter for each patch, an analysis which we term Max-Gabor analysis. In our previous work [1], we also used a Max-Gabor analysis for patch carrier estimation, but in this work we robustify its use even further by estimating the carrier from patch *gradients* rather than from raw patch values.

In the second step of our algorithm, a local amplitude envelope $A(f,t)$ for that patch is estimated using the local carrier $cos(\phi(f,t))$ obtained from the previous step. In prior work [1], we assumed that the ampltitude envelope was constant over the patch, which is clearly an inadequate assumption since amplitude modulations can vary significantly within a single patch. In this work, we estimate a smooth but non-constant local ampltitude envelope for each patch using scattered data interpolation techniques.

We also demonstrate how to overlap-add the estimated local patch carriers and envelopes to construct carriers and envelopes for the entire spectrogram. Additionally, we demonstrate how to obtain an estimate of the smooth phase surface $\phi(f,t)$ for the entire spectrogram. Finally, we demonstrate how the spectrograms in each case may be adequately reconstructed as $S(f,t) = A(f,t)cos(\phi(f,t))$.

It is instructive to note that prior work on AM-FM demodulation of signals has been applied either to 1-D speech signals [2], or to 2-D images [3] [4], but surprisingly never to magnitude spectrograms! Additionally, these previous works rely on using either Kaiser-Teager energy separation algorithms [3] [2] or on analytic Hilbert computations [4] to demodulate the patch AM and FM components *simultaneously*, from all points in the patch. Instead, we rely on a simpler two-step algorithm which first estimates the FM carrier, and then estimates modulating AM envelope.

We discuss the individual steps of our algorithm in the sections below.

2. LOCAL CARRIER ESTIMATION

2.1. 2D Gabor Carrier Family

We define a family of spectro-temporal 2D Gabor carriers $C(f,t)$ parameterized by spectro-temporal frequency F, spectro-

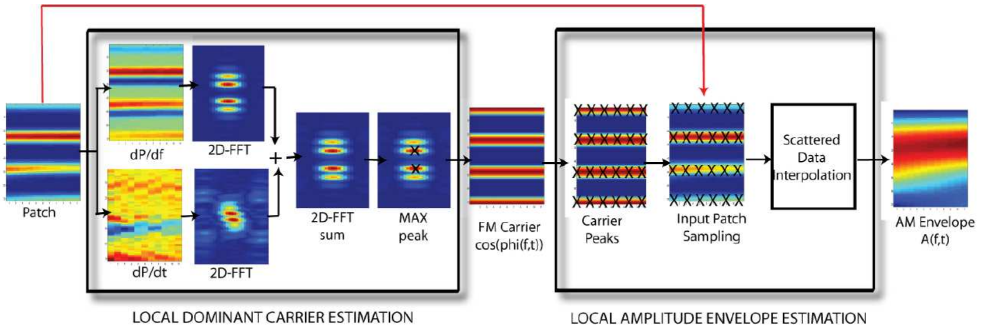

Fig. 1. Overview of AM-FM patch demodulation using Max-Gabor Analysis.

temporal orientation Θ, and phase Φ as:

$$C(f,t) = W(f,t) \cdot cos(2\pi F\hat{x} + \Phi) \quad (2)$$

where

$$\hat{x} = tcos\Theta + fsin\Theta \quad (3)$$

and $W(f,t)$ is a symmetric 2D Gaussian window.

2-D spectro-temporal Gabors look like sets of oriented lines on the 2-D spectro-temporal plane, and as such are especially well-suited to model harmonic carriers in speech. The parameter F controls the spacing between the lines; the parameter Θ controls the orientation of the lines; the parameter Φ controls the position of the lines within the local patch grid.

It is well-known [5] that the Fourier transform of a 2-D Gabor looks like a pair of conjugate Gaussian "peaks", whose distance from each other is proportional to F, and whose orientation is proportional to Θ. As will be evident, Max-Gabor analysis relies heavily on this fact in estimating the local patch carrier. (This same fact was used independently by [6] for pitch-tracking).

It is highly instructive to re-write Equation 2 as

$$G(f,t) = W(f,t) \cdot cos(\phi(f,t)) \quad (4)$$

where

$$\phi(f,t) = 2\pi F\hat{x} + \Phi \quad (5)$$

represents a *local planar phase surface* corresponding to the Gabor carrier. As we are interested in ultimately reconstructing our spectrograms, our AM-FM demodulation algorithm must keep track of this phase surface across the spectrogram.

Finally, we point out that, since magnitude spectrograms are non-negative, all of our computations in Equations 2 or 4 involve *rectifying* the carriers by setting their negative compononents to zero.

2.2. 1D STFT

All of the 16KHz utterances we consider are first STFT analyzed using a 25msec Hamming window with a 1ms frame rate and a zeropadding factor of 4. This yields 1600 dimensional STFT frames, which are truncated to 800 bins due to the symmetry of the Fourier transform. We limit our analysis in this paper to the magnitude spectrogram of each utterance, which we represent notationally as $S(f,t)$. Additionally, we limit our analysis to a linear frequency axis, deferring logarithmic frequency analysis to future work.

2.3. Patch Extraction

At every grid point (i,j), we extract a patch $P_{ij}(f,t)$ of the spectrogram of size df and width dt. The height df and width dt of the local patch are important analysis parameters: they must be large enough to be able to resolve the underlying local dominant carrier, but small enough so that the underlying signal is locally stationary. Suitable parameter ranges are 5-15msec for the dt parameter, and $600Hz - 800Hz$ for the df parameter. Additional analysis parameters are the window hopsizes in time Δi and frequency Δj. Typically we set Δi to be 3-5ms and Δj to 150-350Hz, which creates overlap between the patches.

2.4. Patch Gradients

For every patch $P_{ij}(f,t)$, we compute its spectral gradient $\frac{\delta P_{ij}}{\delta f}$ and its temporal gradient $\frac{\delta P_{ij}}{\delta t}$ using simple local second-order differences. As shown in Figure 1, computing gradients highlights the local edge details in the patch. This allows us to determine the underlying carrier *independent of the local patch ampltitude levels*. The spectral gradient $\frac{\delta P}{\delta f}$ highlights horizontal edges, which usually relate to speech harmonics, while the temporal gradient $\frac{\delta P}{\delta t}$ hightlights vertical edges, which usually relate to speech transients. Of course, we are not just limited to taking vertical and horizontal derivatives, and one can imagine augmenting our analysis with a whole bank of other directional derivative filters. However, for the sake of computational simplicity, we limit ourselves in this work to horizontal and vertical gradient computations.

2.5. 2D Local FFT

A local 2D FFT analysis is then performed on patch gradients $\frac{\delta P_{ij}}{\delta f}$ and $\frac{\delta P_{ij}}{\delta t}$ separately: First, we multiply each patch gradient by a 2D Gaussian window $W(f,t)$ of the same size as the patch. Second, a 2-dimensional Fourier transform of size $N_H \times N_W$ is performed on each windowed patch gradient to produce the local spectral-temporal gradient magnitude spectrum:

$$R^x_{ij}(\Omega,\omega) = \left\| \sum_f \sum_t W(f,t) \frac{\delta P_{ij}(f,t)}{\delta x} e^{-j2\pi \frac{\Omega}{N_H} f} e^{-j2\pi \frac{\omega}{N_W} t} \right\|$$

Finally, we sum the spectro-temporal gradient magnitude spectrum for each patch gradient component to produce the total combined magnitude spectrum for that patch:

$$R_{ij}(\Omega,\omega) = R^f_{ij}(\Omega,\omega) + R^t_{ij}(\Omega,\omega) \tag{6}$$

Typical values N_H and N_W are 256 and 48 respectively.

2.6. Maximum Peak Detection

Visual inspection of $R_{ij}(\Omega,\omega)$ for different patches reveals that most of the spectra exhibit a Gabor-like spectral structure. As shown in Figure 1, this is exemplified by the presence of two dominant Gaussian-like peaks in the spectrum whose location we wish to identify. Additionally, a set of smaller similarly-oriented peaks exist because the magnitude STFT is non-negative. A host of other local peaks also emerge due to noise in the patch.

We use a simple peak-detection strategy to obtain a set C of candidate peak locations and values in the spectral response $R_{ij}(\Omega,\omega)$. We match the conjugate peak locations in C with each other into pairs and throw out any peak candidates which do not have matching conjugate peaks. Finally, we choose among the peak pairs in the set C *the one with the largest peak value*. This pair will comprise our estimate for the underlying dominant Gabor carrier in the patch.

2.7. Local Carrier Parameter Estimation

The local carrier orientation $\Theta(i,j)$ and frequency $F(i,j)$ may be estimated from the chosen peak pair as

$$\Theta(i,j) = tan^{-1}\left(\frac{\Delta\Omega_{max}}{\Delta\omega_{max}}\right) \tag{7}$$

and

$$F(i,j) = \frac{\sqrt{\left(\frac{\Delta\Omega_{max}}{N_H}\right)^2 + \left(\frac{\Delta\omega_{max}}{N_W}\right)^2}}{2} \tag{8}$$

where $\Delta\Omega_{max}$ and $\Delta\omega_{max}$ refers to differences between the conjugate pair location coordinates. Local carrier phase $\Phi(i,j)$ is estimated by projecting the input patch $P_{ij}(f,t)$ onto a complex Gabor $C^*(f,t) = W(f,t) \cdot e^{j(2\pi F\hat{x})}$ with local frequency $F(i,j)$ and local orientation $\Theta(i,j)$:

$$\Phi(i,j) = angle\left(\sum_f \sum_t W(f,t) P_{ij}(f,t) C^*(f,t)\right)$$

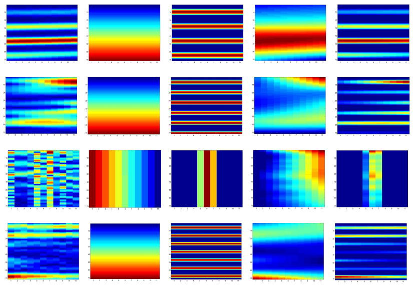

Fig. 2. Left column: Sample input patches $P_{ij}(f,t)$, 2nd column: Estimated phase surface $\phi_{ij}(f,t)$, 3rd column: Estimated carrier $cos(\phi_{ij}(f,t))$, 4th column: Estimated amplitude envelope $A_{ij}(f,t)$, Final column: AM-FM approximation to each patch $A_{ij}(f,t)cos_{ij}(\phi(f,t))$

Given estimates $F(i,j)$, $\Theta(i,j)$, and $\Phi(i,j)$ for each patch, we can synthesize the local phase surface $\phi_{ij}(f,t)$ and the local carrier $cos(\phi_{ij}(f,t))$ using Equation 5.

Shown in Figure 2 are example input patches, as well as the estimated local carriers and phase surfaces for each.

3. LOCAL AMPLITUDE ENVELOPE ESTIMATION

3.1. Carrier Peak Detection and Input Patch Sampling

To estimate the local amplitude envelope $A_{ij}(f,t)$, the local carrier $cos(\phi_{ij}(f,t))$ is first thresholded for values greater than 0.95. This yields a set of N locations $\{f_i, t_i\}_{i=1}^N$ that represent the locations of the carrier peaks. The input patch $P_{ij}(f,t)$ is then *sampled at these locations*, yielding a set of value-location triples $V = \{A_k, f_k, t_k\}_{k=1}^N$. The set V is figuratively shown in Figure 1 as the set of “X” marks on the right-hand side of the figure.

3.2. Scattered Data Interpolation

A *scattered data interpolation* approach is used to interpolate the set of points V and fill in values for the entire local amplitude envelope $A_{ij}(f,t)$. This is done by minimizing an error E that contains a target term which penalizes envelopes that do not match the sampled points in V, and a gradient smoothness term that penalizes local envelopes that are not smooth:

$$E = \underbrace{\sum_k (A(f_k,t_k) - A_k)^2}_{target\ term} + \lambda \underbrace{\sum_{f,t} |\nabla A|^2}_{smoothness} \tag{9}$$

In this work, we implement Equation 9 using the Matlab routine `gridfit`, with λ set to 30. Shown in Figure 2 are examples of estimated local amplitude envelopes for various local input patches.

Intuitively, our algorithm computes the envelope only from points located at *the peaks of the underlying estimated carrier*, throwing out all other samples. This makes our method more robust than other methods which estimate the envelope from all the points in the patch.

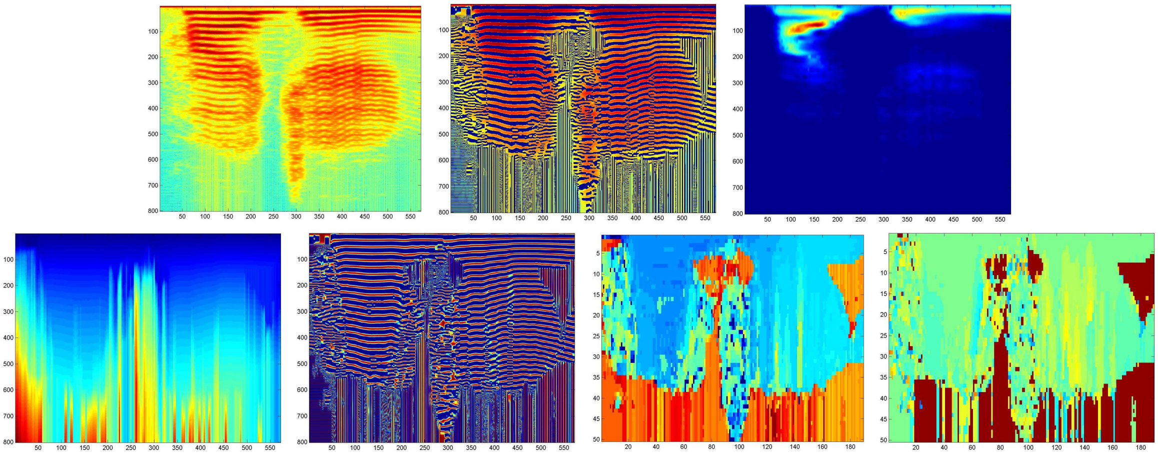

Fig. 3. *Top row, left to right: Original magnitude spectrogram $S(f,t)$, reconstructed spectrogram $\hat{S}(f,t)$, amplitude envelope $A(f,t)$. Bottom row, left to right: smooth phase surface $\phi(f,t)$, rectified harmonic carrier $cos(\phi(f,t))$, Gabor patch frequencies $F(i,j)$, and Gabor patch orientations $\Theta(i,j)$. Our spectrograms are flipped so low frequency is at the top.*

4. AM-FM PATCH OVERLAP-ADD

Given the estimated local amplitude envelopes $A_{ij}(f,t)$ for each patch, we construct the complete envelope $A(f,t)$ for the whole spectrogram using overlap-add:

$$A(f,t) = \frac{\sum_i \sum_j W(f,t) A_{ij}(f,t)}{\sum_i \sum_j W(f,t)} \tag{10}$$

Similarly, we overlap-add the local carriers $cos(\phi_{ij}(f,t))$ for each patch to construct the complete carrier $cos(\phi(f,t))$:

$$cos(\phi(f,t)) = \frac{\sum_i \sum_j W(f,t) cos(\phi_{ij}(f,t))}{\sum_i \sum_j W(f,t)} \tag{11}$$

Obtaining an estimate of the smooth phase surface $\phi(f,t)$ for the entire spectrogram is a bit more involved: We additionally overlap-add local *sine* carriers $sin(\phi_{ij}(f,t))$:

$$sin(\phi(f,t)) = \frac{\sum_i \sum_j W(f,t) sin(\phi_{ij}(f,t))}{\sum_i \sum_j W(f,t)} \tag{12}$$

Then a *principal phase* surface is obtained as:

$$\phi_P(f,t) = atan\left(\frac{sin(\phi(f,t))}{cos(\phi(f,t))}\right) \tag{13}$$

Finally, the desired smooth phase surface $\phi(f,t)$ is obtained by 1-D column-unwrapping the principal phase surface:

$$\phi(f,t) = Unwrap\_1d(\phi_P(f,t)) \tag{14}$$

.

5. RESULTS AND CONCLUSIONS

We analyzed and reconstructed several test utterances of different speakers uttering the phrase `''Hi Jane''`. An example of our results is shown in Figure 3. [1]. The first and second plots in the figure show the real spectrogram $S(f,t)$ and the reconstructed spectrogram $\hat{S}(f,t) = A(f,t)cos(\phi(f,t))$. The following plots depict (consecutively) the estimated amplitude envelope $A(f,t)$, smooth phase surface $\phi(f,t)$, (rectified) harmonic carrier $cos(\phi(f,t))$, Gabor patch frequencies $F(i,j)$, and Gabor patch orientations $\Theta(i,j)$.

In order to perform auditory comparisons, we synthesized time waveforms for both original and reconstructed magnitude spectrograms using sinusoidal analysis/synthesis techniques [7]. Informal listening tests indicated that both were very similar to each other, which suggests that our technique is successful at capturing the important aspects of the spectrogram.

Future work will consist of exploring the use of the extracted parameters for applications such as speech recognition, compression, de-noising, and synthesis.

6. REFERENCES

[1] T. Ezzat, J. Bouvrie, and T. Poggio, "Max-gabor analysis and synthesis of spectrograms," in *Proc. ICSLP*, Pittsburgh, PA, 2006.

[2] P. Maragos, JF Kaiser, and TF Quatieri, "Energy separation in signal modulations with applications to speech analysis," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3024–3051, 1993.

[3] P. Maragos and AC Bovik, "Image demodulation using multidimensional energy separation," *Journal of Optical Society of America*, vol. 12, pp. 1867–1876, 1995.

[4] JP Havlicek, DS Harding, and AC Bovik, "The multicomponent am-fm image representation," *IEEE Trans. on Image Processing*, vol. 5, pp. 1094–1100, 1996.

[5] JG Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional cortical filters," *Journal of Optical Society of America*, vol. 2, pp. 1160–1169, 1985.

[6] T.F. Quatieri, "2-d processing of speech with application to pitch tracking," in *Proc. ICSLP*, September 2001.

[7] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. Vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[1] See `http://cuneus.ai.mit.edu:8000/research/amfm` for more results