Max-Gabor Analysis and Synthesis of Spectrograms

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

10.21437/Interspeech.2006-625

Abstract

We present a method that analyzes a two-dimensional magnitude spectrogram $S(f,t)$ into its local constituent spectro-temporal amplitudes $A(f,t)$, frequencies $F(f,t)$, orientations $\Theta(f,t)$, and phases $\phi(f,t)$. The method operates by performing a two-dimensional local Gabor-like analysis of the spectrogram, retaining only the parameters of the 2D-Gabor filter with *maximal* amplitude response within the local region. We demonstrate the technique over a wide variety of speakers, and show how the spectrograms in each case may be adequately reconstructed using the parameters of the Max-Gabor analysis. Finally, we discuss the nature of the extracted Max-Gabor parameters.

Index Terms: spectrogram analysis, spectrogram reconstruction, two-dimensional Gabor, spectro-temporal frequency, spectro-temporal orientation.

1. Introduction

We observe that within small local two-dimensional patches p in a narrowband magnitude spectrogram $S(f,t)$ *there is usually only one locally dominant spectro-temporal frequency* $F(p)$ *and orientation* $\Theta(p)$. Shown in Figure 1 are several patches A, B, and E which exhibit this local spectro-temporal behavior for a speaker uttering ``Hi Jane''. Secondly, we observe that these locally dominant spectro-temporal frequencies and orientations *change smoothly in time and frequency*. Finally, we observe that there are patches for which this assumption is violated, such as patch F in Figure 1.

Based on these observations, our goal in this work is to present a method which analyzes a magnitude spectrogram $S(f,t)$ into its locally dominant spectro-temporal frequencies $F(f,t)$ and orientations $\Theta(f,t)$. Our method also estimates local patch amplitudes $A(f,t)$ and phases $\Phi(f,t)$ as well. Finally, our method adequately reconstructs the analyzed spectrograms from the extracted parameters. Since the local patches are Gabor-like, the method we performs a two-dimensional Gabor-like analysis of the spectrogram, retaining only the parameters of the 2D-Gabor filter with *maximal* amplitude response within the local region. Hence we call our technique a *Max-Gabor* analysis of spectrograms.

Our hope in performing such an analysis is to factor out various important phenomena occurring during speech. In particular, local Gabor amplitude should relate to formant energies during speech, while local spectro-temporal frequency and orientation should relate to pitch and the underlying pitch changes during speech. At the same time, we hope that the parameters extracted by this representation summarize the spectrogram in a smooth and useful way, and would be useful for further processing in applications such as speech recognition and speech synthesis.

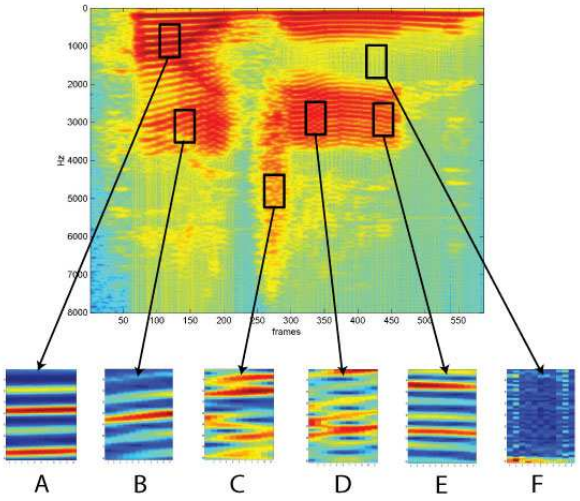

Figure 1: *A magnitude spectrogram of ``Hi Jane'' (vertical axis is low frequency to high frequency). Patches A,B,E exhibit clear locally dominant spectro-temporal frequencies and orientations. Patch F violates this assumption, while patches C,D are somewhere in between.*

2. Background

Our Max-Gabor analysis is inspired by the work of Shamma and colleagues [1] [2], who have developed a two-stage auditory model based on psycho-acoustical and neurophysiological findings in the early and central stages of the auditory pathway.

The first stage of their approach converts a sound into an *auditory spectrogram*, which is similar in nature to a regular magnitude spectrogram. The second stage analyzes local patches of the auditory spectrogram using a filterbank of two-dimensional local spectro-temporal filters that are selective to different frequency scales Ω and to different temporal rates ω. Thus, for each point (i,j) in the spectrogram, the Shamma model produces a two-dimensional output $R_{ij}(\Omega,\omega)$ that is the response of the entire filterbank for that location. Typical values for the dimensionality of Ω and ω are 6 and 26 respectively [2].

In our work we also examine the content of local spectro-

temporal patches of the spectrogram using an analogous response function $R_{ij}(\Omega, \omega)$. However, unlike the Shamma model, *we choose to keep only the maximum output response value* $\max_{\Omega,\omega}(R_{ij}(\Omega, \omega))$*, as well as the parameters of the filter which produced the maximum value:* $argmax_{\Omega,\omega}(R_{ij}(\Omega, \omega))$. In so doing, we are using our assumption that, within a local patch, there is only one *dominant* local frequency and orientation, and consequently the entire filterbank $R_{ij}(\Omega, \omega)$ may be *compressed* to the parameters of one meaningful Gabor filter.

The usage of a MAX operator as the mechanism to select the dominant local frequency and orientation is itself motivated by recent work in visual neuroscience [3], where it was embedded in a hierarchical mechanism for visual object recognition in order to account for visual translation- and scale-invariance. Additionally, a MAX operator was also used in texture analysis [4], where it was used to extract dominant local texture parameters for the purpose of texture segmentation.

3. Max-Gabor Analysis

3.1. Overview

Our MAX-Gabor algorithm is modelled after recent fingerprint enhancement algorithms [5], where it is also necessary to analyze local spectro-temporal frequencies and orientations. Shown in Figure 2 is a schematic of the various stages of our algorithm, which we discuss in detail below.

3.2. 2D Gabor Definition

We define a 2D Gabor $G(f, t)$ with spectro-temporal frequency F, spectro-temporal orientation Θ, amplitude A, and phase Φ as:

$$G(f, t) = A \cdot W(f, t) \cdot cos(2\pi F\hat{x} + \Phi) \quad (1)$$

where

$$\hat{x} = tcos\Theta + fsin\Theta \quad (2)$$

and $W(f, t)$ is a Gaussian-like window (to be defined later). For phase estimation, we will need the notion of a complex Gabor, which will consist of real and imaginary versions of equation 1 in quadrature phase:

$$G^*(f, t) = A \cdot W(f, t) \cdot e^{j(2\pi F\hat{x}+\Phi)} \quad (3)$$

Finally, since magnitude spectrograms are nonnegative, we will also need to rectify Gabors during reconstruction, so we will denote a rectified real Gabor as $\hat{G}(f, t)$.

3.3. 1D STFT

All of the 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.

3.4. 2D Local FFT

At every grid point (i, j), we extract a patch $P_{ij}(f, t)$ of the spectrogram of size df and width dt. First, the patch is multiplied by a 2D Hamming window $W_H(f, t)$ of the same size as the patch.

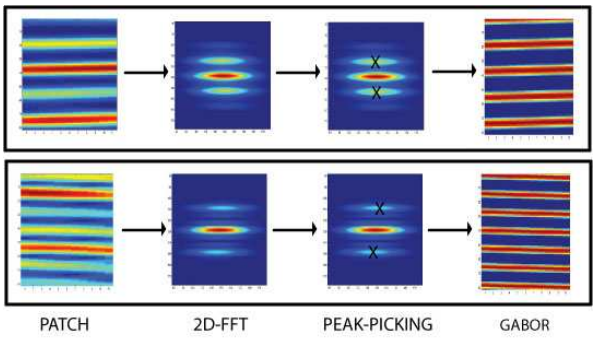

Figure 2: *Two example Max-Gabor analyses: First column depicts sample input patches* $P_{ij}(f, t)$*. Second column depicts sample spectro-temporal spectra* $R_{ij}(\Omega, \omega)$*. Third column depicts estimated Max-Gabor peaks* $\{\Omega_{max}, \omega_{max}\}$*. Fourth column depicts reconstructed Gabors* $G_{ij}(f, t)$*.*

Second, a 2-dimensional Fourier transform of size $N_H \times N_W$ is performed on the patch to produce the local spectral-temporal magnitude spectrum:

$$R_{ij}(\Omega, \omega) = \left\| \sum_f \sum_t W_H(f, t) P_{ij}(f, t) e^{-j2\pi \frac{\Omega}{N_H} f} e^{-j2\pi \frac{\omega}{N_W} t} \right\| \quad (4)$$

It is important to point out that 2D-FFT weighted by $W_H(f, t)$ takes the place of Shamma's 2D Gabor analysis in our case.

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 frequency and orientation, but small enough so that the underlying signal is locally stationary. Suitable parameter ranges are 10-20msec for the dt parameter, and $800Hz - 1.2Khz$ for the df parameter. Male speakers require window heights in the lower end of the df range, while female speakers require a window height in the higher end of the df range.

Additional analysis parameters are the window hopsizes in time Δi and frequency Δj, as well as the FFT sizes N_H and N_W. Typically we set Δi to be 3-5ms and Δj to 50-100Hz, which creates overlap between the patches. N_H and N_W are each set to 256.

3.5. Peak-Finding With Quadratic Interpolation

Visual inspection of the local spectro-temporal magnitude spectrum $R_{ij}(\Omega, \omega)$ for different window patches reveals that most of the spectra exhibit a Gabor-like spectral structure (see Figure 2). This is exemplified by the presence of two Gaussian-like "bumps" in the spectrum whose location we wish to identify. Additionally, a DC "bump" usually exists because the magnitude STFT is nonnegative, so local patches will have a nonzero average value.

We use a *peak-finding* strategy to obtain a set C of candidate locations for the Gabor "bumps" in the spectral response $R_{ij}(\Omega, \omega)$. This set C is composed of the locations $\{\Omega_c, \omega_c\}$ of the local peaks, as well as their corresponding values $\{R_c\}$. The local peaks and their locations are first identified on the original sampling grid of $R_{ij}(\Omega, \omega)$, and then subsequently refined using local quadratic interpolation.

Generally, the peak locations will come in conjugate pairs due to the conjugate symmetry of the Fourier transform, so we match the conjugate peak locations in C with each other into pairs. We also remove the DC peak from the set C if it exists.

As a result, at the end of our peak-processing stage for each patch, the set C will contain a set of candidate local maxima $\{R_c\}$, and their conjugate locations $\{\Omega_c, \omega_c\}$.

3.6. Locally Dominant Frequency and Orientation Estimation Using MAX

We determine the locally dominant Gabor peak by choosing among the peaks in the set C using a MAX operator:

$$R_{max} = \max_c \{R_c\} \tag{5}$$

This identifies the corresponding peak locations as

$$\{\Omega_{max}, \omega_{max}\} = \{\Omega_{c^*}, \omega_{c^*}\} \tag{6}$$

where

$$c^* = argmax_c \{R_c\} \tag{7}$$

Finally, the locally dominant orientation and frequency may be estimated from the chosen peak location as

$$\Theta(i,j) = tan^{-1}\left(\frac{\Delta\Omega_{max}}{\Delta\omega_{max}}\right) \tag{8}$$

and

$$F(i,j) = \frac{\sqrt{\left(\frac{\Delta\Omega_{max}}{N_H}\right)^2 + \left(\frac{\Delta\omega_{max}}{N_W}\right)^2}}{2} \tag{9}$$

where $\Delta\Omega_{max}$ and $\Delta\omega_{max}$ refers to differences between the conjugate pair location coordinates. Shown in Figure 2 in the third column are example peaks extracted by our algorithm.

3.7. Local Phase and Amplitude Estimation

Local phase $\Phi(i,j)$ is estimated for the patch under consideration by first synthesizing a *complex* 2D Gabor signal $G^*_{ij}(f,t)$ with local frequency $F(i,j)$, local orientation $\Theta(i,j)$, amplitude $A = 1$, phase $\Phi = 0$, and window $W_H(f,t)$. The patch $P(i,j)$ is then projected onto the complex Gabor $G^*_{ij}(f,t)$, and the phase value determined from the resulting angle:

$$\Phi(i,j) = angle\left(\sum_f \sum_t W_H(f,t) P_{ij}(f,t) G^*_{ij}(f,t)\right) \tag{10}$$

Similarly, local amplitude is estimated for the patch under consideration by first synthesizing a rectified real 2D Gabor signal $\hat{G}_{ij}(f,t)$ with local frequency $F(i,j)$, local orientation $\Theta(i,j)$, phase $\Phi(i,j)$, and amplitude $A = 1$. An optimal scaling factor $A(i,j)$ which scales the synthetic Gabor to to match the current patch P_{ij} under consideration is estimated as:

$$A(i,j) = \frac{\sum_f \sum_t W_A(f,t) P_{ij}(f,t) \hat{G}_{ij}(f,t)}{\sum_f \sum_t W_A^2(f,t) P_{ij}^2(f,t)} \tag{11}$$

For accurate amplitude estimation, we have found it necessary to use an amplitude Hanning window $W_A(f,t)$ which is *narrower* than the patch Hamming window $W_H(f,t)$. Typically, $W_A(f,t)$ ranges in height from 200Hz-450Hz, and 1-5msec in width.

3.8. Smoothing over Time and Frequency

There are many cases (such as patch F in Figure 1) when a patch has very little or no energy, and our peak-picking algorithm does not find any local peaks in the spectro-temporal response $R_{ij}(\Omega, \omega)$. We handle this case by setting the frequency and orientation for the current patch $P(i,j)$ to be the same as those from the previous patch *in time* $P(i-1,j)$.

There are also cases when a patch (such as patch C in Figure 1) contains a lot of noise-like energy, and the peak-finding algorithm finds many spurious peaks which may throw off the local frequency and orientation estimation in Section 3.6. We handle this case by removing from the set C those candidate peaks whose orientations and frequencies are significantly different from those estimated in the previous patches $P(i-1,j)$ and $P(i,j-1)$ in time and frequency. Two parameters, ΔF_{max} and $\Delta\Theta_{max}$, determine how much patch-to-patch change in frequency and orientation we are willing to tolerate in our Max-Gabor analysis.

4. Max-Gabor Synthesis

Given the estimated local frequencies $F(i,j)$, orientations $\Theta(i,j)$, amplitudes $A(i,j)$, and phases $\Phi(i,j)$, the spectrogram $\hat{S}(f,t)$ is reconstructed by synthesizing local rectified 2D Gabors $\hat{G}_{ij}(f,t)$ with window W_A for each patch, and overlap-adding them together:

$$\hat{S}(f,t) = \frac{\sum_i \sum_j \hat{G}_{ij}(f,t)}{\sum_i \sum_j W_A(f,t)} \tag{12}$$

5. Max-Gabor Results

We analyzed and re-synthesized several test utterances of different speakers uttering the phrase ``Hi Jane''. Two examples of our results are shown in Figures 3 and 4 [1].

The first through third plots in each pair of Figures shows the real spectrogram $S(f,t)$, the reconstructed spectrogram $\hat{S}(f,t)$, and the reconstructed spectrogram $\hat{S}(f,t)$ smoothed with a 150Hz-by-3msec 2D Gaussian filter (to scale the colormap for better comparison with the first plot). Comparing the reconstructions with the original spectrograms reveals that our Max-Gabor analysis faithfully captures the local frequency, orientation, and amplitude of the original harmonics.

The fourth through seventh plots in each figure depict the local amplitudes $A(i,j)$, frequencies $F(i,j)$, orientations $\Theta(i,j)$, and phases $\Phi(i,j)$ estimated by our Max-Gabor analysis. The local amplitudes clearly capture the formant energy behavior in an utterance. The local frequency plots exhibit visible segmentation into distinct but smooth pitch regions (for example, as yellow and blue regions in the 5th plot of Figure 3) Finally, the orientation plots depict distinct vertical orientation *bands* which represent underlying upward or downward pitch shifts. Dark blue bands reflect downward pitch shifts, while yellow and red bands depict upward pitch shifts.

Finally, in order to perform auditory comparisons, we synthesized time waveforms for both original and reconstructed magnitude spectrograms using sinusoidal analysis/synthesis techniques [6]. Informal listening tests indicated that both were very simi-

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

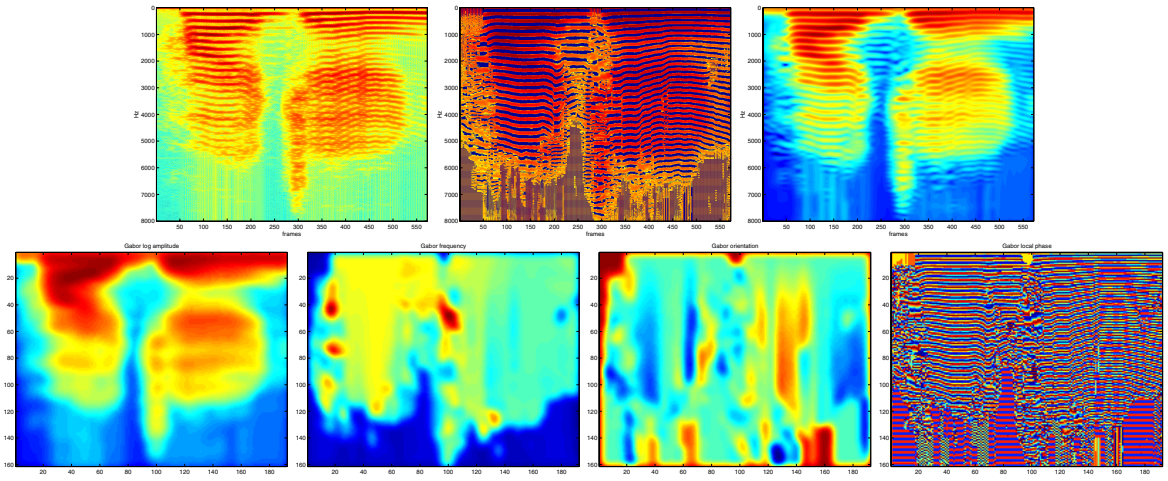

Figure 3: *Top row, left to right: Original magnitude spectrogram $S(f,t)$, reconstructed spectrogram $\hat{S}(f,t)$, reconstructed spectrogram convolved with a small Gaussian filter $\hat{S}(f,t) * W_G(f,t)$. Bottom row, left to right: $A(i,j)$, $F(i,j)$, $\Theta(i,j)$, and $\Phi(i,j)$.*

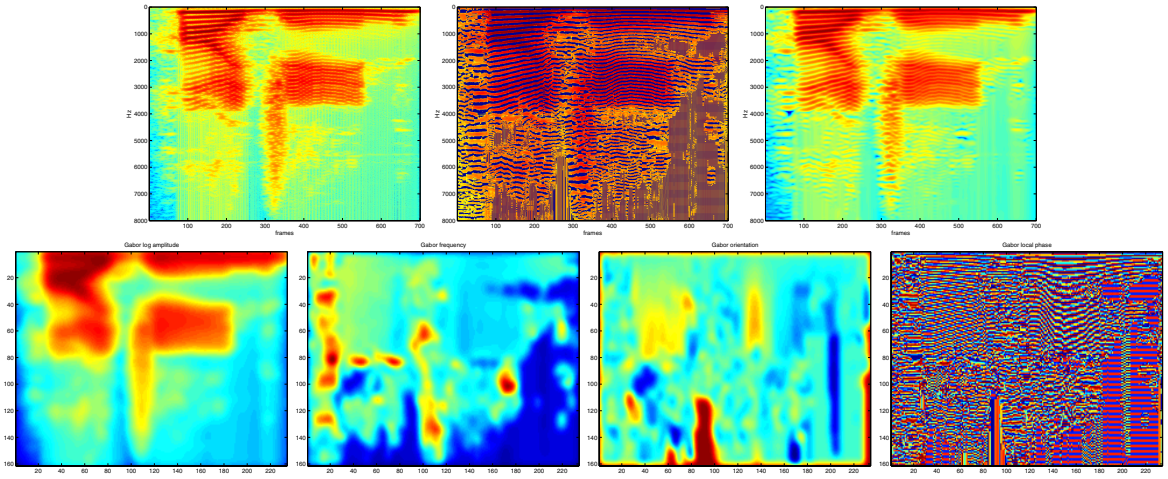

Figure 4: *Top row, left to right: Original magnitude spectrogram $S(f,t)$, reconstructed spectrogram $\hat{S}(f,t)$, reconstructed spectrogram convolved with a small Gaussian filter $\hat{S}(f,t) * W_G(f,t)$. Bottom row, left to right: $A(i,j)$, $F(i,j)$, $\Theta(i,j)$, and $\Phi(i,j)$.*

lar to each other, which suggests that the Max-Gabor technique is successful at capturing the important aspects of the spectrogram.

6. Conclusions and Future Work

We presented a method that analyzes a two-dimensional magnitude spectrogram $S(f,t)$ into its locally dominant spectro-temporal amplitudes $A(f,t)$, frequencies $F(f,t)$, orientations $\Theta(f,t)$, and phases $\phi(f,t)$. In addition, we presented a method that reconstructs a spectrogram from the extracted parameters.

The quality of our reconstructions suggests that assuming only one dominant frequency and orientation within a local patch is in fact a valid assumption, and represents a meaningful compression of the filterbank outputs of [1].

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

7. References

[1] T. Chih, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[2] N. Mesgarani, M. Slaney, and S. Shamma, "Speech discrimination based on multiscale spectro-temporal features," in *Proc. ICASSP*, May 2004.

[3] M. Riesenhuber and T. Poggio, "Hierachical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[4] AC Bovik, M Clark, and WS Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55–73, 1990.

[5] S. Chikkerur, A. Cartwright, and V. Govindaraju, "Fingerprint image enhancement using stft analysis," in *Proc. ICAPR*, 2005.

[6] 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.