Spectro-Temporal Analysis of Speech Using 2-D Gabor Filters

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 2-D spectro-temporal Gabor filterbank based on the 2-D Fast Fourier Transform, and show how it may be used to analyze localized patches of a spectrogram. We argue that the 2-D Gabor filterbank has the capacity to decompose a patch into its underlying dominant spectro-temporal components, and we illustrate the response of our filterbank to different speech phenomena such as harmonicity, formants, vertical onsets/offsets, noise, and overlapping simultaneous speakers.

Index Terms: speech analysis, spectro-temporal filterbanks, 2-D Gabor

1. Introduction

A typical narrowband magnitude spectrogram of speech displays several important and well-known phenomena: harmonicity, which is exemplified by the presence of horizontal lines related to the pitch of the speaker; low-frequency amplitude modulations, related to the formants of the speaker's vocal tract; vertical onset/offset edges in time, related to plosive sounds in speech; and noise, related to fricatives, aspirants, and other phonemes which generate noise.

All of these speech-related phenomena may be viewed as different types of spectro-temporal modulations, and one of the challenges our auditory system faces in processing speech is that it must detect, separate, and recognize these modulations in a fast and reliable manner.

Recent neurophysiological evidence from a number of animals [1] [2] indicates that cells in the auditory cortex are, in fact, tuned to localized spectro-temporal modulations. The spectro-temporal receptive fields (STRFs) of these cortical cells look like 2-D spectro-temporal Gabor filters, and an analogy between these STRFs and a 2-D spectro-temporal Gabor filterbank clearly suggests itself.

Motivated by these studies, we present in this work a simple 2-D Gabor filterbank, and use it to analyze localized patches of a spectrogram. We show in this work how such a filterbank responds to the different types of phenomena commonly occuring in spectrograms, and further argue that, in principle, it has the capacity to detect, separate, and recognize these phenomena.

Our 2-D Gabor filterbank is implemented using the 2-D FFT. We do this for two main reasons: firstly, the 2-D FFT is fast, since it makes use of the Fast Fourier Transform. Additionally, the 2-D FFT organizes its outputs into a 2-D grid, which allows us to easily visualize the filterbank's response as an image.

In prior work, Kleinschmidt, Gelbart, and colleagues [3] [4] also applied 2-D spectro-temporal Gabors to mel-spectrograms. However, in their approach, no organizational map of the Gabor filter responses was formed. Instead, the 2-D Gabor outputs were lumped together into a one-dimensional feature vector for use in recognition experiments. As a consequence, it is hard to interpret their results and see how the 2-D Gabor filterbank analyzed the various types of spectrogram phenomena.

Shamma and colleagues [1] [5] have also applied 2-D spectro-temporal Gabor filterbanks for speech discrimination and enhancement. In their work the spectro-temporal responses were organized into a very large multi-dimensional tensorial representation which is very hard to visualize and interpret. We present an alternative filterbank decomposition in our work here which we believe is simpler and easier to interpret.

Finally, in our own previous work [6], we applied 2-D Gabor analysis using the 2-D FFT to spectrogram patches. However, in that work we only noted the filterbank's response to harmonic phenomena, and failed to document how it responds to other very important phenomena such as formants and vertical edges. This was partly due to the fact that patch DC values were not removed prior to performing 2-D Gabor analysis, which made it difficult to see the response of the filterbank for these other phenomena.

In the following sections, we briefly review how our spectrograms are constructed (Section 2), and how are patches are selected (Section 3). Then we describe our 2-D Gabor filterbank in Sections 4 and 5. Finally, in Section 6 we describe the response of the filterbank for different types of speech-related phenomena.

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.

3. STFT Patches

At every grid point (i,j) in the spectrogram, 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 spectro-temporal components in the patch, but small enough so that the underlying signal is stationary. Suitable parameter ranges are 5-15msec for the dt parameter, and $600-800Hz$ 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. Additionally, we subtract the patch DC value $\frac{1}{df\,dt}\sum_{f,t} P_{ij}(f,t)$ from the patch $P_{ij}(f,t)$

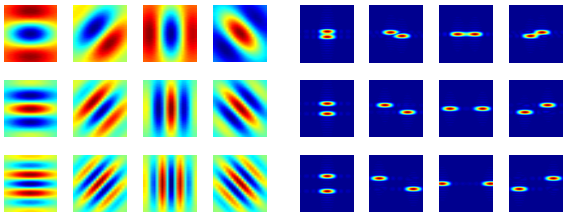

Figure 1: Left: Examples of spectro-temporal Gabor filters $G_{F,\Theta}(f,t)$ for $F = \{1,2,3\}$ and $\Theta = \{0, \frac{pi}{4}, \frac{pi}{2}, \frac{3pi}{4}\}$. Right: The magnitude Fourier transform of each corresponding Gabor filter on the left.

before any further processing.

4. 2-D Gabor Filterbank Using the 2-D FFT

To perform 2-D Gabor filterbank analysis, we first multiply each patch $P_{ij}(f,t)$ by a 2-D Gaussian window $W(f,t)$ located at the patch center (f_0, t_0):

$$W(f,t) = \frac{1}{2\pi\sigma_f\sigma_t} e^{-\frac{1}{2}\left(\frac{(f-f_0)^2}{\sigma_f^2} + \frac{(t-t_0)^2}{\sigma_t^2}\right)} \quad (1)$$

Typically, we fix the window bandwidth (σ_f, σ_t) to be one-third of the patch height and width respectively.

Secondly, a 2-D Fourier transform of size $N_H \times N_W$ is applied to the windowed patch to produce the 2-D spectro-temporal Gabor filterbank output:

$$R_{ij}(\Omega,\omega) = \sum_f \sum_t W(f,t) P_{ij}(f,t) e^{-j2\pi\frac{\Omega}{N_H}f} e^{-j2\pi\frac{\omega}{N_W}t}$$

Typical values N_H and N_W are 512 and 256 respectively.

By exchanging the terms $W(f,t)$ and $P_{ij}(f,t)$, we can re-write the filterbank output as

$$R_{ij}(\Omega,\omega) = \sum_f \sum_t P_{ij}(f,t) G^*_{\Omega,\omega}(f,t) \quad (2)$$

where

$$G_{\Omega,\omega}(f,t) = W(f,t) e^{j2\pi(\frac{\Omega}{N_H}f + \frac{\omega}{N_W}t)} \quad (3)$$

Equation 3 above is the equation of a 2-D spectro-temporal Gabor filter. $R_{ij}(\Omega,\omega)$ may thus be viewed as the projection of a patch $P_{ij}(f,t)$ on an entire bank of spectro-temporal 2-D Gabor filters $G_{\Omega,\omega}(f,t)$.

It is sometimes desirable to re-parameterize the spectro-temporal Gabors $G_{\Omega,\omega}(f,t)$ in terms of their spectro-temporal frequency F and orientation Θ. This can be done in a straight-forward manner through the forward trigonometric mapping $(F,\Theta) = (\sqrt{\Omega^2+\omega^2}, tan^{-1}(\Omega,\omega))$ and the backward trigonometric mapping $(\Omega,\omega) = (Fcos\Theta, Fsin\Theta)$. Shown in Figure 1 on the left are example 2-D Gabors $G_{F,\Theta}(f,t)$ for different values of F and Θ.

Finally, it is well-known [7] 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. Shown in Figure 1 on the right are example 2-D Fourier transforms of the Gabor filters on the left.

5. Patch 2-D Gabor Analysis & Synthesis

Shown in Figure 2 on the left is a representative patch from a narrowband magnitude spectrogram. The magnitude of the 2-D spectro-temporal response $|R_{ij}(\Omega,\omega)|$ for that patch is plotted in the middle. In general, 2-D spectro-temporal Gabor responses of spectrogram patches exhibit multiple Gaussian "peaks" that come in conjugate pairs. Each peak pair corresponds to a different spectro-temporal modulation contained in the patch. In effect, the 2-D Gabor filterbank decomposes a patch into its spectro-temporal components. "Peaks" with large amplitude in the spectro-temporal response $R_{ij}(\Omega,\omega)$ correspond to dominant spectro-temporal modulations in the patch.

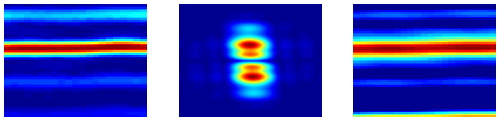

Figure 2: Left: example spectrogram patch $P_{ij}(f,t)$. Middle: magnitude spectro-temporal response $|R_{ij}(\Omega,\omega)|$. Right: reconstructed patch $\hat{P}_{ij}(f,t)$ using the top 5 spectro-temporal components.

In order to examine what spectro-temporal component each of these peak pairs corresponds to, a simple peak-detection strategy is used to obtain a set C of candidate peak locations $(\Omega_{peak}, \omega_{peak})$ and values (amplitude A_{peak} and phase Φ_{peak}) from the spectro-temporal 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.

For each of these matched peak pairs, we can estimate a spectro-temporal orientation Θ_k and frequency F_k associated with that spectro-temporal component k as

$$\Theta_k = tan^{-1}\left(\frac{\Delta\Omega_{peak}}{\Delta\omega_{peak}}\right) \quad (4)$$

and

$$F_k = \frac{\sqrt{\left(\frac{\Delta\Omega_{peak}}{N_H}\right)^2 + \left(\frac{\Delta\omega_{peak}}{N_W}\right)^2}}{2} \quad (5)$$

where $\Delta\Omega_{peak}$ and $\Delta\omega_{peak}$ refers to differences between the conjugate pair location coordinates. The amplitude A_k and phase Φ_k of the spectro-temporal component are just equal to the amplitude A_{peak} and phase Φ_{peak} of the peaks in $R_{ij}(\Omega,\omega)$ itself.

The spectro-temporal component k associated with a certain peak pair may thus be synthesized as $A_k e^{-j\Phi_k} G_{F_k,\Theta_k}(f,t)$. If there are k components in a patch then the patch may be approximately reconstructed as

$$\hat{P}_{ij}(f,t) = \Re\left(\sum_k A_k e^{-j\Phi_k} G_{F_k,\Theta_k}(f,t)\right) \quad (6)$$

Shown in Figure 2 at right is the reconstuction of the patch on the left using the top 5 spectro-temporal components from the 2-D spectro-temporal Gabor response in the middle.

6. Phenomenological Analysis of 2-D Gabor Responses

We now examine more carefully how different types of commonly-occurring phenomena in spectrograms are analyzed by the 2-D Gabor transform. The phenomena we will examine are harmonicity; low-frequency amplitude modulations related to formants; vertical onset/offset phenomena related to plosives; phonetic noise; the effect of adding white background noise; and finally, the effect of overlapping simultaneous speakers. In each of the following sections, we look at each phenomenon individually.

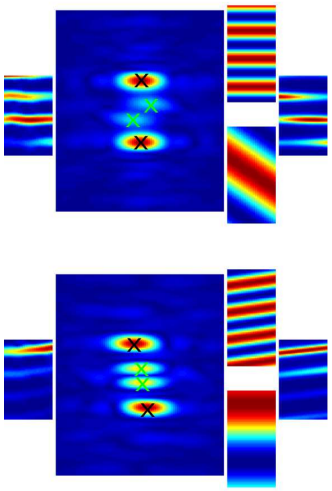

Figure 3: First column: example harmonic patches $P_{ij}(f,t)$. Second column: magnitude spectro-temporal responses $|R_{ij}(\Omega,\omega)|$ for each patch. Third column: spectro-temporal components corresponding to largest (black "X") and next largest (green "X") peak pairs. Fourth column: reconstructed patches $\hat{P}_{ij}(f,t)$ using the top 2 spectro-temporal components.

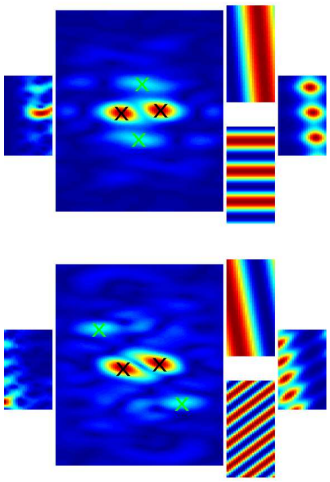

Figure 4: First column: example plosive patches $P_{ij}(f,t)$. Second column: magnitude spectro-temporal responses $|R_{ij}(\Omega,\omega)|$ for each patch. Third column: spectro-temporal components corresponding to largest (black "X") and next largest (green "X") peak pairs. Fourth column: reconstructed patches $\hat{P}_{ij}(f,t)$ using the top 2 spectro-temporal components.

6.1. Harmonicity

Shown in Figure 3 on the far left are two different representative harmonic patches from a spectrogram. In the second column, we plot the magnitude of the spectro-temporal response $|R_{ij}(\Omega,\omega)|$ for each patch. In the third column we plot the re-synthesized spectro-temporal components associated with the top two peak pairs in the spectro-temporal response. The first spectro-temporal component corresponds to peak pair with the largest amplitude (marked with a black "X"), while the second spectro-temporal component corresponds to peak pair with the second largest amplitude (marked with a green "X"). Finally, in the last column we plot the reconstruction $\hat{P}_{ij}(f,t)$ of the input patch using only the top two components (as in Equation 6).

Clearly we can see that the spectro-temporal components associated with the largest peaks are in fact the harmonic components. In general, harmonicity in a patch emerges as a set of dominant conjugate peaks spaced with a distance proportional to spectro-temporal frequency F and with an angle proportional to the spectro-temporal orientation Θ. This fact was noticed and employed for the purposes of carrier estimation in [6] and pitch-tracking in [8].

6.2. Formants

Further inspection of the spectro-temporal magnitude responses $|R_{ij}(\Omega,\omega)|$ in Figure 3 reveals that the patches contain a second component exemplified by the presence of two smaller peaks located closer to the origin in the spectro-temporal responses. Synthesizing the spectro-temporal component associated with these secondary peaks reveals that they correspond to the low-frequency amplitude modulations associated with formants.

In general, low-frequency amplitude modulations in a patch emerge as a set of conjugate peaks that are spaced closer to the origin in the spectro-temporal response. As with the harmonic peaks, the distance and angle of the peaks corresponds directly to the frequency and orientation of the modulation.

It is worthwhile to note here that 2-D Gabor filterbank analysis is capable of *separating* harmonic from formant spectro-temporal components. Such a property could be used by our auditory system to endow us with the ability to recognize speech in a manner that is invariant to the speaker uttering it: our auditory system could have evolved to focus its attention mainly on the low-frequency components (which carry the phonetic information), ignoring the higher-frequency harmonic components related to the speaker.

On the other hand, the response to both harmonic and formant components is simultaneously *preseved* in the Gabor filterbank's outputs. Such a property could be used by our auditory system to enable us to recognize speech in a noise-robust manner: whenever distinct harmonic peaks emerge in the filterbank's response, the auditory system can identify that response as a distinct signature of speech, and it can then attend to the information contained in the low-frequency peaks related to formants.

6.3. Vertical/Plosive Edges

Shown in Figure 4 on the far left are two different representative patches which contain plosive phenomena. These phenomena are characterized by rapid onset/offset of voicing or noise, and look like vertical edges in a patch.

Inspection of the spectro-temporal magnitude responses $|R_{ij}(\Omega,\omega)|$ for these plosive patches reveals the presence of two dominant peaks which are *horizontal* in their angular orientation, in contrast to the *vertical* angular orientation of both harmonic and formant spectro-temporal modulations shown in Figure 3. Synthesizing the spectro-temporal component associated with these peaks reveals that they do in fact correspond to a vertical amplitude modulation associated with the plosive edge.

We note that vertical plosive onset/offset edges *cannot* be detected unless a filterbank is used whose filters have a *finite extent in time*. Consequently, it is quite heartening that, within the same 2-D Gabor filterbank framework, all three types of harmonic, formant, and plosive phenomena can be separately detected.

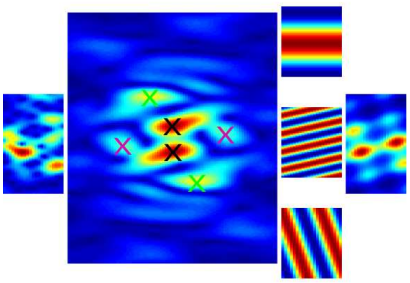

Figure 5: First column: example noisy patch $P_{ij}(f,t)$. Second column: magnitude spectro-temporal responses $|R_{ij}(\Omega,\omega)|$. Third column: spectro-temporal components corresponding to largest (black "X"), second largest (green "X"), and third largest (magenta "X") peak pairs. Fourth column: reconstructed patch $\hat{P}_{ij}(f,t)$ using the top 3 spectro-temporal components.

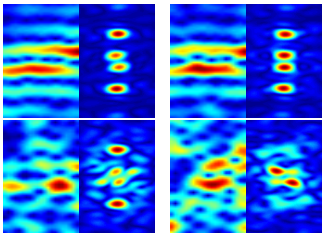

Figure 6: Patch $P_{ij}(f,t)$ with increasing levels of white noise added, along with the corresponding magnitude spectro-temporal response $|R_{ij}(\Omega,\omega)|$ for each patch. (Top left: 10 dB SNR, Top right: 5 dB SNR, Bottom left: -5 dB SNR, Bottom right: -10 dB SNR).

6.4. Phonetic Noise

Shown in Figure 5 on the left is a representative noisy patch corresponding to the phoneme sh. As may be seen, the spectro-temporal response of a noisy patch contains multiple peaks at multiple orientations and frequencies. In general we have found that this is a signature propery of noisy patches, in contrast to harmonic patches or plosive patches which usually contain one or at most two dominant spectro-temporal components. For example, we need at least three spectro-temporal components in order to begin to faithfully reconstruct the patch in Figure 5.

6.5. Background Noise

In Figure 6 we investigate the effect of adding white noise to a sound. Shown on the left of each figure pair is the same identical patch but with increasing amounts of white noise added[1]. On the right hand side of each figure pair are the corresponding spectro-temporal responses. As may be seen, even though the SNR ratios are quite low, the spectro-temporal responses show very clear harmonic peaks up to about -5 SNR ratios. At -10 SNR, the spectro-temporal response begins to lose the dominant harmonic peaks.

Even though white noise adds a large number of uncorrelated peaks to the spectro-temporal response, peaks corresponding to the relevant harmonic, formant, or plosive phenomena may still be detectable. This is because the output of any one spectro-temporal Gabor filter is obtained by *integrating information from the entire 2-D patch*, which allows the 2-D spectro-temporal filterbank to be more robust to noise than purely 1-D spectral counterparts.

[1] White noise added in the time domain before the STFT is recomputed.

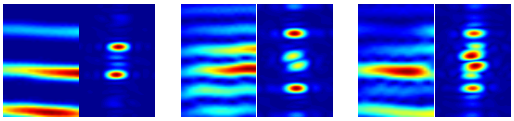

Figure 7: Patches $P_{ij}(f,t)$ and associated magnitude spectro-temporal response $|R_{ij}(\Omega,\omega)|$ for speaker A (left), speaker B (middle), and speaker A+B (right).

6.6. Overlapping Speakers

Finally, we examine the effect of simultaneous speakers on the spectro-temporal response. Shown in the left of Figure 7 is a patch from speaker A and its associated spectro-temporal response. In the middle is the patch and response from speaker B (with same spectrogram coordinates (i,j) as the patch from speaker A). Finally, in the right of the figure is the patch and response from a spectrogram of speaker A and B. As may be seen, the spectro-temporal response of the simultaneous speakers contains identifiable peaks from both speakers. *A strategy suggests itself for speaker separation which is based on identifying the peaks in the combined spectro-temporal response, and assigning those peaks across frequency and time to either speaker.*

7. Discussion and Conclusion

In this work, we showed that the 2-D spectro-temporal Gabor response $R_{ij}(f,t)$ contains many useful and important properties. In particular, we showed that 1) harmonicity emerges as a pair of very dominant vertical peaks; 2) formants emerge as a pair of vertical peaks spaced closer to the origin; 3) plosive onsets/offsets emerge as a pair of horizontal peaks; and 4) noise emerges as a large number of peaks at multiple orietations and frequencies.

While this paper merely presented initial observations, future work will consist of more systematic exploration of the use of 2-D Gabor spectro-temporal responses for applications such as speech recognition, de-noising, separation, and synthesis.

8. 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] F.E. Theunissen, K. Sen, and A. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *Journal of Neuroscience*, vol. 20, pp. 2315–2331, 2000.

[3] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. ICSLP*, 2002.

[4] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. Eurospeech*, 2003.

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

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

[7] 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.

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