Using the Gammachirp Filter for Auditory Analysis of Speech

18.327: Wavelets and Filterbanks

Alex Park
malex@sls.lcs.mit.edu

May 14, 2003

Abstract

Modern automatic speech recognition (ASR) systems typically use a bank of linear filters as the first step in performing frequency analysis of speech. On the other hand, the cochlea, which is responsible for frequency analysis in the human auditory system, is known to have a compressive non-linear frequency response which depends on input stimulus level. Irino and Patterson have developed a theoretically optimal auditory filter, the gammachirp, whose parameters can be chosen to fit observed physiological and psychophysical data. The gammachirp impulse response can be used as the kernel for a wavelet transform which approximates the frequency response of the cochlea. This paper implements the filter design described by Irino and examines its application to a specific example of speech. Implications for noise robust speech analysis are also discussed.

1 Introduction

Speech is a natural and flexible mode of communication for humans. For transmission of information, speech is very efficient; conversational speaking rates can be as high as 200 words per minute. For reception of information, speech offers advantages as well. The auditory system allows us to perceive and understand speech omnidirectionally over a wide variety of background noise conditions, including situations where multiple speakers may be talking.

Because of the important role of speech in human-human interaction, automatic speech recognition (ASR) and understanding is considered a critical component of systems which seek to enable flexible and natural user interaction. Over the past 30 years, advancements in speech recognition technology have led to the adoption of ASR in large vocabulary applications such as dictation software, as well as in limited domain tasks such as voice control of non-critical automobile functions. Despite its deployment in specialized applications, automatic speech recognition is typically not viewed as a mature and reliable technology.

One of the characteristic weaknesses of ASR systems, and a reason they are not more widely used, is their lack of robustness to noise. In [1], Lippmann compared the recognition performance of ASR systems with that of humans and found that humans outperform automatic systems significantly on clean, noisefree data. At higher noise levels, or under mismatched training and testing conditions, the performance gap is much higher.

A contributing factor to the lack of robustness may be in the front-end processing used by ASR systems to analyse incoming sounds. This paper is motivated by the hypothesis that the poor robustness of ASR systems is partly due to inadequate modeling of the human auditory periphery. Specifically, the absence of a compressive cochlear non-linear component, which is common to automatic systems and some hearing impaired humans, may explain similar conditions experienced by both in noisy environments.

The purpose of this paper is twofold. First, we review the work of Irino and Patterson in developing the gammachirp auditory filter as a possible filtering model for the cochlea. We compare this new technique with traditional approaches to speech analysis and with a simpler auditory model from a wavelet-filterbank perspective. Second, we propose a framework for incorporating the compressive non-linear effects of the gammachirp and illustrate the resulting representation for a specific example of speech.

2 Auditory system

In this section we give a brief and simplified overview of relevant components of the auditory periphery. More detailed information can be found in [2].

2.1 Processing of sound in the auditory periphery

Sound travels through the air as a longitudinal pressure wave. After passing through the outer ear, pressure variations impinge upon the ear drum and are transduced mechanically by bones in the middle ear onto a round window at the base of the cochlea. The cochlea is a rigid, fluid-filled tube which is located in the inner ear. A simplified view of the auditory periphery is shown in Figure 1.

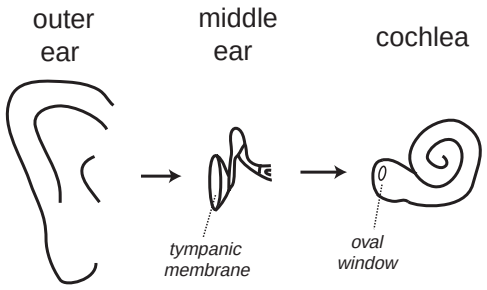

Figure 1: *Pathway of sound through outer ear to tympanic membrane, transduced through the bones of the middle ear, into the cochlea by way of the oval window at the base of the cochlea*

The cochlea is depicted in its uncoiled state in Figure 2. The basilar membrane runs along the length of the cochlea, separating the tube into two chambers. In response to the mechanical action of the input at the base of the cochlea, a standing wave like pattern passes down the basilar membrane. Because of the hydrodynamics of the cochlear fluid and stiffness variation in the membrane, the displacement patterns along the membrane vary depending upon the frequency of the input at the round window. High frequency inputs cause maximal displacement closer to the base of the cochlea, while low frequencies cause maximal displacement at the apex. Inner hair cells situated along the length of the membrane convert the mechanical displacement into neural signals by increasing the firing rates of connected nerve fibers when they are sheared by vertical membrane motion.

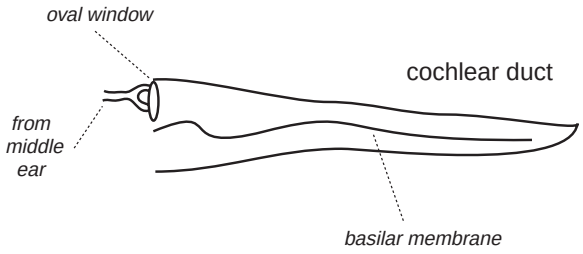

Figure 2: *Caricature of basilar membrane motion in response to pressure at oval window viewed when the cochlear duct is unwrapped*

Outer hair cells, which are collocated with the inner hair cells, are believed to actively dampen or enhance the displacement of the basilar membrane due to input characteristics. Cochlear non-linearity refers to the fact that the displacement due to combined inputs can not be explained by superposition of responses to constituent inputs. One result of this non-linearity is filter responses do not scale directly according to input stimulus level. This nonlinear behaviour is believed to be an important factor which allows humans to hear over a large dynamic range.

Hearing impaired subjects who have damaged outer hair cells lose the compressive non-linearity in their cochlea. A perceptual result of this is abnormal growth of loudness at higher sound intensity levels, known as "loudness recruitment." Because compression does not occur at the physical level in the basilar membrane, the firing rate of auditory nerve fibers saturate at lower sound levels than in normal ears. This can lead to a smaller dynamic range of hearing

2.2 Characteristics of cochlea

The cochlea is often thought of as a bank of filters because it performs frequency analysis using a frequency to place mapping along the basilar membrane. That is, each place along the membrane has a characteristic frequency, f_c, for which it is maximally displaced when a pure tone of that frequency is presented as an input. As a filterbank, the cochlea exhibits the following characteristics:

(a) Non-uniform filter bandwidths. Frequency resolution is higher at the lower frequencies (near the apical end of the cochlea) than at high frequencies (near the basal end of the cochlea). For an equivalent filter bank representation, this implies narrower filters that are more closely spaced together for low frequencies, and broader filters that are spaced further apart for high frequencies.

(b) Asymmetric frequency response of individual filters. For a particular place along the basilar membrane with characteristic frequency f_c, the response to $f_c + \Delta f$ is lower than the response to $f_c - \Delta f$. For a bandpass filter centered at f_c, this can be interpreted as an asymmetric magnitude response, with sharper cutoff on the high frequency side.

(c) Level-dependent frequency response of individual filters. As mentioned in the previous section, basilar membrane motion is compressive and non-linear, meaning that doubling the input stimulus intensity does not result in doubling of membrane displacement. From a filtering perspective, this implies that the peak gain of the filter centered at f_c decreases as the level of the input stimulus increases. Another observation is that the magnitude response of the becomes broader and more symmetric with increasing sound levels.

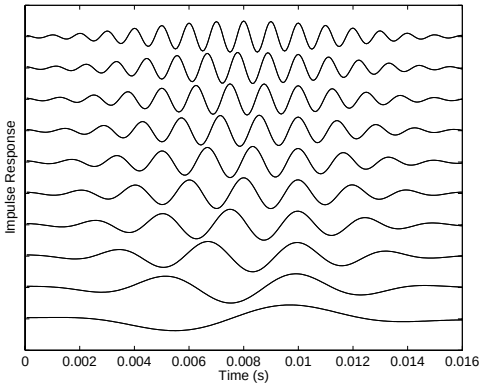

Figure 3: *STFT impulse responses*

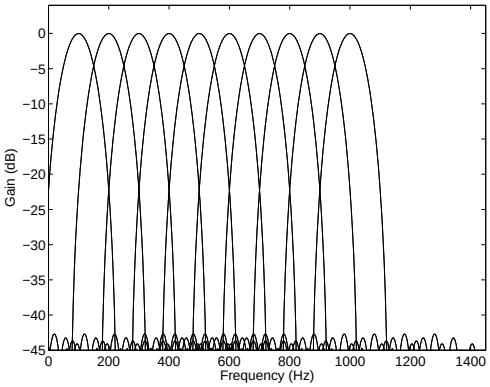

Figure 4: *STFT filterbank*

3 STFT vs. Auditory Wavelet Transforms

In this section, we compare the joint time-frequency representation produced by the short time Fourier transform (STFT) with the joint time-scale representation produced by the auditory wavelet-like transforms produced by the gammatone and gammachirp filters.

3.1 Short Time Fourier Transform

The spectrogram, derived from the short time Fourier transform (STFT), is a common visualization tool used in speech analysis. The STFT is obtained by taking the Fourier transform of localized segments of the time domain signal at fixed time intervals. The signal is localized by multiplying with a shifted window of finite duration. The spectrogram is then obtained by taking the log magnitude of the resulting spectral slices.

In the discrete domain, the STFT is computed using the Fast Fourier Transform (FFT), which computes the frequency content of the windowed signal at uniform frequency intervals. It is possible to think of the STFT as passing the signal through a bank of linear bandpass filters. Each filter has an impulse response which is a modulated version of the window function. In Figure 3, impulse responses are shown which were obtained by modulating a short Hanning window with center frequencies ranging from 100 Hz to 1 kHz. In Figure 4, the same filters are shown in the frequency domain. Each filter has the same magnitude response, but is centered around its modulation frequency.

According to the uncertainty principle, there is an inherent tradeoff between time and frequency resolution which is governed by the duration of the window function. Under the constraints presented by the STFT, Gabor showed that a modulated Gaussian window is optimal for producing minimum uncertainty in the joint time-frequency representation of a signal [3].

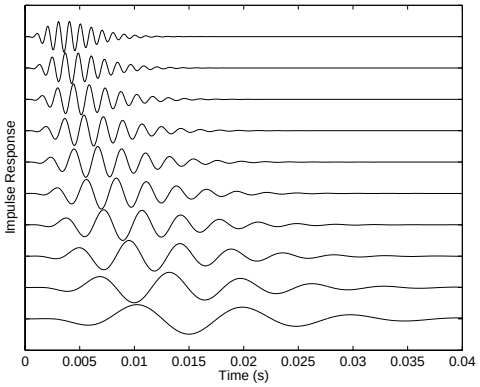

Figure 5: *Gammatone impulse responses*

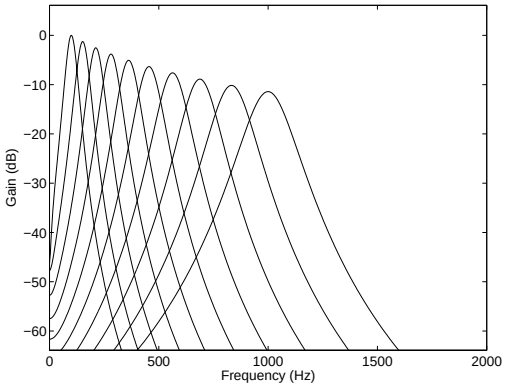

Figure 6: *Gammatone filterbank*

3.2 Gammatone Wavelet Transform

The filtering view described in the previous section illustrated that the filterbanks associated with the STFT have constant bandwidths and are centered at uniformly spaced locations along the frequency axis. In order to better model the frequency response characteristics of the human ear, many researchers use filters inspired by the auditory system which have non-uniform bandwidths and non-uniform spacing of center frequencies. The gammatone filter, developed by Patterson et al [4], is one such filter. Its name is due to the nature of its impulse response, which is a gamma envelope modulated by a tone carrier centered at f_c Hz.

$$g_t(t) = at^{n-1}e^{-2\pi bB(f_c)t}e^{j2\pi f_c t}$$

In this equation, $B(f)$ is the Equivalent Rectangular Bandwidth (ERB) of the center frequency

$$B(f) = 0.1039 \cdot f + 24.7$$

Impulse responses for the gammatone filter are shown at several different center frequencies in Figure 5. The corresponding frequency responses are shown in Figure 6. Passing a signal through a gammatone filterbank is similar to a wavelet transform in that all of the basis functions are scaled and compressed versions of the kernel function at the first center frequency. Narrower support in time corresponds directly to the differences in bandwidth. The center frequencies are chosen by logarithmically sampling points along the frequency axis that lie between the lowest center frequency and the highest center frequency.

3.3 Gammachirp

The gammachirp filter was derived by Irino as a theoretically optimal auditory filter that can achieve minimum uncertainty in a joint time-scale representation. This derivation, which is described in [5], essentially parallels Gabor analysis, but for the wavelet transform. The gammachirp impulse response, shown below,

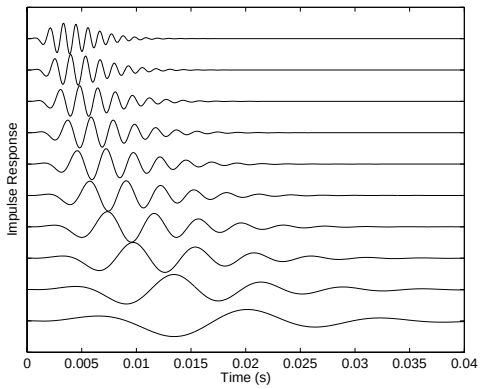

Figure 7: *Gammachirp impulse responses*

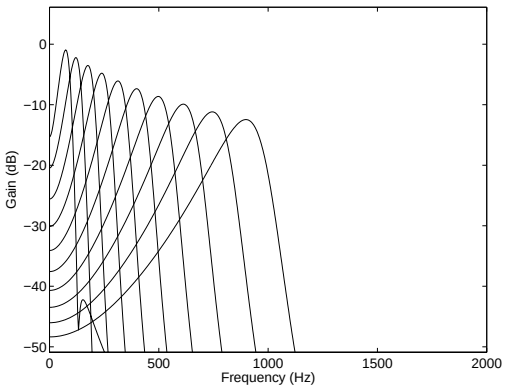

Figure 8: *Gammachirp filterbank*

is essentially identical to that of the gammatone, but also includes a chirp term, c, in the carrier tone.

$$g_c(t) = at^{n-1}e^{-2\pi B(f_c)t}e^{j(2\pi f_c t + c\log t)}$$

The impulse response of the gammachirp at several frequencies are illustrated in Figure 7. The frequency responses of the gammachirp filters, as seen in Figure 8, are asymmetric and exhibit a sharp drop off on the high frequency side of the center frequency. This corresponds well to auditory filter shapes derived from masking data.

The amplitude spectrum of the gammachirp can be written in terms of the gammatone as

$$|G_c(f)| = a_\Gamma(c)|G_T(f)| \cdot e^{c\theta}$$

where $G_C(f)$ is the Fourier transform of the gammachirp function, $G_T(f)$ is the Fourier transform of the corresponding gammatone function, c is the chirp parameter, $a_\Gamma(c)$ is a gain factor which depends on c, and θ is given by

$$\theta = \tan^{-1}\left(\frac{f - f_c}{B(f_c)}\right)$$

This decomposition, which was shown by Irino in [5], is beneficial because it allows the gammachirp to be expressed as the cascade of a gammatone filter, $G_T(f)$, with an asymmetric compensation filter, $e^{c\theta}$. Figure 9 shows the framework for this cascade approach. The spectrum of the overall filter can then be made level-dependent by making the parameters of the asymmetric component depend on the input stimulus level.

4 Implementation

Although basilar membrane impulse response data are available for fitting gammachirp parameters to animal data, human data is only available in the fre-

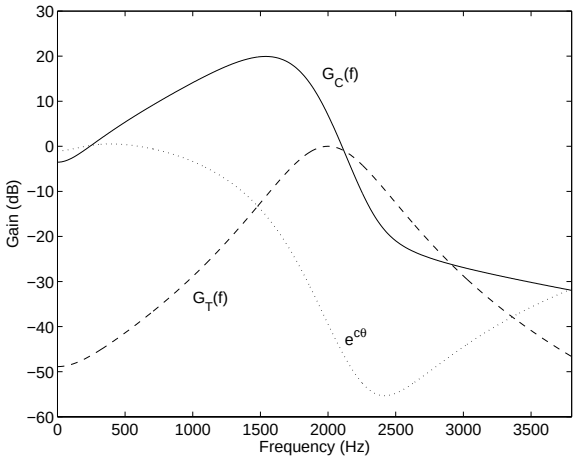

Figure 9: *Composition of gammachirp, $G_C(f)$, as a cascade of a gammatone, $G_T(f)$, with an asymmetric function, $e^{c\theta}$*

quency domain, in the form of data from psychophysical masking experiments. In order to better model this human psychophysical data, a passive gammachirp was used as the level-independent base filter, and a second asymmetric function with varying center frequency was used as the level-dependent component.

For this project, the level-independent, or passive gammachirp, component was specified in the time domain and normalized for the peak gain. The form of the passive gammachirp was

$$g_{pc}(t) = t^3 e^{-2\pi b_1 \cdot B(f_c)t} e^{j(2\pi f_c t + c_1 \log t)}$$

The values for the constants b_1 and c_1 were derived by Irino and Patterson by fitting the frequency curves to notched noise masking data. The numerical values for these parameters are shown in Table 1. This passive linear filter was then cascaded with a asymmetric level-dependent filter to obtain the active compressive gammachirp filter, $g_{CA}(t)$. The amplitude spectrum of this filter is given by

$$|G_{CA}(f)| = |G_{PC}(f)|H_A(f)$$

where $H_A(f)$ is the Fourier transform of the asymmetric level-dependent filter

$$H_A(f) = \exp\left(c_2 \tan^{-1}\left(\frac{f - f_2}{b_2 B(f_2)}\right)\right)$$

In this equation, b_2 and c_2 are constants whose values are shown in Table 1, and f_2 is a level-dependent parameter which specifies the center frequency of the asymmetry.

$$f_2(P_s) = (f_c + c_1 b_1 B(f_c)/3) \times (0.573 + 0.0101(P_s - 80)) \quad (1)$$

Parameter	Value
b_1	2.02
c_1	-3.70
b_2	1.14
c_2	0.979

Table 1: *Parameters used for passive and active gammachirp*

By changing the center frequency of the asymmetry in relation to that of the passive filter, the gain and asymmetry of the overall filter are made level-dependent in a way that agrees with psychophysical data [6]. Figure 10 demonstrates the combination of the component filters to produce the active gammachirp at several gain levels.

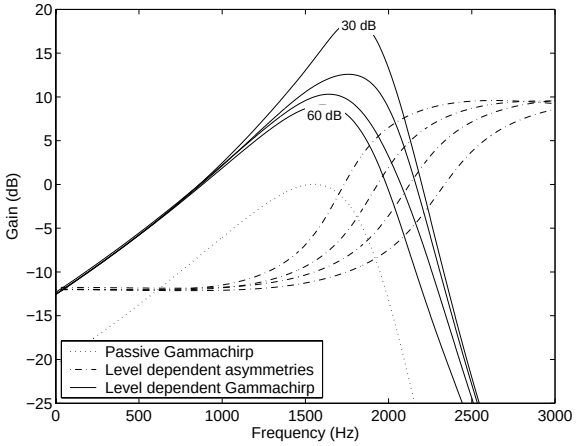

Figure 10: *Composition of gammachirp as a cascade of a gammatone with an asymmetric function*

4.1 IIR approximation of Asymmetry

Because the form of the asymmetric component, $H_A(f)$, is difficult to specify in the time domain, a fourth-order IIR approximation to the asymmetric component was developed in [7]. The discrete filter, $H_c(z)$, was designed to provide a close fit to the compensation filter, $H_A(f)$, in the region of interest around the center frequency, f_2.

$$H_A(f) \approx H_c(z)|_{z=e^{j2\pi f/f_s}}$$

$$H_c(z) = \prod_{k=1}^{4} H_{ck}(z)$$

$$H_{ck}(z) = \frac{1}{|H_{ck}(e^{j2\pi f_k/f_s})|} \frac{(1 - 2r_k \cos(\psi_k) z^{-1} + r_k^2 z^{-2})}{(1 - 2r_k \cos(\phi_k) z^{-1} + r_k^2 z^{-2})}$$

For each second order filter, $H_{ck}(z)$, the parameters are:

$$r_k = \exp\left(\frac{-k p_1 2\pi b_2 B(f_2)}{f_s}\right)$$

$$\phi_k = 2\pi \frac{f_2 + p_0^{k-1} p_2 c_2 b_2 B(f_2)}{f_s}$$

$$\psi_k = 2\pi \frac{f_2 - p_0^{k-1} p_2 c_2 b_2 B(f_2)}{f_s}$$

$$f_k = f_2 + k \cdot p_3 c_2 b_2 B(f_2)/3$$

In these equations, f_s is the sampling rate, and p_0, p_1, and p_2 are positive coefficients which were determined heuristically in terms of c_2, and

$$\begin{array}{ll} p_0 = 2, & p_1 = 1.35 - 0.19|c_2| \\ p_2 = 0.29 - 0.0040|c_2|, & p_3 = 0.23 + 0.0072|c_2| \end{array}$$

Figure 11 shows a comparison between the actual compensation filter and the fourth order approximation filter at several center frequencies. Within the bandpass region for the center frequencies, the approximation error is small. In this project, the approximation filter was used for the level-dependent component filter in the active gammachirp.

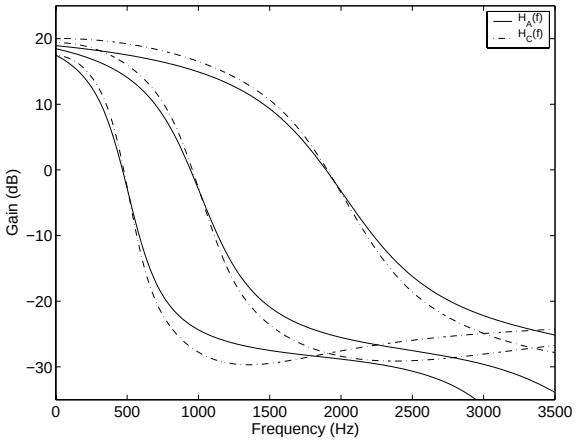

Figure 11: *Amplitude spectra of the asymmetric compensation filters, $H_A(f)$ and $H_C(f)$, at several center frequencies together with their associated IIR approximation filters*

4.2 Incorporating Level Dependency

Because the gammachirp is level-dependent, an estimate of the current input stimulus level must be obtained in order to specify the filter characteristics. In other words, the gammachirp filterbank must be adaptive. Irino has proposed two schemes for incorporating level dependency into frequency analysis by gammachirp filterbanks [7] [8]. However, in both of these schemes, the chirp term, c, was used as the level-dependent parameter.

The approach used in this paper keeps all parameters fixed except for the center frequency of the asymmetric approximation filter. A block diagram of the system is shown in Figure 12. To estimate the value of P_s for equation 1, we calculated a moving average of the energy in each frequency channel. For each center frequency, f_c, the input signal was passed through a second order Butterworth bandpass filter with bandwidth $B(f_c)$. The moving average was then calculated over a windowed segments of the waveform. The duration of each segmented portion of the waveform was 10 milliseconds.

An alternative to updating parameters is to simply generate level estimates for each channel by averaging over the entire utterance. This strategy involves significantly less computation, but is also less adaptive to non-stationary noise.

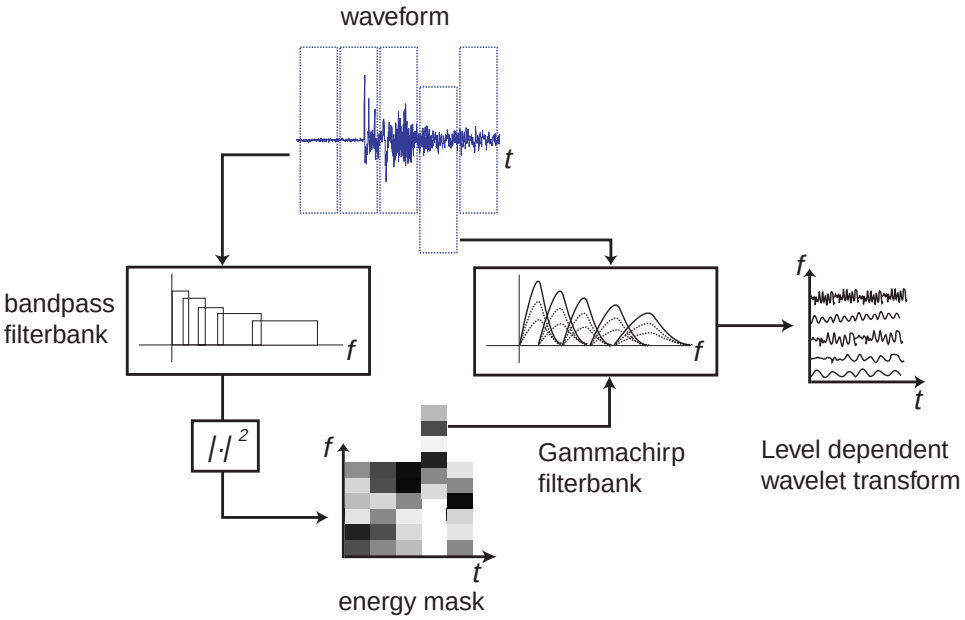

Figure 12: *Framework for estimating energy level for parameter control of gammachirp filterbank*

5 Results and Discussion

In this section, we illustrate the various output representations that are generated by the gammachirp filterbank and compare them to gammatone and STFT representations.

Figures 13, 14, and 15 show the STFT spectrogram and the gammatone and gammachirp scalograms for the spoken digit string "One Two Eight" amidst varying levels of background noise. One immediate difference that can be noticed in comparing the spectrogram and scalogram outputs is the scaling of the frequency axis. Due to difficulties in estimating peaks for the non-uniform characteristic frequencies for the wavelet filters, we were unable to label the frequency axis with the correct center frequencies. By looking at spectral landmarks however, it is evident that the non-uniform spacing of the center frequencies for the scalograms result in a larger gap between the first and second formants when compared to the STFT spectrogram. The higher resolution of the low frequency region is likely to be useful for determining vowel type, since vowels are typically defined by the relative positions of the first two formants.

Because the values of the scalograms and spectrogram are log compressed, it is difficult to observe the compressive effect of the gammachirp. However, for both the gammatone and gammachirp outputs, spectral peaks for voiced segments of speech appear to be more prominent against the background in all three noise conditions than for the STFT spectrogram. Since voicing tends to be a cue that is easily distinguished even at relatively low SNR levels, more spectral detail for voiced segments of speech would be for speech analysis in noise.

Although the gammatone and gammachirp scalograms appear very similar, there are several noticeable differences. First, in the segment between 0.2 and 0.3 seconds, the gammatone output exhibits a more pronounced second formant than for the gammachirp. On the other hand, the low frequency resonances appear to be more strongly emphasized by the gammachirp, and the bandwidths of most resonances also appear to be much narrower.

For a more detailed comparison of the two wavelet transforms on clean speech, Figure 16 shows the gammatone and gammachirp scalograms for the spoken utterance "tapestry". In the sonorant region between 0.1 and 0.2 seconds, the gammatone output appears to have a more continuous transition of spectral peaks. The temporal discontinuity observed in the gammachirp scalogram at 0.15 seconds could likely be smoothed away by using a shorter time window for level estimation.

6 Conclusions and Future Work

This paper reviewed the background and theory of the compressive gammachirp auditory filter proposed by Irino and Patterson. The motivation for studying this auditory filter is to improve the front end signal processing strategies employed by automatic speech recognition systems. The gammachirp was com-

pared to both the short time Fourier transform and the gammatone filter from a wavelet perspective and a level-dependent version of the gammachirp filterbank was implemented in Matlab. Preliminary investigation of speech representations derived from these filtering approaches indicate that both wavelet transforms appear to preserve salient spectral features across several noise conditions.

Although this project focused on the compressive properties of the gammachirp, it would be useful to examine how well it models the multi-tone suppression effect. The suppression effect may be helpful for enhancing maxima in the amplitude spectrum, thus making formant peaks more salient relative to neighbouring frequency channels. An immediate direction for future work would be to utilize this effect to improve formant extraction.

A second possibility for future work is to integrate the level dependent filterbank with the second and third stages of a more complex auditory model proposed by Seneff [9]. In that model, linear auditory filterbanks were designed which had characteristics similar to the passive gammachirp, but were not level-dependent.

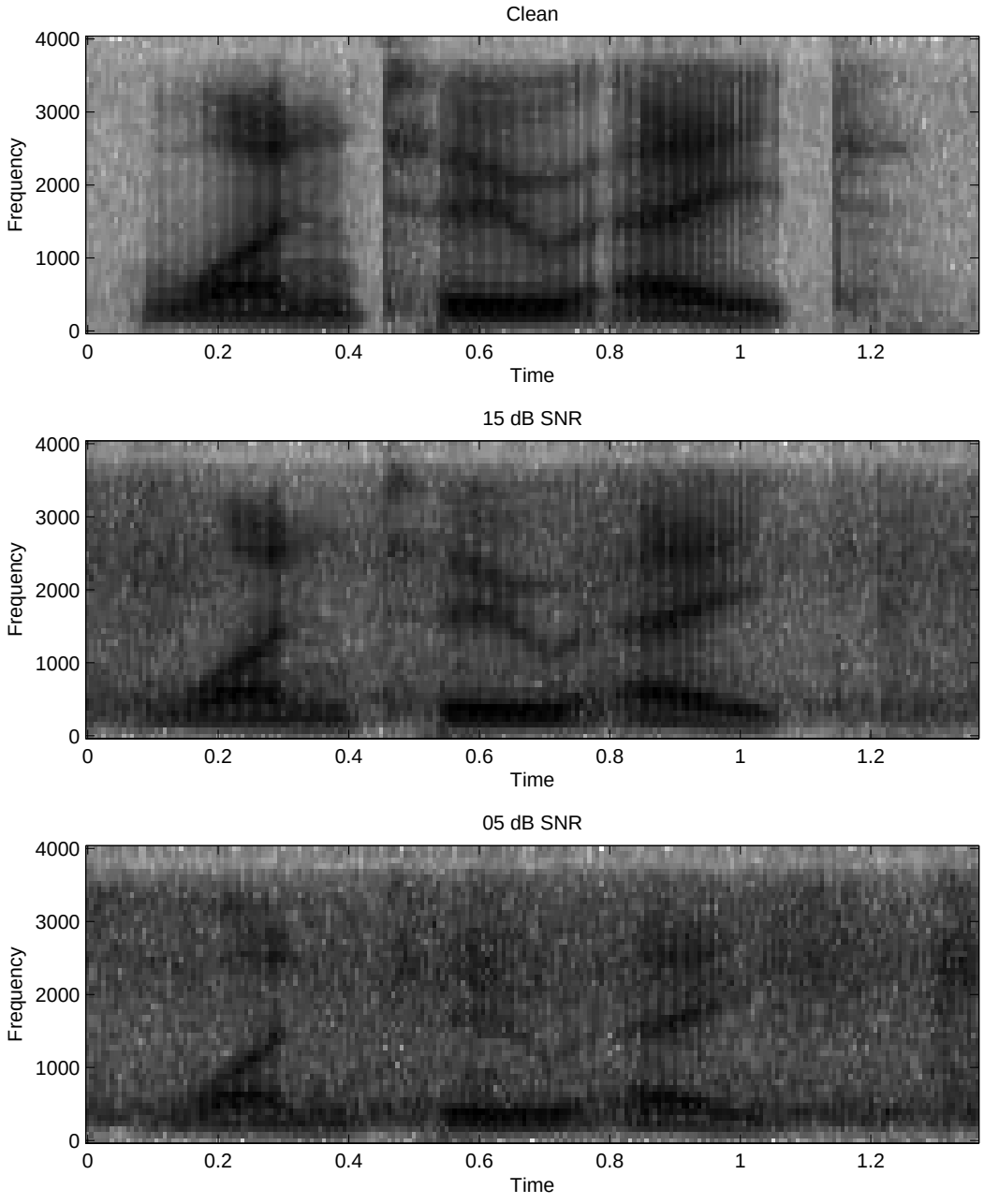

Figure 13: *STFT spectrograms of the digit string "One Two Eight" in varying noise levels*

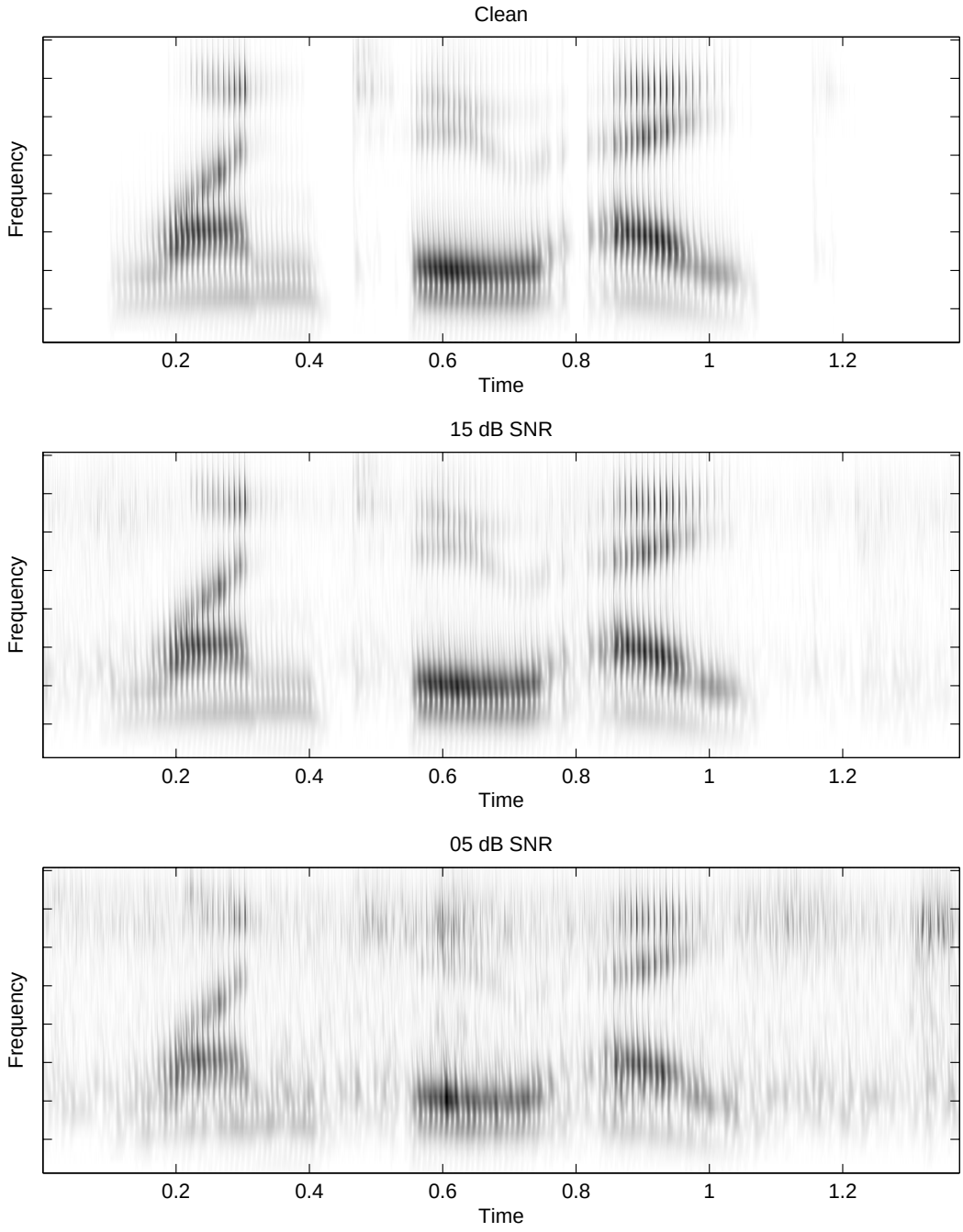

Figure 14: *Gammatone scalograms of the digit string "One Two Eight" in varying noise levels*

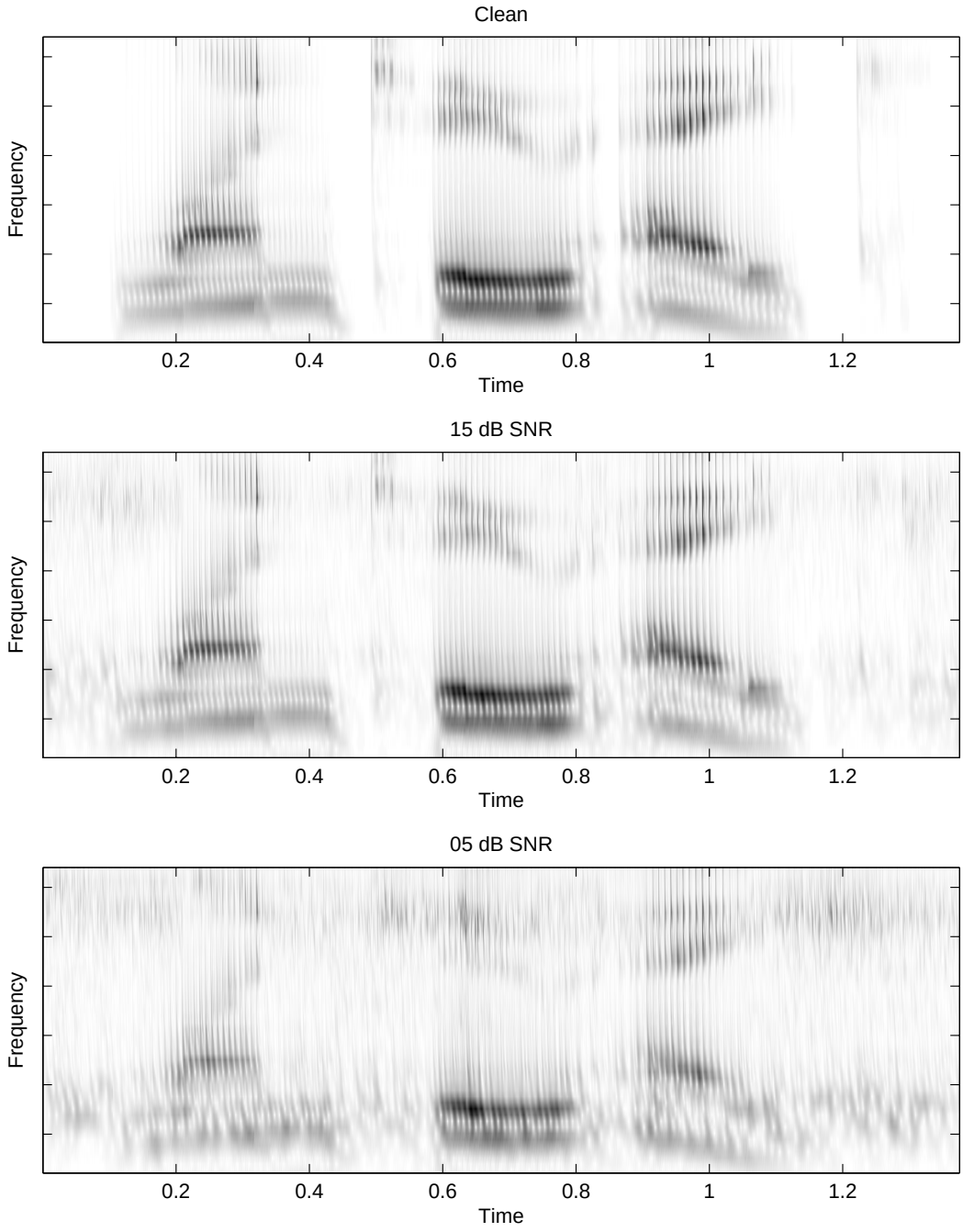

Figure 15: *Gammachirp scalograms of the digit string "One Two Eight" in varying noise levels*

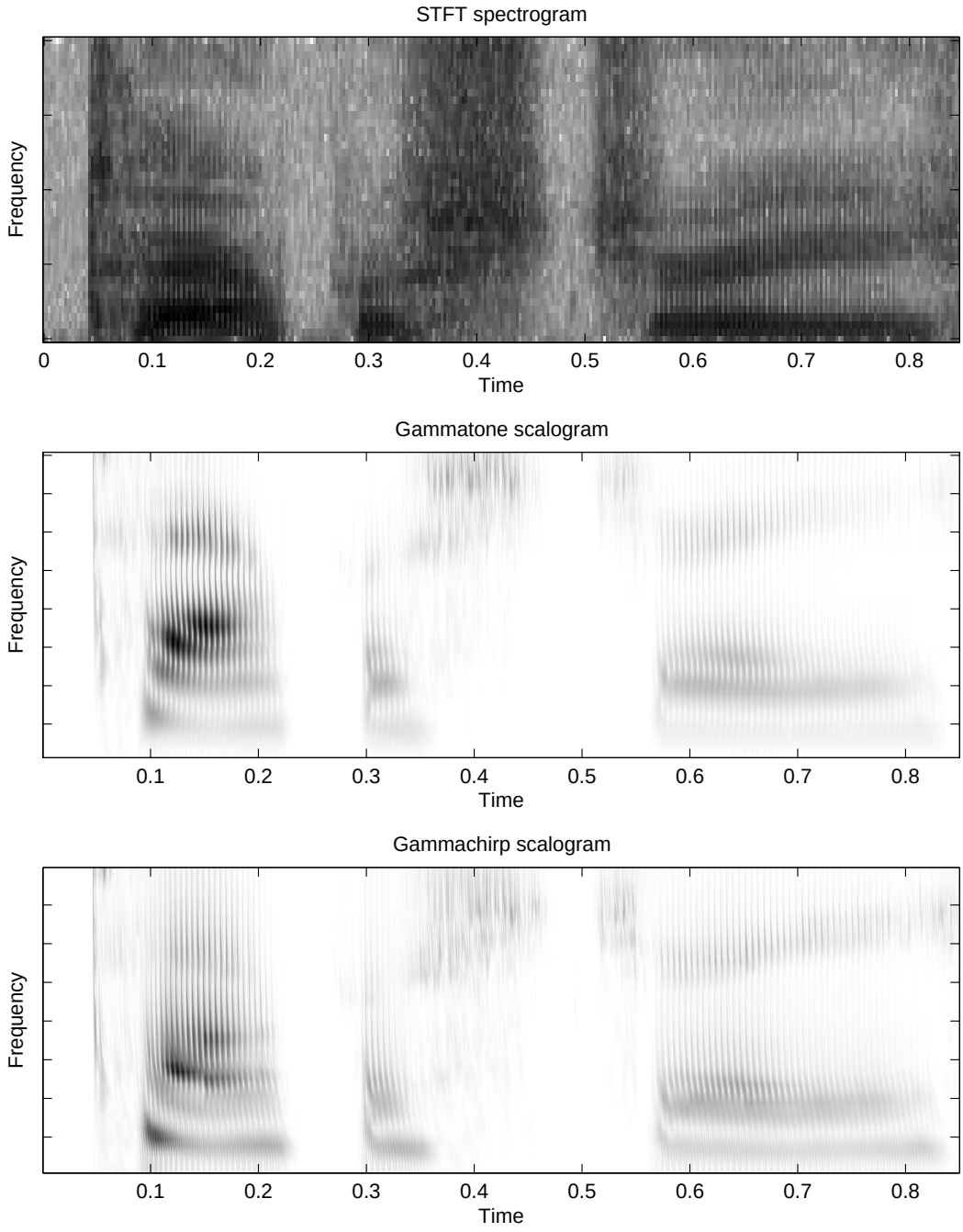

Figure 16: *Spectrogram and scalograms for the spoken utterance "tapestry" in clean background conditions*

References

[1] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[2] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, 2nd edition edition, 1988.

[3] L. Cohen, "Time-frequency distributions - A review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–980, July 1989.

[4] R. D. Patterson, K. Robinson, J. W. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., pp. 429–446. Pergamon, Oxford, 1992.

[5] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory fiilter: the gammachirp," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 412–419, January 1997.

[6] T. Irino and R. D. Patterson, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2008–2022, May 2001.

[7] T. Irino and M. Unoki, "An analysis/synthesis auditory filterbank based on an IIR implementation of the gammachirp," *J. Acoust. Soc. Jap.*, vol. 20, no. 5, pp. 397–406, November 1999.

[8] T. Irino, "Noise suppression using a time-varying, analysis/synthesis gammachirp filterbank," in *Proc. ICASSP*, Phoenix, AZ, 1999.

[9] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.