TECHNICAL RESEARCH REPORT

Spectro-Temporal Modulation Transfer Functions and Speech Intelligibility

by Taishih Chi, Yujie Gao, Matthew C. Guyton, Powen Ru, Shihab Shamma

CAAR T.R. 99-2
(ISR T.R. 99-61)

The Center for Auditory and Acoustic Research (CAAR) is a consortium of researchers from six universities working in partnership with Department of Defense laboratories and industry. CAAR is funded by the Office of Naval Research through a 1997 Department of Research Initiative.

Web site http://www.isr.umd.edu/CAAR/

Spectro-Temporal Modulation Transfer Functions and Speech Intelligibility

Taishih Chi, Yujie Gao, Matthew C. Guyton, Powen Ru, and Shihab Shamma
Center for Auditory and Acoustics Research, Institute for Systems Research
Electrical Engineering Department, University of Maryland, College Park, MD 20742

ABSTRACT

Detection thresholds for spectral and temporal modulations are measured using broadband spectra with sinusoidally rippled profiles that drift up or down the log-frequency axis at constant velocities. Spectro-temporal Modulation Transfer Functions (MTF) are derived as a function of ripple peak density (Ω cycles/octave) and drifting velocity (ω Hz). MTFs exhibit a lowpass function with respect to both dimensions, with 50% bandwidths of about 16 Hz and 2 cycles/octave. The data replicate (as special cases) previously measured purely temporal MTFs ($\Omega = 0$) [Viemeister, 1979] and purely spectral MTFs ($\omega = 0$) [Green, 1986]. We present a computational auditory model that exhibits spectro-temporal MTFs consistent with the salient trends in the data. The model is used to demonstrate the potential relevance of these MTFs to the assessment of speech intelligibility in noise and reverberant conditions.

INTRODUCTION

The most obvious feature of a speech spectrogram is the energy modulations, both in time in any given frequency channel, and along the spectral axis at any instant, due to formant peaks and their transitions, spectral edges, and rapid amplitude modulations at onsets/offsets. These modulations occur at relatively slow temporal rates (few Hz) reflecting the speed of the articulatory gestures, and hence the phonetic and syllabic rates of speech. Speech intelligibility is critically dependent on the clarity of these spectro-temporal modulations. Thus speech reconstructed from smoothed spectrograms along either dimension suffers from progressive loss of intelligibility [Shannon et al., 1995, Arai et al., 1996, Drullman, Festen, and Plomp, 1994].

Human sensitivity to spectral and temporal modulations has been studied extensively in various experimental settings. In most cases, these two measurements are treated separately. For instance, sensitivity measurements to *purely temporal* modulations - usually referred to as a temporal "modulation transfer function (**MTF**)" are illustrated in Figure 1(a). They employ either amplitude modulated white noise [Viemeister, 1979] (Fig.1(a), left panel) or temporally modulated harmonic-like spectra [Yost and Moore, 1987, van Zanten and Senten, 1983] (Fig.1(a), right panel). While both MTFs are lowpass in character, they exhibit substantially different upper cutoff rates, with flat noise being detectable to much higher rates (exceeding 64 Hz compared to less than 10 Hz for the noise-delayed stimulus). Complementary tests of *purely spectral* sensitivity are shown in Figure 1(b). They employ stationary (static) spectra with sinusoidal envelopes along the logarithmic frequency axis - also called

ripples [Hillier, 1991, Green, 1986]. This spectral MTF demonstrates that our ability to detect closely spaced ripple peaks deteriorates above about 4 cycles/octave[1]. Spectral MTFs in birds also exhibit similar trends and upper limits, although there is some variability across different species [Amagai et al., 1999].

All MTFs described in Fig.1 are essentially one-dimensional in that they are measured by varying either the spectral or temporal modulation rates while holding the other constant. This is the case even for the spectrally complex stimuli in [Yost and Moore, 1987, van Zanten and Senten, 1983] because such harmonic-like spectra preserve their shape against the tonotopic (logarithmic frequency) axis of the auditory system regardless of the change in their frequency spacing. Consequently, MTFs are effectively always measured with the same spectral pattern (except for a translation to a different frequency region)[2].

Modulations in speech spectrograms are usually combined spectro-temporal modulations. Thus, speech is rarely a flat modulated spectrum nor is it a stationary peaked spectrum, but rather it is both - a spectrum with dynamic peaks. Therefore, sensitivity to these types of combined modulations relates directly to speech perception. But, are spectro-temporal MTFs separable? That is, can the combined spectro-temporal MTF be derived from a product of purely temporal and spectral MTFs?

A particularly useful and simple example of a combined spectro-temporal modulation is the spectral ripple that drifts upwards or downwards at a constant velocity, as illustrated in Figure 2(a). By varying the density of the peaks along the spectral axis (Ω, cyc/oct), and the drifting speed (ω, Hz) and directions, it is possible to measure a full combined spectro-temporal MTF. These stimuli are also interesting from a theoretical perspective in that they form a complete set of orthonormal basis functions for the spectrogram. Thus, any arbitrary spectrogram can be decomposed (by a two-dimensional Fourier Transform) into a linearly weighted sum of such drifting ripples with different spectral densities, velocities and directions (Fig.2(b)). Because of this property, ripples have played a useful role in characterizing the linear aspects of the spectro-temporal response fields in the auditory cortex [Simon, Depireux, and Shamma, 1998]. Moreover, there is neurophysiological evidence that these dynamically rippled spectra are especially effective in eliciting responses in the auditory cortex [Kowalski, Depireux, and Shamma, 1996, deCharms, Blake, and Merzenich, 1998].

In order to characterize the role of these modulations in the perception of speech and other complex sounds, we have measured the sensitivity of human subjects to spectro-temporal modulations over the perceptually important range of 0.25-8 cycles/octave and 1-128 Hz. The experimental methods and results are discussed in the next section (II). A simplified model of early and central auditory processing that accounts for the major trends in the

[1]Sensitivity also decreases to very low ripples (below about 1 cyc/oct) although this is probably due to the design of the test which uses a flat spectrum as the standard.

[2]This is the case for one of two stimuli presented in [Yost and Moore, 1987]. In the other, the spectrum is translated periodically at different rates along the linear frequency axis. Along the logarithmic frequency axis, the spectral pattern periodically changes shape in a complicated manner, but thresholds are measured only with respect to the change in temporal rates. Thus, strictly speaking this stimulus does not fit neatly in the purely temporal vs. spectral classification.

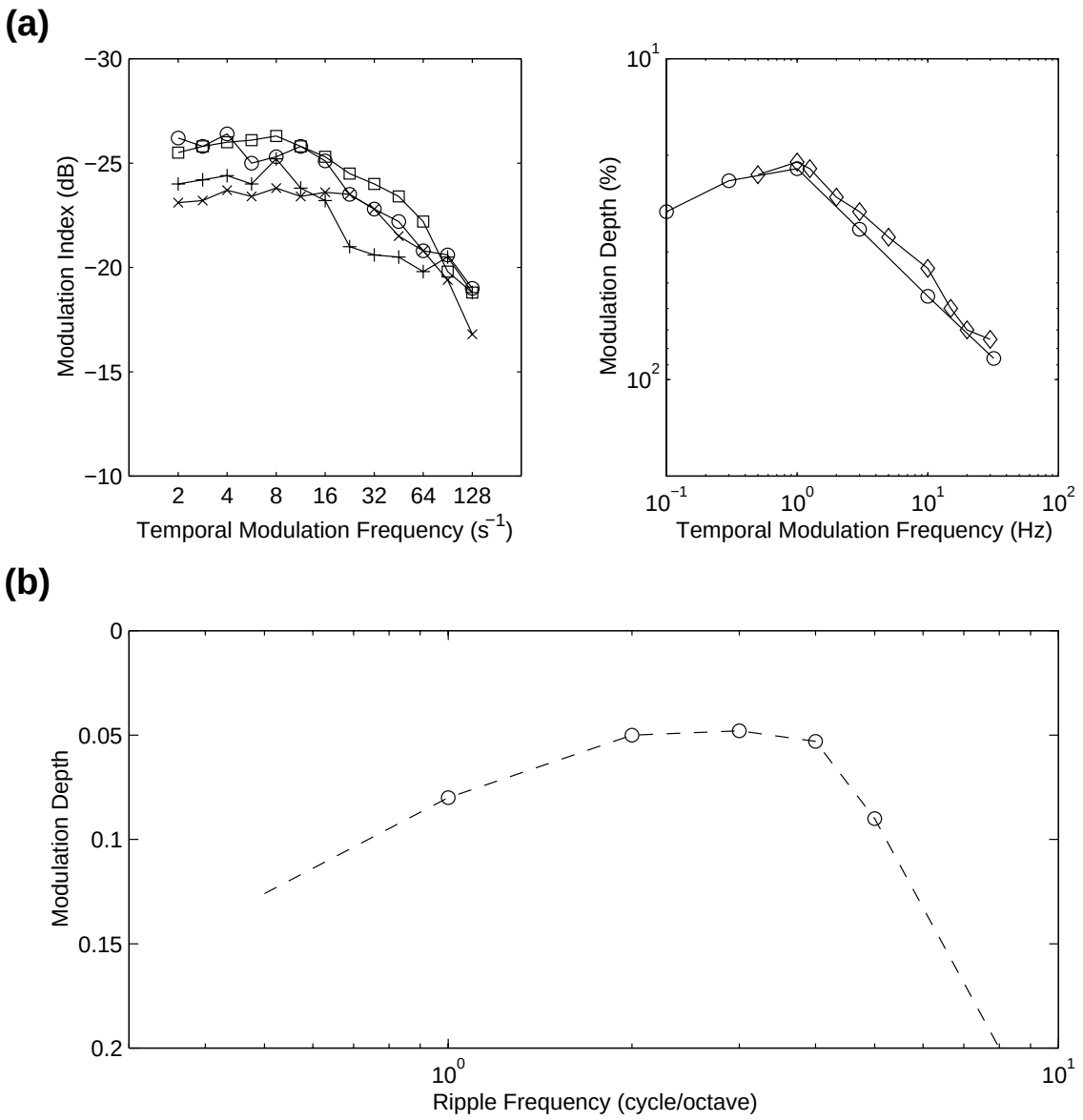

Figure 1: Temporal and spectral modulation transfer functions (MTF). (**a**) *Left panel* - Temporal MTFs measured using amplitude modulated white noise [Viemeister, 1979]. *Right panel* - Temporal MTFs measured using rippled noise with sinusoidally modulated delays [van Zanten and Senten, 1983]. (**b**) Spectral modulation transfer functions measured using ripples [Green, 1986].

data is presented in section III. In section IV, the utility of the model is demonstrated by a preliminary evaluation of the intelligibility of speech in different kinds of noise. Finally, we discuss the relevance of these results in the wider context of visual and other auditory tasks.

I: METHODS

Psychoacoustical MTFs were measured for four subjects who are graduate students familiar with this task. The results shown below followed a period of training after which subjects' performance stabilized with little further improvements.

A. Testing procedures

A "two-alternative two-interval" forced choice adaptive procedure was used to estimate thresholds. Each trial consisted of two 1 second long observation intervals separated by 200 ms pause. After listener's response, a short visual feedback was provided and a new trial started until all 50 trials that comprise one block were presented.

The discrimination task was to distinguish between a spectrally flat *standard*, which did not change over a block of trials, and the *signal*, which resembled the *standard* except for an added modulation on the profile whose amplitude changed in steps adaptively. On the first trial the signal was three step sizes away from the standard. On each subsequent trial the signal was changed according to the "two down-one up" procedure in order to estimate the level that produces 70.7% correct answers [Levitt, 1971]. The step size was halved after three reversals and the threshold was estimated as the average of the signal across the last even number of reversals, excluding the first three. Signal and standard occurred with equal *a priori* probability in one of the two intervals. The overall presentation level was randomized across trials and within a trial over a 20 dB range in 1 dB resolution, in order to ensure that listeners based their judgement on a change in spectral shape rather than on absolute level change in a particular frequency band [Green, 1986].

B. The moving ripple stimulus

In all tests, sounds were generated digitally with 16-bit resolution and 16 kHz sampling rate. They were low-pass filtered at 8 kHz. Before presentation to listeners, sounds were gated for a 1 sec duration, including 10 ms rise and decay ramps. Sounds were delivered inside an acoustic chamber through a loudspeaker (ADS L470).

The broadband ripple spectra consisted of 92 tones equally spaced along the logarithmic frequency axis and spanning 5.75 octaves (0.14-7.34 kHz), as illustrated in Fig.2(a). The *spectral envelope (or profile)* of the complex was modulated as a single sinusoid along the logarithmic frequency axis on a linear amplitude scale (Fig.2(a)). The amplitude of the ripple profile (A) is defined relative to the unit base or flat spectrum. Thus, $A = 0$ to 1 corresponds to 0 to 100% modulation of the flat ripple profile. The ripple density Ω is in units of cycles/octave (cyc/oct). The ripple phase is given in radians or degrees relative to a sine wave starting at the low frequency edge of the complex (Fig.2(a)). Therefore, the profile of a *stationary* ripple spectrum is given by

$$S(x) = A \cdot \sin(2\pi \cdot \Omega \cdot x + \Phi) \ , \tag{1}$$

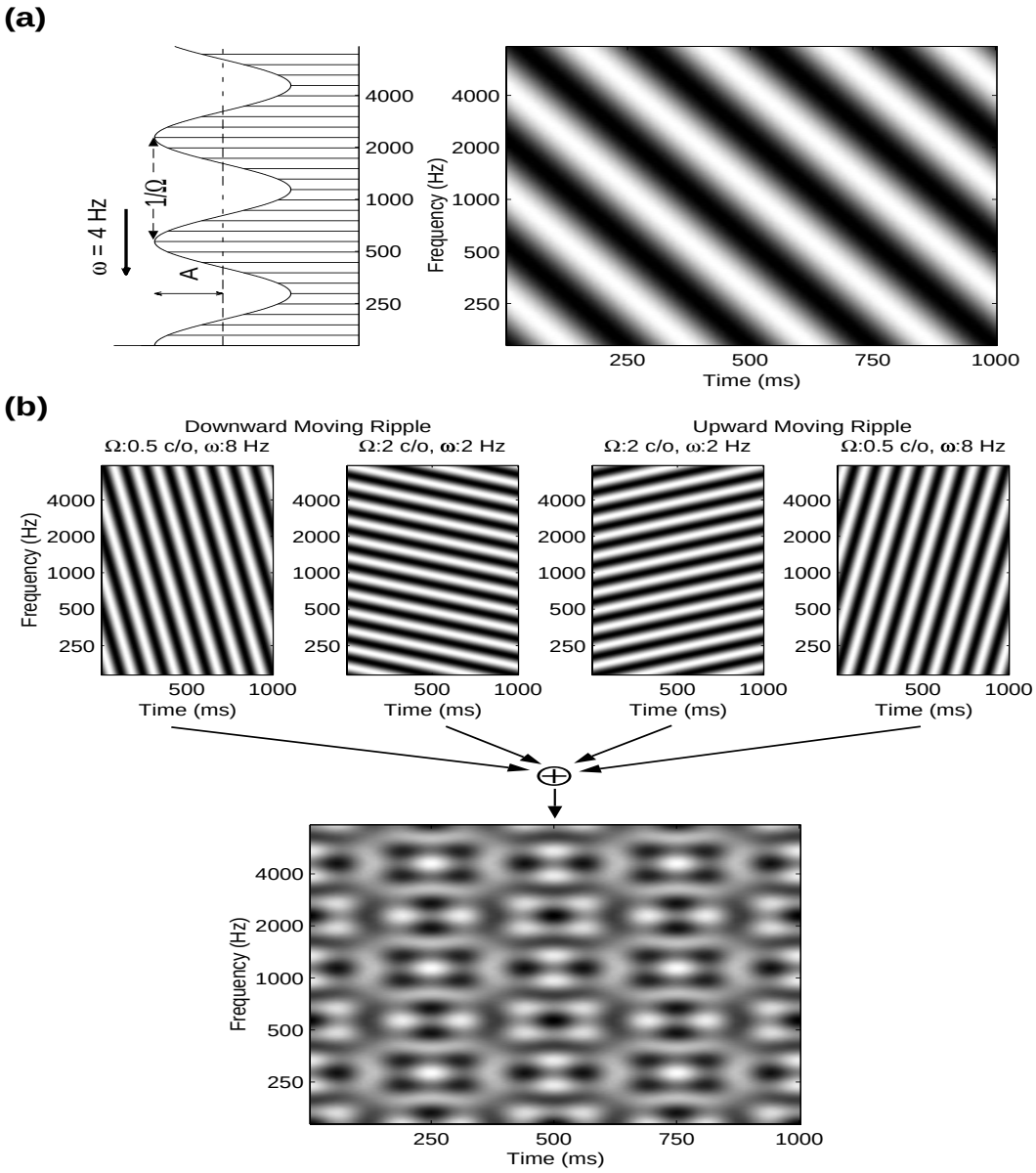

Figure 2: Moving ripples: Parameters and motivations. (**a**) Definition of moving ripple spectrum parameters. *Left panel* illustrates a ripple spectrum with an envelope of amplitude A, sinusoidally modulated along the spectral axis x with a density of $\Omega = 0.5$ cyc/oct, and is moving at velocity $\omega = 4$ Hz. *Right panel* displays the ripple spectrogram . (**b**) A weighted sum of ripples can be used to construct any arbitrary spectrogram. *Top row* illustrates spectrograms of four upward/downward moving ripples with different Ω, ω combinations. Adding these ripples produces a complex spectrogram.

where x is the position on the logarithmic frequency axis (in octaves) defined as: $x = log_2(\frac{f}{f_o})$ with f_0 the lower edge of the spectrum (0.14 kHz), and f as frequency; Φ is the phase of the profile.

The stimuli of interest in this study, however, are also modulated in time by having the ripple profile move up or down the spectral axis at a constant velocity. Ripple velocity (ω) is defined as the number of ripple cycles-per-second (Hz) sweeping past the low frequency edge of the spectrum. The resulting moving ripple profile is fully characterized by:

$$S(x,t) = A \cdot \sin(2\pi \cdot (\omega \cdot t + \Omega \cdot x) + \Phi) \ . \tag{2}$$

Therefore, a positive (negative) ω corresponds to a ripple envelope drifting downward (upward) in frequency.

Fig.2(b) illustrates spectrograms of moving ripple profiles with different (Ω,ω) combinations. Note that the spectrograms appear as two-dimensional gratings with orientations determined by the ratio of the spectral to temporal modulation rates (Ω/ω). A program to generate these stimuli interactively is available at *http://www.isr.umd.edu/CAAR/pubs.html*.

II. RESULTS

Threshold measurements can be conceptually inverted and interpreted as sensitivity measures to different spectro-temporal modulations, hence reflecting the gain of the system or its modulation transfer function, e.g., as in [Viemeister, 1979, Yost and Moore, 1987]. The average thresholds for four subjects are presented in Figure 3 as a function of Ω and ω for upward and downward drifting ripples.

The data generally exhibit a lowpass function in both dimensions. Sensitivity slightly peaks in a small region around 2-8 Hz. Subjects maintain high sensitivity to temporal modulations of low Ω spectra up to 32 Hz; in fact, temporal MTFs at $\Omega = 0.25 - 2$ cyc/oct are almost identical to those measured by [Viemeister, 1979] with flat spectra. The data also suggest that, apart from an overall decrease in sensitivity, the temporal transfer functions approximately preserve their lowpass shape at higher Ω. For instance, temporal MTFs at Ω =0.25, 4, and 8 cyc/oct are approximately shifted upwards relative to each other reflecting the rising detection thresholds to high Ω (Fig.3). This implies that the spectro-temporal MTF is approximately the product of purely temporal [Viemeister, 1979] and purely spectral [Green, 1986] MTFs, i.e., it is *separable*.

To confirm this impression, we have applied the Singular Value Decomposition (SVD) method to analyze the data in Fig.3(b). Specifically, the MTF is diagonalized as $\Lambda = U \cdot MTF \cdot V$, where Λ is the eigenvalue matrix, and U, V are the corresponding eigenvectors [Haykin, 1996]. If the MTF matrix is separable, i.e., expressible as the product of two vectors (a purely temporal and spectral MTF), then it should have only one nonzero eigenvalue; otherwise, relatively sizable secondary eigenvalues occur. Fig.3(c) illustrates the results of such an analysis on the MTFs shown in Fig.3(a,b). Only one significant eigenvalue is found ($\frac{\lambda_i}{\lambda_1} < 15\%$ for $i = 2...6$). The corresponding purely temporal and spectral threshold functions (or MTFs) are shown in Fig.3(c). The spectro-temporal MTF surface reconstructed by a

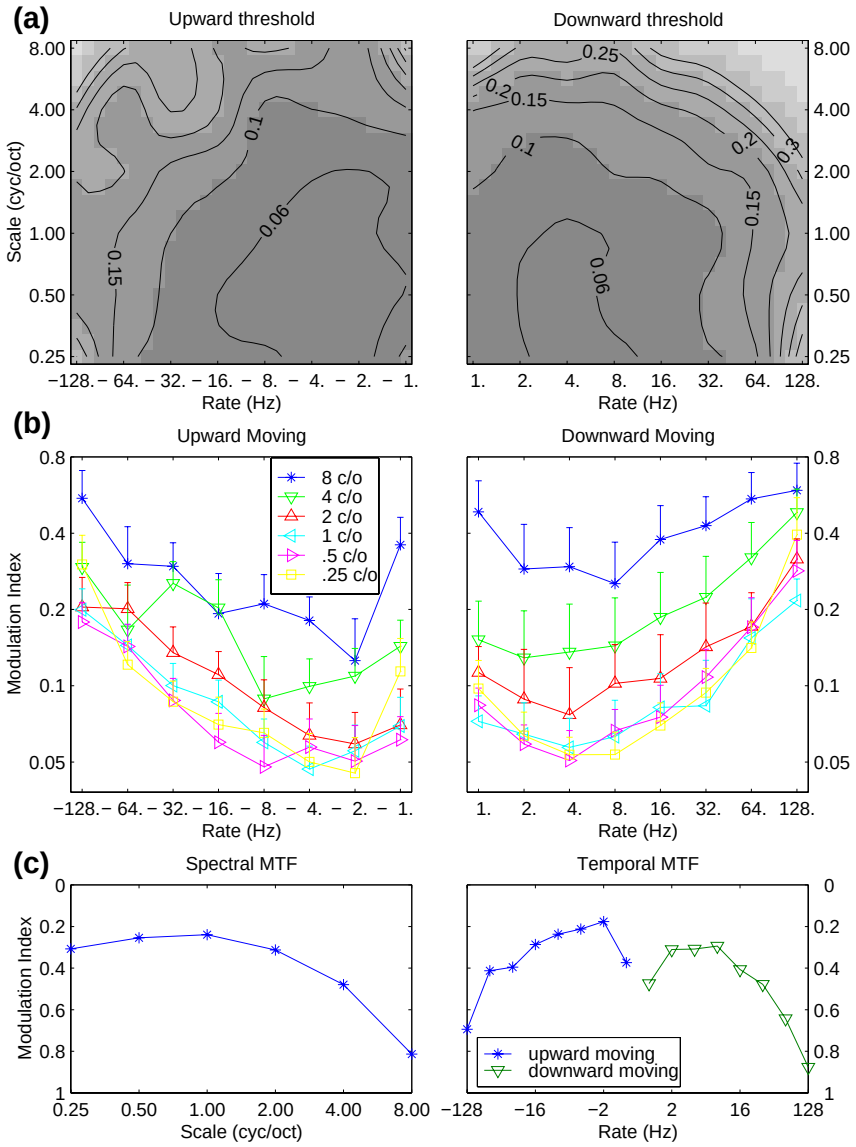

Figure 3: Detection thresholds of moving ripples as a function of ripple density (Ω), velocity (ω), and direction. (**a**) Contour and gray scale interpolated representation of the ripple amplitudes at threshold. (**b**) Same thresholds as above plotted as a function of ripple velocity, with density as parameter. For display purpose, only the upper error bars with half the standard deviation error ranges are shown. (**c**) The one-dimensional temporal (left) and spectral (right) MTFs (or thresholds) derived by singular-value-decomposition procedure from the combined data in (**b**) above (see text for details).

pure product of these two functions produces an MTF that is to within 3.4% of the original data (in mean-square-error sense) . This error is well within the bounds of the experimental errors indicated by the bars. These results strongly argue for the full separability of the MTFs in Fig.3(a,b).

III: MODEL

A model of spectro-temporal modulation sensitivity is developed here to explain the origin of the salient trends in the data, and to serve as a computational module in applications requiring analysis of the spectro-temporal modulations in sound spectrograms. The model is inspired and is consistent with known biophysics of the peripheral auditory system, and with single unit responses in the primary auditory cortex [Kowalski, Depireux, and Shamma, 1996]. It consists conceptually of two parts: (1) An early auditory portion which models the transformation of the acoustic signal into an *auditory spectrogram*, and (2) a central portion which further analyzes the auditory pattern into a modulation *scale-rate* plot using a family of cortical-like filters.

We first describe schematically these two stages. Next, a precise mathematical formulation is presented, followed by a brief illustration of the model overall MTF. All model stages are available as a MATLAB library on *http://www.isr.umd.edu/CAAR/pubs.html* (NSL Tools Package).

A. The early auditory spectrogram

The early stages of the auditory system transform sound into a pattern of neural activity that represents an enhanced and noise-robust version of the acoustic spectrum, henceforth called the *auditory spectrogram*. Extensive details of the biophysical basis, anatomical structures, and computational implementation of the model used here to generate the auditory spectral profiles are available in [Yang, Wang, and Shamma, 1992, Wang and Shamma, 1994]. Figure 4 illustrates the various stages of the model. Briefly, it consists of a bank of 120 asymmetric critical overlapping bandpass filters that are equally spaced over a 5 octave frequency range (24 filters/octave) (see [Wang and Shamma, 1994] for details of the filter parameters and implementations). The output of each filter is processed by a hair cell stage which consists of a highpass filter, followed by nonlinear compression (optional), and then a lowpass filter [Shamma et al., 1986]. The final stage mimics the action of a lateral inhibitory network [Shamma, 1988] which sharpens the filter outputs, and hence the filter bank frequency selectivity. It is implemented by a first-difference operation across the channel array, followed by a half-wave rectifier, and a short-term integration to estimate the final output [Yang, Wang, and Shamma, 1992]. Figure 5(a) shows the auditory spectrogram of a speech sentence "*Come home right away*". Spectrograms of other sentences and moving ripples are illustrated in Figures 5 and 6(b,c).

B. The modulation scale-rate plot

The auditory spectrum is relayed to the primary auditory cortex (AI) through several stages of processing [Clarey, Barone, and Imig, 1992]. AI responses integrate influences from

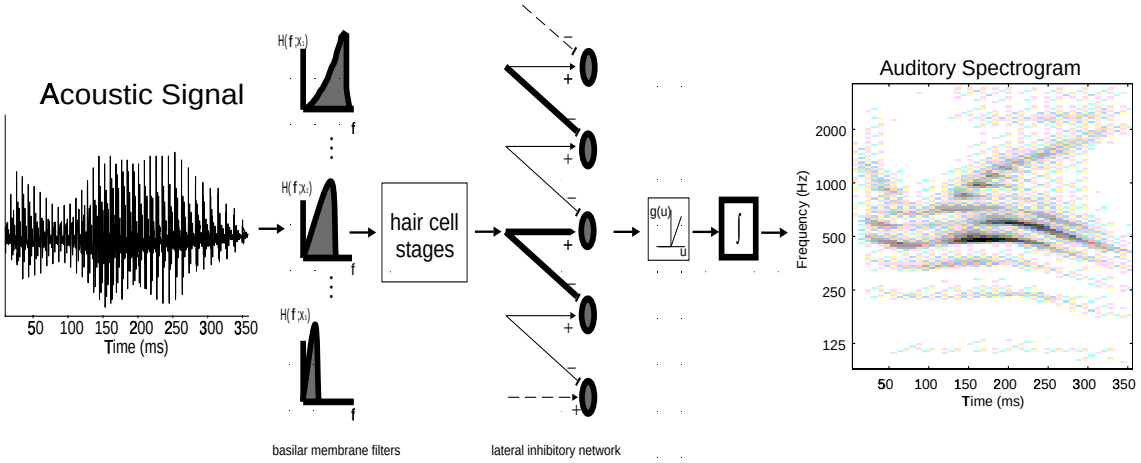

Figure 4: Schematic of processing in the early auditory stages (details in [Yang, Wang, and Shamma, 1992, Wang and Shamma, 1995]). The acoustic signal is analyzed by a bank of constant-Q cochlear-like filters. The output of each filter is then processed by a hair cell model, and then by a lateral inhibitory network. The output at each point is then rectified and integrated to produce the *auditory spectrogram*.

preceding nuclei which likely are involved in a host of other perceptual tasks such as binaural localization and pitch estimation.

Neural responses in AI exhibit a complex and highly varied pattern of spectro-temporal selectivity. For instance, an AI unit is usually tuned to a range of frequencies around a "best frequency" (BF). Within this range, responses change from excitatory to inhibitory in a pattern that varies from one cell to another in its width and asymmetry around the BF [Kowalski, Depireux, and Shamma, 1996]. AI units also exhibit a similar selectivity and variability in their *temporal* responses [Simon, Depireux, and Shamma, 1998]. Some units are best responsive to fast changing spectra, while others are rather sluggish. In addition, AI units often respond selectively to the direction of movement of a spectral peak near their BF. These response properties are summarized by the so-called spectro-temporal response field (STRF), which is a generalization of the classic response areas in auditory physiology [Clarey, Barone, and Imig, 1992] or receptive fields in the retina and visual cortex [De-Valois and De-Valois, 1990]. It represents the spectro-temporal pattern that best excites the cell. Fig.5(a) displays two model STRFs that are sensitive to very different spectro-temporal patterns. On *top*, the STRF is relatively broadly tuned (responds best to ripples of $\Omega = 0.5$ cyc/oct), dynamically agile (responds best to ripples drifting at $\omega = 4$ Hz), and exhibits a downward directional selectivity. In comparison, the STRF on the *bottom* is spectrally narrowly tuned (best $\Omega = 2$ cyc/oct), temporally slow (best $\omega = 2$ Hz), and is upwardly sensitive. STRFs in AI vary along these multiple dimensions, exhibiting spectral bandwidths from 0.5 to 2 octaves (in ferrets and cats), temporal selectivity that ranges from rapid (over 16 Hz) to very slow (under 2 Hz), and directional sensitivities to upwards, bi-directional, and downwards moving spectral energy [deCharms, Blake, and Merzenich, 1998,

Simon, Depireux, and Shamma, 1998, Kowalski, Depireux, and Shamma, 1996].

Therefore, from a functional and computational point of view, AI can be considered a bank of *modulation filters* which analyzes the spectro-temporal modulation rates of its input spectrogram. This view is illustrated in Fig.5(b) where we construct the *scale-rate* plot to summarize the AI responses. The computations consist of two stages. First, the auditory spectrum is analyzed by a bank of STRFs with varying spectro-temporal (Ω-ω) selectivities. Then we estimate the total output power from the STRFs at each Ω-ω combination and plot the results in a *scale-rate* plot as shown in Fig.5(b). For example, a downward moving ripple (Ω=2 cyc/oct, ω= 4 Hz) evokes a fairly circumscribed pattern of *scale-rate* activation centered around the corresponding Ω-ω location (Fig.5(b)). A non-stationary speech spectrogram evokes a series of short-time *scale-rate* plots reflecting the changing modulation content of the utterance (Fig.5(c)).

C. Mathematical formulation of the cortical model

The cortical response (r) produced by the STRF analysis of the auditory spectrogram $y(x,t)$ is defined as:

$$r(x,t;\Omega,\omega) = y(x,t) *_t \cdot_x STRF(x,t;\Omega,\omega) \tag{3}$$

where $*_t \cdot_x$ denotes convolution with respect to t and multiplication with respect to x. Note that the $STRF(x,t;\Omega,\omega)$ is parameterized by its most sensitive spectral and temporal modulations (Ω,ω), and these in turn reflect the bandwidth, dynamics, and orientation of its excitatory and inhibitory fields.

The *scale-rate* plot is derived from the cortical response by integrating the output over the whole spectrum x (5 octaves):

$$SR(t;\Omega,\omega) = \int_0^5 |r(x,t;\Omega,\omega)|dx \tag{4}$$

An intuitive interpretation of the *scale-rate* plot (SR) is that it displays the total amount of modulation that $y(x,t)$ contains at each Ω,ω combination *regardless of its distribution along the spectral axis (x).*

We assume that each upward or downward STRF can be represented by the product of two separate spectral and temporal functions: a response field $RF(x)$ along the frequency (tonotopic) axis; and a temporal impulse response $h_{IR}(t)$. Therefore,

$$STRF(x,t;\Omega,\omega) = RF(x;\Omega,\phi) \cdot h_{IR}(t;\omega,\theta). \tag{5}$$

$RF(x)$ is defined by a symmetric seed function $h_s(x;\Omega)$ and its Hilbert transform

$$RF(x;x_c,\Omega,\phi) = h_s(x-x_c;\Omega)\cos\phi + \hat{h}_s(x-x_c;\Omega)\sin\phi$$

where x_c, Ω and ϕ are respectively the center frequency, density and phase of the most sensitive ripple; its Hilbert transform is

$$\hat{h}_s(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{h_s(x)}{z-x}dz$$

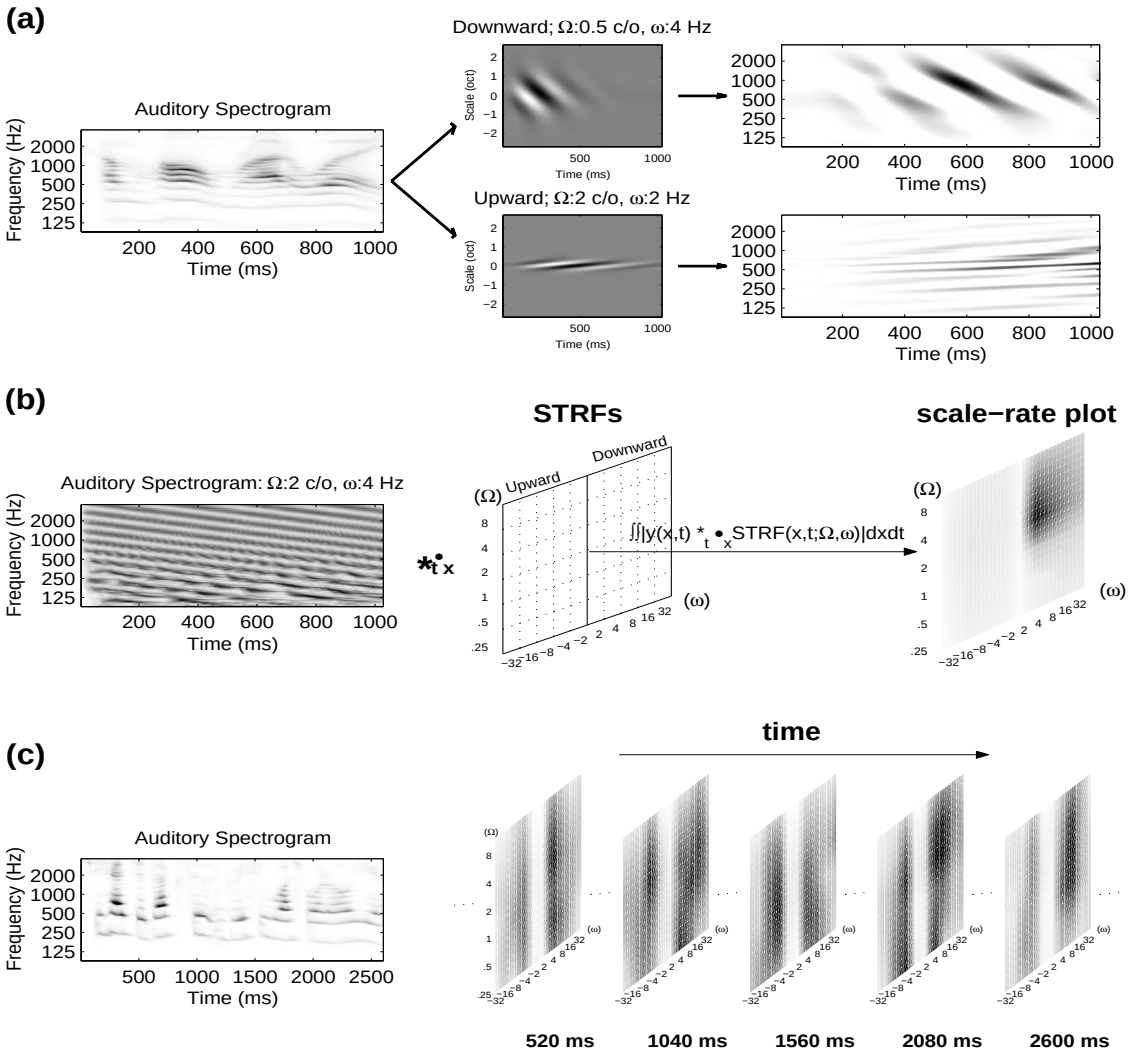

Figure 5: Model of central auditory processing and the *scale-rate plot*. (**a**) Cortical analysis of the auditory spectrogram. The spectrogram of the sentence "Come home right away" is analyzed by a bank of cortical spectro-temporal response fields (STRF). Two model STRFs are shown; in each, black (white) color represents excitatory (inhibitory) regions of the STRF. The panels on the right illustrate the results of processing the original auditory spectrogram through each of the STRFs. Only downward broad spectral transitions survive in the upper panel; the opposite spectro-temporal features are seen in the lower panel. (**b**) Measuring the output of cortical STRFs. The auditory spectrogram (*left*) is projected to a bank of STRFs with all Ω, ω parameters represented *middle*. The ripple spectrogram activates maximally the STRF that matches its outline best (i.e., the STRF at 2 cyc/oct and 4 Hz). The output from all STRF's is summarized in the scale-rate plot (*right*), which shows a peak at $\Omega = 2$ cyc/oct and $\omega = 4$ Hz. (**c**) The scale-rate plots vary as a function of time reflecting the changing ripple content of a speech spectrogram

Here we choose a Gabor-like function to approximate $h_s(\cdot)$:

$$h_s(x) = (1 - x^2)e^{-\frac{x^2}{2}}$$

$$h_s(x;\Omega) = \Omega h_s(\Omega x)$$

Similarly, the temporal impulse response can be expressed as

$$h_{IR}(t;\omega,\theta) = h_t(t;\omega)\cos\theta + \hat{h}_t(t;\omega)\sin\theta$$

where ω is the most sensitive ripple velocity. $h_t(\cdot)$ is modelled by one gamma probability density function

$$h_t(t) = t^3 e^{-4t}\cos(2\pi t)$$

$$h_t(t;\omega) = \omega h_t(\omega t)$$

In many instances, we will need to consider the *average* scale-rate plot of a sentence or even an entire corpus of speech. This is simply defined as

$$SR_{avg}(\Omega,\omega) = \frac{1}{T}\int_T |SR(t;\Omega,\omega)|dt,$$

where T denotes the entire interval over which SR is averaged.

D. Spectro-temporal MTF of the auditory model

The MTF of the full auditory model is shown in Figure 6(a). It is measured by presenting single ripples of all Ω,ω combinations and noting for each output at the corresponding $SR(\Omega,\omega)$. The model responses capture the main trends seen in the threshold MTF illustrated earlier in Fig.3. Specifically, the model exhibits lowpass MTFs with very similar spectral and temporal rate cutoffs. The origin of these response characteristics is the effectively narrow bandwidths used to compute the auditory spectrogram. The cochlear filters have a typical critical-band gammatone shape; However, the lateral inhibition stage in the model effectively narrows the filter bandwidths and slows down its dynamics (see [Wang and Shamma, 1995] for details).

The MTF loss of sensitivity at high ripple velocities is evident in the auditory spectrograms (Fig.6(b)) of a 4 cyc/oct ripple at $\omega = 4$, 8, 16 and 32 Hz . In all panels, the temporal modulations are poorly represented at the lowest BFs where the auditory filters are narrowest and dynamically slowest. At higher modulation rates, the disruption spreads towards higher, hence reducing the corresponding SR outputs.

The decrease in output at high density ripples is due to the finite bandwidths of the cochlear filters. As ripple peaks become closely spaced, they are less resolved by the cochlear filters, and their amplitudes decrease as illustrated in Fig.6(c). Based on these arguments, it is evident that the upper limits of the temporal and spectral modulation rates are inversely related through the effective bandwidths of the cochlear filters.

Finally, the model responses exhibit a slight asymmetry with respect to ripple directions at high rates (e.g., responses at -32 Hz are smaller than at 32 Hz). Experimental data, however, do not replicate such a preference. The model asymmetry is due to the staggered cochlear filter group delays which give rise to a (basilar membrane) travelling wave towards the lower frequency channels, i.e., in the downward direction. These delays disrupt the flow of responses to upward ripple, but less so to downward ripples.

IV: APPLICATION TO SPEECH INTELLIGIBILITY

Auditory spectrograms of speech are rich in spectro-temporal modulations which are important in preserving its intelligibility. Numerous tests have estimated a critical range of temporal modulations in speech at between 2 and 8 Hz [Greenberg, Hollenback, and Ellis, 1996]. In fact, filtering out temporal modulations outside of this range has proven to be an effective strategy for combating the deleterious effects of noise and reverberations in real world speech signals [Hermansky and Morgan, 1994, Greenberg and Kingsbury, 1997].

However, not all spectro-temporal modulations of speech (or other environmental sounds) are equally perceived by humans. To determine the most perceptually salient range of speech modulations, we computed the long-term average scale-rate plot of 380 spoken sentences. These sentences were extracted from a subset of TIMIT corpus[3] (the training portion of the New England dialect region) which contains a total of 24 male speakers (240 sentences) and 14 female speakers (140 sentences). Figure 7 illustrates that spectro-temporal ripples bounded by the range between 4-8 Hz and <4 cyc/oct are the critical perceptible modulations in speech. It is important to note that in marking this range, we consider *both* the spectral and temporal dimensions simultaneously. For instance, temporal modulations at 4-8 Hz are not important for densely rippled spectra (e.g., > 4 cyc/oct).

Speech often suffers significant loss of intelligibility in noisy or reverberant environments. Presumably this is partly because these conditions disrupt the modulations of normal clean speech. Therefore, the scale-rate plot could serve as a useful indicator of this disruption by providing a sensitive spectro-temporal representation from which an "intelligibility index" could be derived. This is analogous to the way traditional critical-band spectra and single-tone temporal MTFs have been utilized to derive the classical *articulation index* [Kryter, 1962] and *speech transmission index* [Houtgast, Steeneken, and Plomp, 1980].

Figures 7(b-c) illustrate the effects of noise and reverberations on the scale-rate plots of clean speech. Fig.7(b) shows a series of average scale-rate plots ($SR_{avg}(\Omega, \omega)$) for a sentence contaminated with increasing levels of white noise (decreasing S/N). Another series of similar plots for speech with increasing reverberation delays is shown in Fig.7(c)[4]. In both cases, a suitably defined measure of the similarity between the clean and distorted plots could serve as an indicator of the noise perceived by the listener as described next.

[3] A more extensive description of corpus design, collection, and transcription can be found in the printed documentation from National Institute of Standards and Technology (NIST# PB91-100354).

[4] The reverberation model is an exponentially decaying function, $m(F) = [1 + (2\pi F\tau)^2]^{-1/2}$, applied to the envelope of each channel in the spectrogram, where F is the temporal modulation frequency of the envelope [Houtgast, Steeneken, and Plomp, 1980].

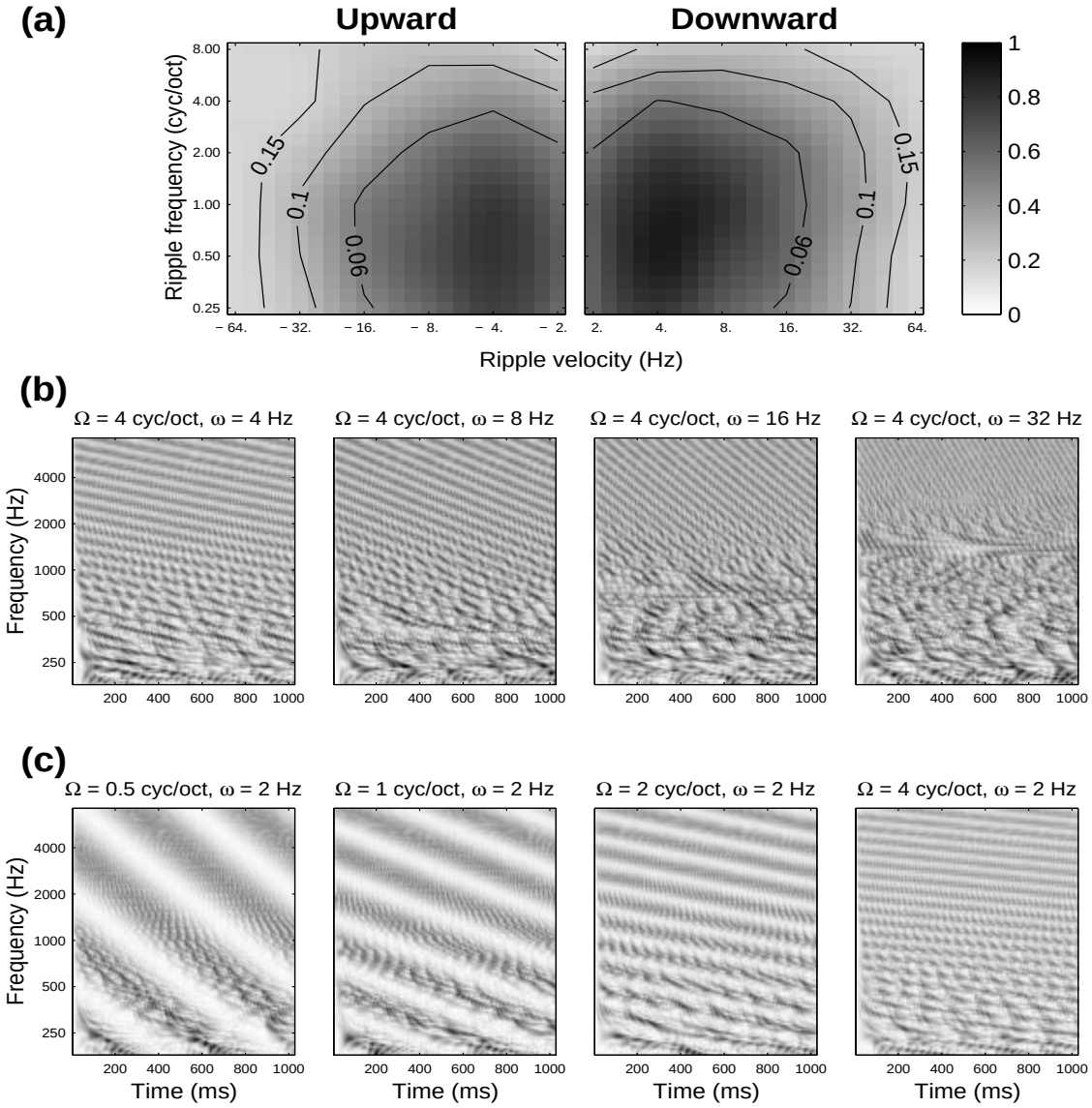

Figure 6: Spectro-temporal MTFs of the full auditory model. (**a**) The model MTF shown in arbitrary linear gray scale (bar on the right). The contour surfaces depict thresholds predicted by the model; They are derived by normalizing the minimum perceptual threshold in Fig.4 (=0.035) by the model MTF. (**b**) Origin of the temporal lowpass shape in the MTF. With increasing ripple velocity, the low BF regions fail to follow rapid modulations and hence become progressively more distorted, thus causing the decrease in the corresponding STRF outputs. (**c**) Origin of the spectral lowpass shape in the MTF. Cochlear filters gradually fail to resolve higher density ripples causing a decrease in overall output.

A. Deriving an intelligibility measure from the scale-rate plots

Assume that the scale-rate plots of a clean speech utterance $SR_c(t;\Omega,\omega)$ and its noisy version $SR_n(t;\Omega,\omega)$ are available, then the similarity (or correlation) between any corresponding (Ω,ω) channels is defined as [Duda and Hart, 1973]:

$$\rho_0(\Omega,\omega) = \frac{< SR_c(t;\Omega,\omega) - \mu_{SR_c(t;\Omega,\omega)}, SR_n(t;\Omega,\omega) - \mu_{SR_c(t;\Omega,\omega)} >}{\parallel SR_c(t;\Omega,\omega) - \mu_{SR_c(t;\Omega,\omega)} \parallel \cdot \parallel SR_n(t;\Omega,\omega) - \mu_{SR_c(t;\Omega,\omega)} \parallel} \tag{6}$$

where μ_{SR_c} is the mean of the random variable SR_c and the inner product and the induced norm are defined as : (*Note that dependence on* Ω,ω, *and* t *in all quantities below is suppressed to simplify notation*)

$$< SR_c - \mu_{SR_c}, SR_n - \mu_{SR_c} >= \int_T (SR_c(t) - \mu_{SR_c}) \cdot (SR_n(t) - \mu_{SR_c}) dt$$

and

$$\parallel SR - \mu \parallel = \sqrt{< SR - \mu, SR - \mu >}$$

This similarity measure ρ_0 compares the $SR(\cdot)$ of the clean and noisy signals *frame-by-frame*, and not simply through the time-averaged plot SR_{avg}. Consequently, such a measure is only useful if *both* clean and noisy samples of the *same* sentence are available.

A slightly modified measure can be defined which requires only knowledge of the mean and variance of the clean speech scale-rate plot (SR_c in Fig.7(a)). It is used in situations where only noisy speech samples are provided, or if clean and noisy samples are of different utterances. We assume that the effect of added noise and reverberations at a given channel can be modeled as a change in the mean and variances of the random variables SR_c :

$$SR_n(t;\Omega,\omega) = A_{\Omega\omega} \cdot SR_c(t;\Omega,\omega) + C_{\Omega\omega} \tag{7}$$

where $A_{\Omega\omega}$ and $C_{\Omega\omega}$ are measurable from the long-term average scale-rate output of the clean and noisy speech samples as:

$$\mu_{SR_n} = A \cdot \mu_{SR_c} + C \tag{8}$$

$$\sigma_{SR_n} = A \cdot \sigma_{SR_c} \tag{9}$$

where μ_{SR} and σ_{SR} are the mean and standard deviation of the random variable SR at each channel (Ω,ω).

For discrete time interpretation, substituting Eqs. (7), (8), (9) into Eq. (6) and noting that $\sum_n (SR_c[n] - \mu_{SR_c}) = 0$, the numerator of Equation 6 can be written as :

$$\begin{aligned} & \sum_n (SR_c[n] - \mu_{SR_c})(SR_n[n] - \mu_{SR_c}) \\ = & \sum_n (SR_c[n] - \mu_{SR_c})(A \cdot SR_c[n] - \mu_{SR_c} + C) \\ = & A \cdot \sum_n (SR_c[n] - \mu_{SR_c})^2 \end{aligned} \tag{10}$$

and the denominator can be simplified to :

$$\begin{aligned} & \sqrt{(\sum_n (SR_c[n]-\mu_{SR_c})^2)(\sum_n (SR_n[n]-\mu_{SR_c})^2)} \\ = & \sqrt{(\sum_n (SR_c[n]-\mu_{SR_c})^2)(\sum_n [A\cdot(SR_c[n]-\mu_{SR_c})+(A-1)\mu_{SR_c}+C]^2)} \\ = & \sqrt{(\sum_n (SR_c[n]-\mu_{SR_c})^2)(A^2\sum_n (SR_c[n]-\mu_{SR_c})^2+N[(A-1)\mu_{SR_c}+C]^2)} \end{aligned} \quad (11)$$

where N is the total number of samples along n axis. Therefore, ρ_0 can be simplified to ρ_s:

$$\begin{aligned} \rho_s(\Omega,\omega) &= \sqrt{\frac{1}{1+(\frac{(A_{\Omega\omega}-1)\cdot\mu_{SR_c(\Omega,\omega)}+C_{\Omega\omega}}{A_{\Omega\omega}\cdot\sigma_{SR_c(\Omega,\omega)}})^2}} \\ &= \sqrt{\frac{1}{1+(\frac{\mu_{SR_n(\Omega,\omega)}-\mu_{SR_c(\Omega,\omega)}}{\sigma_{SR_n(\Omega,\omega)}})^2}} \end{aligned} \quad (12)$$

A final indicator ρ of the total similarity between clean and noisy samples is derived as the average (over all Ω and ω of all $\rho_o(\Omega,\omega)$ or $\rho_s(\Omega,\omega)$):

$$\rho = \frac{1}{N_\Omega \cdot N_\omega}\sum_\Omega\sum_\omega \rho_0(\Omega,\omega) \quad (13)$$

or

$$\rho = \frac{1}{N_\Omega \cdot N_\omega}\sum_\Omega\sum_\omega \rho_s(\Omega,\omega) \quad (14)$$

where N_Ω, N_ω are the number of channels along Ω and ω axes.

The validity of the assumptions leading to the derivation of the simplified ρ (Eq. 14) is demonstrated by the approximate correspondence between the two ρ measures (Eqs. 13 and 14) for the noisy speech samples shown in Figs.7(b-c). It is important to note here that the proposed measures are defined and computed irrespective of the nature (or model) of the distorting process, i.e., ρ reflects the perceptual change the same way whether it is caused by broadband noise, reverberations, or any other process. Therefore, ρ serves as a *spectro-temporal modulation index (STMI)*, a measure of the total change in the modulations away from those of clean speech.

B. Assessing speech intelligibility with STMI (ρ)

We discuss the use of the STMI in two situations: (1) determining the quality of a transmission medium (e.g. a telephone channel or an auditorium); (2) assessing the intelligibility of a given noisy speech sample.

In the first case, clean speech is transmitted through a noisy or reverberant channel (or recorded in an auditorium). The clean and transmitted versions of the *same* sentence are

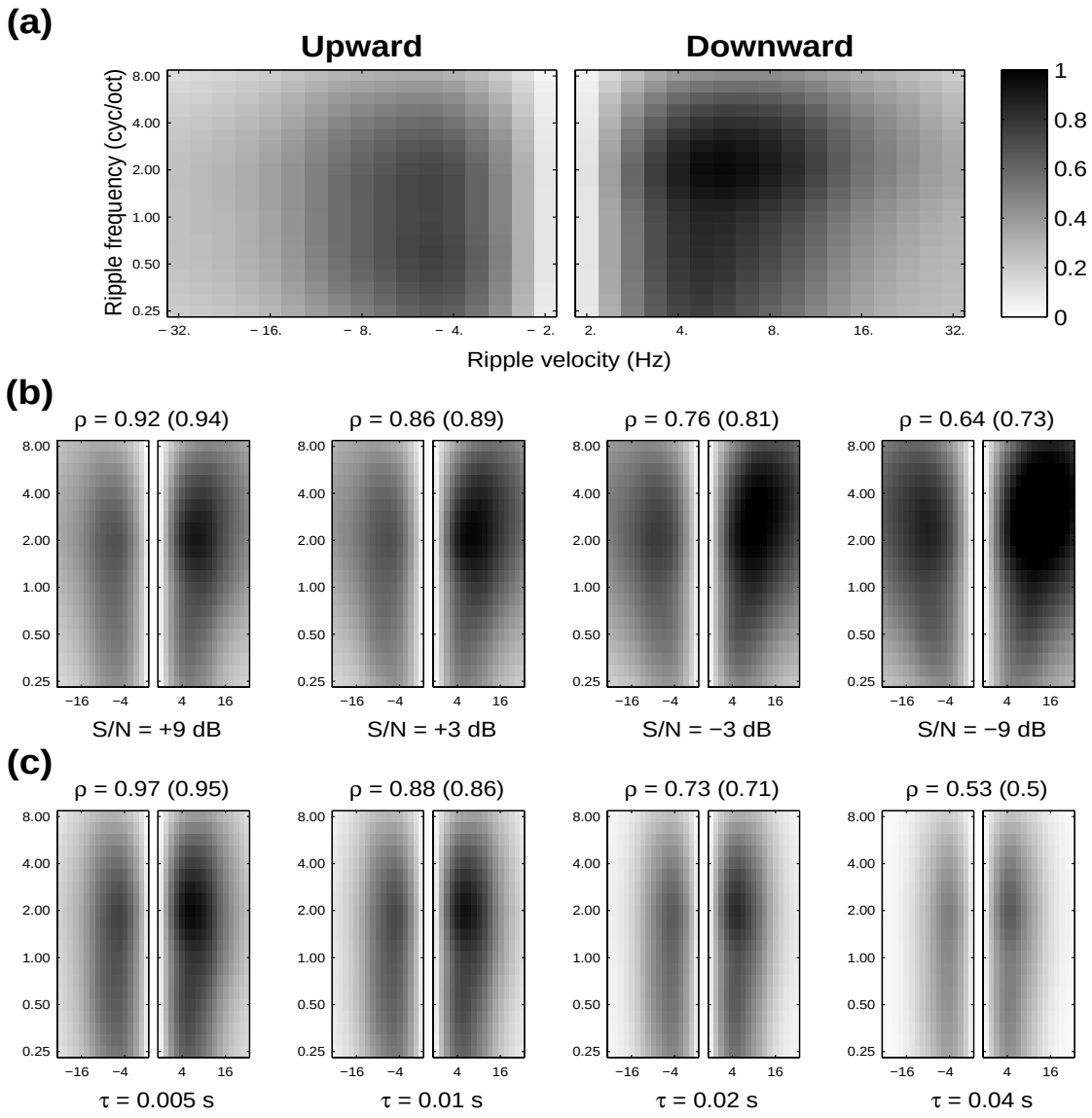

Figure 7: Spectro-temporal modulations in speech. (**a**) The average (mean) scale-rate plot of modulations in speech (μ_{SR_c}). On average, the strongest modulations are downward, and in the 4-8 Hz range and under 4 cyc/oct. (**b**) Average scale-rate plots (SR_{avg}) of speech contaminated by stationary white noise at different S/N ratios. Above each panel is a measure of the similarity (ρ) between the *frame-by-frame* scale-rate plots of the clean and noisy speech signals; The number in the parenthesis is a simplified version of ρ that uses averaged SR only (see text for details on both measures). As the S/N ratio decreases, ρ values decrease reflecting distortions due to the noise. (**c**) The effect of reverberation delays: ρ decreases gradually with increasing reverberation delays (τ).

compared by computing ρ (Eq. 13). The resulting values predict the quality of the transmission medium. Informal experiments in our laboratory indicate that ρ values near 0.75 reflect marginally intelligible speech, corresponding to S/N ratios of about -3 dB (Fig.7(b)) or reverberation delays of 20 ms (Fig.7(c)).

In the second situation, the goal is to estimate the intelligibility of noisy speech samples recorded in different S/N conditions. No clean samples of the noisy speech are available. Instead, we rely on the long-term average of the clean speech in Fig.7(a) to provide the reference against which to measure the effects of the noise. To illustrate this procedure, we have utilized data supplied to us by the Southwest Research Institute based on speech intelligibility tests in different S/N conditions conducted with 5 subjects. Each test consisted of the following conditions and test materials:

(1) A specific S/N condition.

(2) A total of eight non-sense sentences.

(3) Each sentence consisted of 5 randomly chosen monosyllabic words.

(4) Each word consisted of 3 phonemes with approximately balanced presentation of vowels vs. consonants (typically 40% vs. 60% over the entire set of sentences). Examples of words used are: GAB, BAR, WHET, BUG, LOT.

In a given test, each subject was presented the 5-word sentences and asked to enter the phonemes heard in each word, including the option of "not certain" at any position in a word. The subjects also had the option of hearing the sentences again before responding. Numerous measures of the responses were compiled including the correct percentage of responses for vowels and consonants combined and separately, total number of 100% correct words, and the average correct phonemes all subjects entered correctly. For the purposes of this paper, we focus on the correct percentage of phonemes perceived as a function of the S/N ratio of the test. A more elaborate analysis of the data is underway in which ρ is computed separately over specific phonemes (e.g., vowels, consonants, or restricted vowel types), and compared to the corresponding identification results in the tests.

Figure 8 illustrates the correspondence between the ρ (Eq. 14) and the percentage of correct phonemes at each of the 7 S/N conditions tested. The ρ evidently provides a fair average measure of the integrity of the phoneme percepts.

V: Discussion

We have reported new measurements of the spectro-temporal MTFs using moving ripple spectra. The MTFs exhibit a lowpass function with respect to both dimensions, with 50% bandwidths of about 16 Hz and 2 cyc/oct (Fig.3). We have also formulated a computational auditory model that exhibits spectro-temporal MTFs consistent with the salient trends in the data. The model was used to demonstrate the potential relevance of these MTFs to the assessment of speech intelligibility.

A. Relation to psychoacoustic modulation transfer functions

Modulation transfer functions have been measured with different types of acoustic stimuli. For purely spectrally modulated spectra, two sets of largely comparable measurements

are available [Green, 1986, Hillier, 1991], and both are in agreement with the spectral transfer functions in Fig.3(a) at low temporal modulations (< 32 Hz). For *temporal* modulation transfer functions, the data using modulated flat noise [Viemeister, 1979] or rippled noise [Yost and Moore, 1987, van Zanten and Senten, 1983] give quite different results. With flat noise, subjects detect modulation rates well over 64 Hz. By contrast, modulation of rippled noise is not detectable beyond about 10 Hz. Yost and Moore [Yost and Moore, 1987] discussed and discounted several hypotheses that could account for this difference. More recently, Yost (personal communication) suggested that a possible reason for this disparity is that temporal modulations of a flat noise yield amplitude modulations that are in-phase over the whole spectrum. By comparison, modulations of the linear ripple produce out-of-phase amplitude modulations in different parts of the spectrum, and hence will cancel out if they are centrally integrated.

Our data indicate that Viemeister's temporal MTFs are valid not just for flat spectra, but for any spectra composed of ripples up to 2 cyc/oct. The data also suggest that the low pass form of the MTFs persist at higher ripple densities (> 2 cyc/oct) (Fig.3), but with significant overall loss of sensitivity presumably due to the higher thresholds associated with high density ripples in general [Green, 1986, Hillier, 1991]. This finding argues for the separability of the temporal and spectral dimensions of the MTFs, a conjecture that is strongly supported by the SVD analysis.

The stimuli of [Yost and Moore, 1987, van Zanten and Senten, 1983] are fundamentally different in that they are not pure ripples (on the tonotopic axis), but rather a collection of ripples of many densities. Consequently, to predict their detection thresholds, we need to specify further detection procedures for arbitrary spectra, which are beyond the scope of the present model formulation. However, it is intuitively possible to see that our results are at least consistent with the notion that ripple-noise thresholds are high because they contain high density profiles ($\Omega > 4$ cyc/oct) which are poorly perceived in the MTFs (Fig.3). Our model further suggests that the locus of spectral integration that gives rise to this phenomenon is not necessarily central, but is instead the limited resolution of the cochlear filters.

B. Spectro-temporal modulations and speech intelligibility

The experiments reported in section IV illustrate the potential utility of spectro-temporal MTFs in quantifying speech intelligibility. Their promise seems to derive from their integration of spectral and temporal factors into one measure. In this sense, they can be viewed as closely related to, or in fact combining two widely used intelligibility measures: *articulation index* [Kryter, 1962] and *speech transmission index* [Houtgast, Steeneken, and Plomp, 1980]. These two measures represent in some sense extreme versions of the STMI. The *articulation index* captures effectively the distortions due to stationary broadband noise and hence is purely spectral. The *speech transmission index* was specifically designed to deal with problems arising from severe reverberations and is therefore mostly temporal. We have demonstrated in Fig.7 that the *scale-rate* plot (and its associated measure ρ) are sensitive to both kinds of distortions, and that they seem to reflect sensibly the perceptual degradation

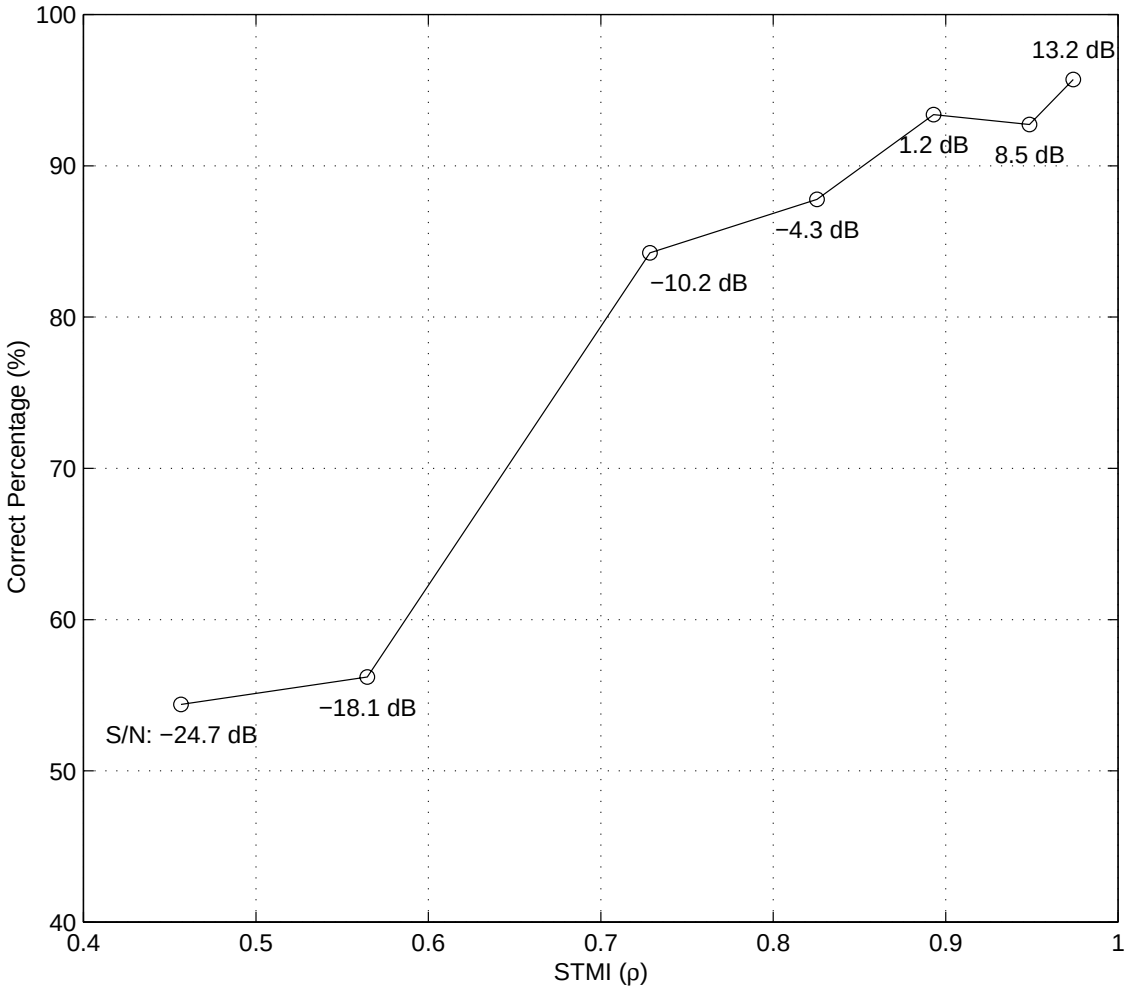

Figure 8: The spectro-temporal modulation index ρ plotted against the percentage of correct phonemes (as reported by 5 subjects) in 7 different S/N conditions.

of the speech signal (Fig.8). Therefore, the STMI has potentially the advantage of being useful in a wide range of noisy environments and applications. Moreover, it can be used for *frame-by-frame* comparisons and with *long-term* averages as illustrated in section IV. Clearly, much more experimental work is needed to "calibrate" the numerical values of STMI against human speech perception under controlled noisy situations, and against accepted estimates of the *articulation index* and the *speech transmission index*.

C. Relation to vision spatio-temporal MTFs

Visual spatio-temporal MTFs are usually measured with sinusoidally modulated gratings, with various orientations and drifting velocities [Dong and Atick, 1995, Kelly, 1961]. These measurements are analogous to our spectro-temporal MTFs if we consider the spatial axis of the retina as analogous to the tonotopic axis of the auditory system. Visual and auditory

MTFs are generally similar in that they both exhibit an overall lowpass function in both dimensions. There are, however, three important details to note about the data:

(1) Both visual and auditory MTFs exhibit a small but consistent highpass edge at the lowest modulation rates, giving the MTF more of a bandpass shape (see Fig.3(c)).

(2) Temporal cut-off rates of visual and auditory MTFs are quite comparable, contrary to the common assumption that auditory processes are generally faster.

(3) Unlike auditory MTFs, visual MTFs seem to be inseparable, i.e., they cannot be reduced to a product of purely temporal and spatial MTFs [Dong and Atick, 1995]. This conclusion may be revised based on the criteria and tests one accepts for separability. To first order, however, it is clear that our auditory spectro-temporal MTFs can be derived from a simple product of [Viemeister, 1979] temporal MTFs and [Green, 1986] spectral MTFs (assuming that the highpass shape of the latter curve is ignored).

D. Further refinements and considerations of the auditory model

The auditory model described here (section III) combines a simplified version of an early auditory model, and a computational module that captures the main features so far observed in auditory cortical responses. It is evident that most of the MTF features are due to the early auditory processing stages (e.g., bandwidths of cochlear filters, and the lateral inhibitory network). The primary purpose of the cortical module is to derive an estimate of the distribution of power in the spectrogram modulations. A major simplification in our analysis and displays is the integration of the spectrogram over the entire tonotopic axis x (Eq. 4). This allows us to generate the two dimensional scale-rate plots, and to ignore the contribution of different frequency ranges to the final displays. Clearly, this variable must be considered in future work in ways that reflect the specific applications. For instance, the effect of noise should be weighted more heavily in the intermediate frequency ranges (near 1 kHz) where hearing thresholds are lowest and speech spectra are concentrated. Another useful addition to the model is a detection criteria that will enable us to predict sensitivity to changes in arbitrary spectrograms, and to derive MTFs for arbitrarily complex spectra and manipulations such as Yost's ripple noise.

Acknowledgement

This work is supported by a contract with the Southwest Research Institute, and partially by the Office of Naval Research through a MURI grant (Center for Auditory and Acoustic Research). We are grateful to Dr. Brian Zook (at Southwest Research Institute) for supplying the speech intelligibility data.

References

[Amagai et al., 1999] Amagai, S., Dooling, R., Shamma, S., Kidd, T., and Lohr, B. (1999). "Perception of rippled spectra in the parakeet and zebra finches," J. Acoust. Soc. Am. (in press).

[Arai et al., 1996] Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (1996). "Intelligibility of speech with filtered time trajectories of spectral envelopes," Proc. ICSLP pp. 2490–2492.

[Clarey, Barone, and Imig, 1992] Clarey, J., Barone, P., and Imig, T. (1992). "Physiology of thalamus and cortex," in *The Mammalian Auditory Pathway: Neurophysiology*, edited by D. Webster, A. Popper, and R. Fay (Springer Verlag, New York), pp. 232–334.

[De-Valois and De-Valois, 1990] De-Valois, R. and De-Valois, K. (1990). *Spatial Vision* (Oxford University Press, New York).

[deCharms, Blake, and Merzenich, 1998] deCharms, R., Blake, D., and Merzenich, M. (1998). "Optimizing sound features for cortical neurons," Science **280**, 1439.

[Dong and Atick, 1995] Dong, D. and Atick, J. (1995). "Statistics of natural time-varying images," Network: Computation in Neural Systems **6**, 345–358.

[Drullman, Festen, and Plomp, 1994] Drullman, R., Festen, J., and Plomp, R. (1994). "Effect of envelope smearing on speech reception," J. Acoust. Soc. Am. **95(2)**, 1053–1064.

[Duda and Hart, 1973] Duda, R. and Hart, P. (1973). *Pattern Classification* (Wiley-Interscience, New York).

[Green, 1986] Green, D. (1986). "'Frequency' and the detection of spectral shape change," in *Auditory Frequency Selectivity*, edited by B. C. J. Moore and R. Patterson (Plenum Press, Cambridge), pp. 351–359.

[Greenberg, Hollenback, and Ellis, 1996] Greenberg, S., Hollenback, J., and Ellis, D. (1996). "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *ICSLP-96 Proc. 4th Int. Conf. Spoken Lang.* (IEEE, New York), pp. S32–35.

[Greenberg and Kingsbury, 1997] Greenberg, S. and Kingsbury, B. (1997). "The modulation spectrogram: In pursuit of an invariant representation of speech," ICASSP-97 pp. 1647–1650.

[Haykin, 1996] Haykin, S. (1996). *Adaptive Filter Theory* (Prentice Hall, New Jersey).

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). "RASTA processing of speech," IEEE Trans. Speech and Audio Proc. **2(4)**, 578–589.

[Hillier, 1991] Hillier, D. (1991), "Auditory processing of sinusoidal spectral envelopes," Ph.D. thesis, The Washington University and Severn Institute.

[Houtgast, Steeneken, and Plomp, 1980] Houtgast, T., Steeneken, H., and Plomp, R. (1980). "Predicting speech intelligibility in rooms from the Modulation transfer function: General room acoustics," Acoustica **46**, 60–72.

[Kelly, 1961] Kelly, D. H. (1961). "Visual responses to time-dependent stimuli," J. Opt. Soc. Am. **51**, 422–429.

[Kowalski, Depireux, and Shamma, 1996] Kowalski, N., Depireux, D., and Shamma, S. (1996). "Analysis of dynamic spectra in ferret primary auditory cortex: Characteristics of single unit responses to moving ripple spectra," J. Neurophysiology **76(5)**, 3503–3523.

[Kryter, 1962] Kryter, K. (1962). "Methods for the calculation and Use of the articulation index," J. Acoust. Soc. Am. **34(11)**, 1689–2147.

[Levitt, 1971] Levitt, W. (1971). "T Transformed Up-Down Methods in Psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

[Shamma, 1988] Shamma, S. (1988). "The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives," Journal of Phonetics **16**, 77–91.

[Shamma et al., 1986] Shamma, S., Chadwick, R., Wilbur, J., Morrish, K., and Rinzel, J. (1986). "A biophysical model of cochlear processing: Intensity dependence of pure tone responses ," J. Acoust. Soc. Am. **80(1)**, 133–145.

[Shannon et al., 1995] Shannon, R., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science 1995 **270**, 303–304.

[Simon, Depireux, and Shamma, 1998] Simon, J., Depireux, D. A., and Shamma, S. A. (1998). "Representation of complex spectra in auditory cortex," in *Psychophysical and Physiological advances in hearing. Proceedings of the 11th international symposium on hearing*, edited by A. R. Palmer, A. Ress, A. Q. Summerfield, and R. Meddis (Whurr Publishers, London), pp. 513–520.

[van Zanten and Senten, 1983] van Zanten, G. and Senten, C. (1983). "Spectro-temporal modulation transfer functions (STMTF) for various types of temporal modulation and a peak distance of 200 Hz," J. Acoust. Soc. Am. **74(1)**, 52–62.

[Viemeister, 1979] Viemeister, N. (1979). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66(5)**, 1364–1380.

[Wang and Shamma, 1995] Wang, K. and Shamma, S. (1995). "Representation of spectral profiles in primary auditory cortex," IEEE Transactions on Speech and Audio Processing **3**, 382–395.

[Wang and Shamma, 1994] Wang, K. and Shamma, S. A. (1994). "Self-normalization and noise-robustness in early auditory representations," IEEE Transactions on Speech and Audio Processing pp. 421–435.

[Yang, Wang, and Shamma, 1992] Yang, X., Wang, K., and Shamma, S. A. (1992). "Auditory representations of acoustic signals," IEEE Transactions on Information Theory, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis **38**, 824–839.

[Yost and Moore, 1987] Yost, W. and Moore, M. (1987). "Temporal changes in a complex spectral profile," J. Acoust. Soc. Am. **81**, 1896–1905.