ELSEVIER

Speech Communication 41 (2003) 331–348

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A spectro-temporal modulation index (STMI) for assessment of speech intelligibility

Mounya Elhilali, Taishih Chi, Shihab A. Shamma *

Department of Electrical and Computer Engineering, Institute for Systems Research (ISR), A.V. Williams Building (115), Room # 2202, University of Maryland, College Park, MD 20742, USA

Received 27 November 2001; received in revised form 27 March 2002; accepted 16 July 2002

Abstract

We present a biologically motivated method for assessing the intelligibility of speech recorded or transmitted under various types of distortions. The method employs an auditory model to analyze the effects of noise, reverberations, and other distortions on the *joint* spectro-temporal modulations present in speech, and on the ability of a channel to transmit these modulations. The effects are summarized by a spectro-temporal modulation index (STMI). The index is validated by comparing its predictions to those of the classical STI and to error rates reported by human subjects listening to speech contaminated with combined noise and reverberation. We further demonstrate that the STMI can handle difficult and nonlinear distortions such as phase-jitter and shifts, to which the STI is not sensitive.

Résumé

Nous présentons une approche inspirée par la biologie du système auditif humain, qui prédit l'intelligibilité d'enregistrements directes de paroles ou après transmissions sous différentes conditions de bruit propre, réverbérations, et autres déformations. La méthode est basée sur un modèle auditif qui analyse les effets du bruit sur les modulations conjointes de temps et fréquences, présentes dans la parole. Par ailleurs, cette méthode analyse la capacité d'un canal à transmettre fidèlement ces modulations. Les effets sur les modulations sont convertis en un indice des modulations spectro-temporelles, appelé STMI. La validité de cet indice est établie en comparant ses prédictions à celles du STI classique; ainsi qu'aux résultats expérimentaux des taux d'erreurs de sujets humains qui écoutent de la parole contaminée par des combinaisons de bruit propre et de réverbération. Nous démontrons également que le STMI est capable de manipuler des conditions encore plus sévères, comme les déformations non-linéaires, tels les décalages et autres instabilités des phases; conditions auxquelles le STI classique s'avère être insensible.

Keywords: Modulation transfer function; Spectro-temporal modulations; Speech intelligibility; STMI

1. Introduction

The articulation index (AI) and *speech transmission index* (STI) are the most widely used predictors of speech intelligibility (ANSI, 1969;

* Corresponding author. Tel.: +1 301 405 6842.
E-mail address: sas@eng.umd.edu (S.A. Shamma).

0167-6393/02/$ - see front matter © 2002 Published by Elsevier B.V.
doi:10.1016/S0167-6393(02)00134-6

Houtgast and Steeneken, 1980; Kryter, 1962), and have proven to be extremely valuable in a wide range of applications ranging from architectural designs to vocoder characterization (Bradley, 1986; Houtgast and Steeneken, 1980; Houtgast and Steeneken, 1985; Steeneken and Houtgast, 1979). In an effort to understand the underlying biological mechanisms that render such measures meaningful, and how noise in general compromises the perception of speech and other complex dynamic signals, we have developed earlier a computational model to represent spectral and temporal modulations in the auditory system (Chi et al., 1999). The model is grounded on extensive neurophysiological data from mammalian auditory cortex and earlier stages of auditory processing (Kowalski et al., 1996; Depireux et al., 2001), and on psychoacoustical measurements of human spectro-temporal modulation transfer functions (MTF) (Chi et al., 1999).

Based on the premise that faithful representation of these modulations is critical for perception (Drullman et al., 1994; Dau et al., 1996), we derived an intelligibility index, the spectro-temporal modulation index (STMI), which quantifies the degradation in the encoding of spectral and temporal modulations due to noise regardless of its exact nature. The STI, as we shall discuss below, can best describe the effects of spectro-temporal distortions that are *separable* along these two dimensions, e.g. static noise (purely spectral) or reverberation (mostly temporal). The STMI is an elaboration on the STI in that it incorporates explicitly the *joint* spectro-temporal dimensions of the speech signal. As such, we expect it to be consistent with the STI in its estimates of speech intelligibility in noise and reverberations, but also be applicable to cases of *joint* (or inseparable) spectro-temporal distortions that are unsuitable for STI measurements (as with certain kinds of channel phase-distortions) or severely nonlinear distortions of the speech signal due to channel phase-jitter and amplitude clipping. Finally, like the STI, the STMI effectively applies specific weighting functions on the signal spectrum and its modulations; these assumptions arise naturally from the properties of the auditory model and hence can now be ascribed a biological interpretation.

In an earlier report (Chi et al., 1999), we presented a simplified derivation of the STMI and its application to classic distortions such as white stationary noise or reverberation. Here, we elaborate on the derivation, validation, and application of the STMI in combined stationary noise and reverberation conditions. We also demonstrate STMI performance for noise conditions under which current formulations of the STI would fail such as phase-jitter and joint spectro-temporal distortions. Finally, we shall discuss how the STMI can be used for intelligibility assessment of both transmission channels and in the case of noisy recordings (where there is no access to the channel).

We shall start by giving a brief review of the auditory model and its parameters (Section 2), then define the STMI and compare it to the STI and to results of intelligibility tests with human subjects under various noise conditions (Section 3). Finally, we discuss the performance of the STMI in more difficult noise conditions under which the STI fails and the fundamental differences and similarities between these indices.

2. Methods

Conceptually, the STMI *is a measure of speech integrity as viewed by a model of the auditory system*. In this section, we review briefly the structure of the auditory model employed in this study. We then define the intelligibility index, and describe two practical modes for its application. A more complete description of this model is available in (Chi et al., 1999).

2.1. The auditory model

The computational auditory model is based on neurophysiological, biophysical, and psychoacoustical investigations at various stages of the auditory system (see Lyon and Shamma, 1996; Wang and Shamma, 1994; Yang et al., 1992 for a detailed description). It consists of two basic stages.

- An early stage, which models the transformation of the acoustic signal into an internal neural representation referred to as an *auditory spectrogram*.
- A central stage, which analyzes the spectrogram to estimate the content of its spectral and temporal modulations using a bank of modulation selective filters mimicking those described in the mammalian primary auditory cortex (Chi et al., 1999; Wang and Shamma, 1995).

2.1.1. The early auditory system

The early stages of auditory processing are modeled as a sequence of three operations depicted in Fig. 1 (Lyon and Shamma, 1996; Shamma et al., 1986).

- The acoustic signal entering the ear produces a complex spatio-temporal pattern of vibrations along the basilar membrane of the cochlea (Fig. 1, left panel). The maximal displacement at each cochlear point corresponds to a distinct tone frequency in the stimulus, creating a tonotopically ordered response axis along the length of the cochlea. Thus, the basilar membrane can be thought of as a bank of constant-Q highly asymmetric bandpass filters ($Q = 4$) equally spaced on a logarithmic frequency axis. Our model employs 24 filters/octave over a 5 octave range.
- The basilar membrane outputs are then converted into inner hair cell intra-cellular potentials. This process is modeled as a 3-step operation: a highpass filter (the fluid-cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels), and then a lowpass filter (hair cell membrane leakage). Detailed description of the mechanisms involved in each step can be found in Lyon and Shamma (1996) and Shamma et al. (1986).
- Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array (Shamma, 1998). It is modeled as a first difference operation across the channel array, followed by a

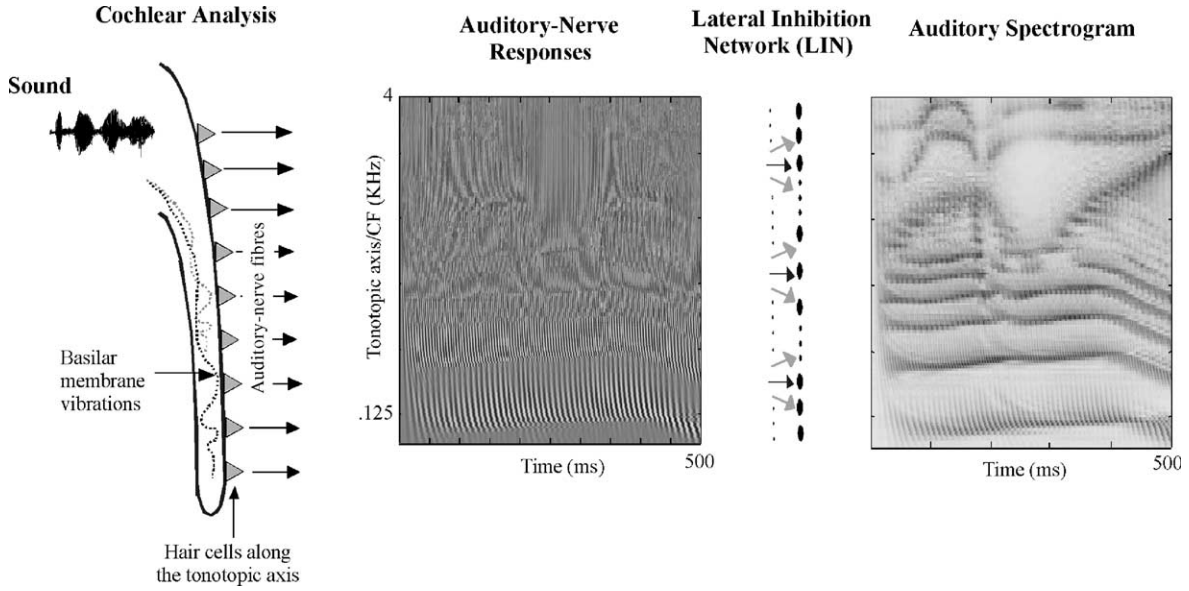

Fig. 1. *Schematic of the early stages of auditory processing*. Sound is analyzed by a model of the cochlea (depicted on the left) consisting of a bank of 128 constant-Q bandpass filters with center frequencies equally spaced on a logarithmic frequency axis (tonotopic axis) spanning 5.2 octaves (e.g., 0.1–4 kHz). Each filter output is then half-wave rectified and lowpass filtered by an inner hair cell model to produce the auditory-nerve response patterns (middle panel). A spatial first-difference operation is then applied mimicking the function of a lateral inhibitory network (LIN) which sharpens the spectral representation of the signal and extracts its harmonics and formants (Shamma, 1998). In this study, the short-term integration is performed over 8 ms intervals. A final smoothing of the responses on each channel results in the *auditory spectrogram* depicted on the right.

half-wave rectifier, and then a short-term integrator. This stage effectively sharpens the bandwidths of the cochlear filters from about $Q = 4$ to 12, as explained in detail in (Wang and Shamma, 1994).

The above sequence of operations *effectively* computes a spectrogram of the speech signal (Fig. 1, right panel) using a bank of constant-Q filters, with a bandwidth tuning Q of about 12 (or just under 10% of the center frequency of each filter). Dynamically, the spectrogram also encodes explicitly all temporal "envelope modulations" due to interactions between the spectral components that fall within the bandwidth of each filter. The frequencies of these modulations are naturally limited by the maximum bandwidth of the cochlear filters.

2.1.2. The central auditory system

Higher central auditory stages (especially the primary auditory cortex) analyze further the auditory spectrum into more elaborate representations, interpret them, and separate the different cues and features associated with different sound percepts. Specifically, from a conceptual point of view, these stages estimate the spectral and temporal modulation content of the auditory spectrogram as illustrated in Fig. 2. They do so computationally via a bank of modulation-selective filters centered at each frequency along the tonotopic axis (Chi et al., 1999). Each filter is tuned ($Q = 1$) to a range of temporal modulations (ω, also referred to as rates or velocities (in *Hz*)) and spectral modulations (Ω, also referred to as densities or scales (in *cyc/oct*)). It has a spectro-temporal impulse response (usually called spectro-temporal response field, STRF) in the form of a spectro-temporal Gabor function (see Eq. (5) in Chi et al., 1999). An example of an STRF is shown in Fig. 2B, together with the result of convolving it with the auditory spectrogram to the left. Since the response is a 4 dimensional complex-valued function (time, frequency, rate and scale); then, for display purposes, we shall sometimes provide only

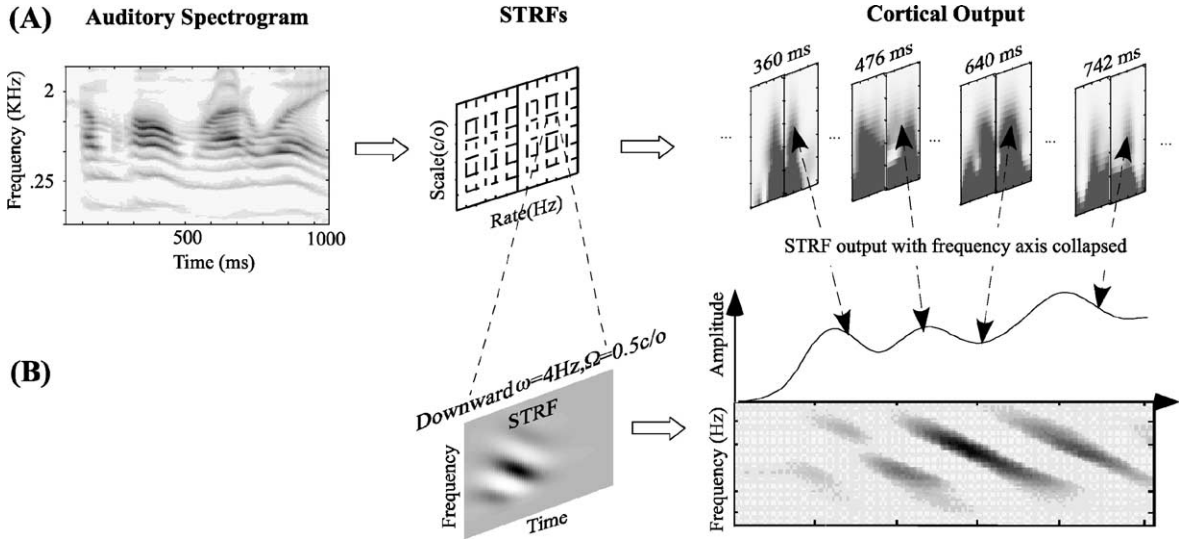

Fig. 2. *The cortical multi-scale representation of speech*. (A) and (B) The auditory spectrogram of a speech sentence /come home right away/, spoken by a male is analyzed by a bank of spectro-temporal modulation selective filters. The spectro-temporal response field (STRF) of one such filter (tuned to $\omega = 4$ Hz and $\Omega = 1$ cyc/oct) is shown in (B, left panel) below. The output from such a filter is computed by convolving the STRF with the input spectrogram, to produce a new spectrogram as shown in (B, right panel). The total output as a function of time from the model is therefore indexed by three parameters: scale-Ω, rate-ω, and frequency-x. We often collapse (integrate over) the frequency axis (x) for display purposes which reduces the output from each filter to a one dimensional time-function as shown on top of the spectrogram in (B, right panel). The total output in this case becomes a series of two-dimensional scale-rate plots as shown in (A, right panels).

the *magnitude* of the response as a function of frequency at each time instant (or for all time if it is constant, as for a stationary stimulus). Where the spectro-temporal modulation content of the spectrogram is of particular interest, we shall display the summed output from all filters with identical modulation selectivity or STRFs (i.e., integrate the *x*-axis out) to generate the *scale-rate* plots as shown in Fig. 2A (right panel). The final view that emerges is that of a continuously updated estimate of the spectral and temporal modulation content of the auditory spectrogram. All parameters of this model are derived from physiological data in animals and psychoacoustical data in human subjects as explained in detail in Chi et al. (1999); Kowalski et al. (1996); and Depireux et al. (2001).

2.2. The spectro-temporal modulation index (STMI)

Broadly speaking, the STMI is a measure of the *changes* in the auditory model output when noise, reverberations, or other distortions are applied to the sound signal. Thus, to measure the intelligibility of a noisy token of speech or other complex sounds, or to characterize a channel (e.g., a recording or transmission medium, a room, or a vocoder), we use the auditory model to estimate the change in the spectro-temporal modulations that a test speech signal undergoes. We propose here two types of STMI: the first (denoted later as STMI^{T}) is derived directly from the speech samples; the second (STMI^{R}) is based on characterizing the integrity of spectro-temporally modulated test signals (called *ripples*) when transmitted through the channel under study. These two indices are analogous to two versions of the STI: one is derived directly from the speech signal and uses the clean speech modulations as the reference (Payton and Braida, 1999); the second is based on the standard definition using narrowband noise carriers (Houtgast and Steeneken, 1980).

2.2.1. Computing the STMI of speech samples ($STMI^T$)

The STMI quantifies the difference between the spectro-temporal modulation content of the *noisy* and *clean* speech signals. The procedure is depicted in Fig. 3. We first analyze the clean speech sentence through the auditory model as in Fig. 2. The 4-D output is averaged over the stimulus duration to generate the 3-D template of the speech token $\{T\}$. Similarly, the averaged output $\{N\}$ of the noisy speech token is computed. In both cases, the auditory outputs must be adjusted by subtracting from each the output due to its own "base" spectrum. The "base" is a stationary noise with a spectrum identical to that of the long-term average spectrum of the appropriate signal (clean or noisy speech). The STMI is then computed, and denoted by STMI^{T} to emphasize that a speech template is used as the clean reference:

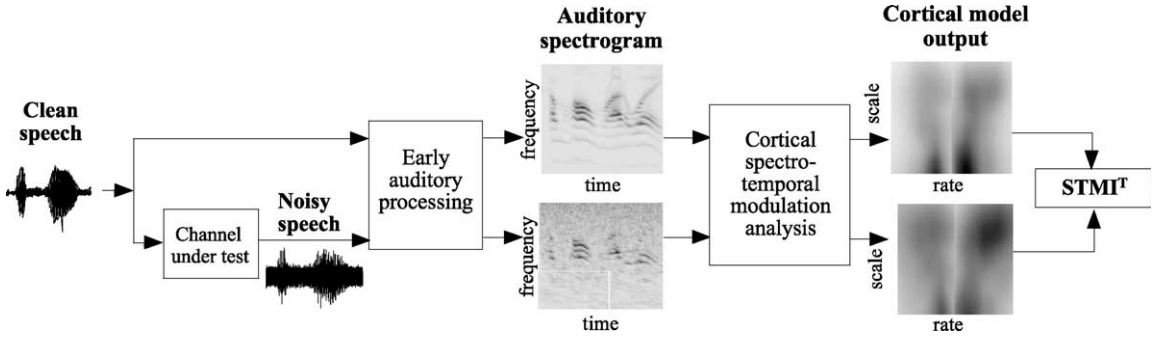

Fig. 3. *Schematic of the $STMI^T$ computation*. The clean and noisy speech signals are given as inputs to the auditory model. Their outputs are normalized by the base signals as explained in the text. The right panel shows the cortical output of both clean and noisy inputs. These cortical patterns are then used to compute the template-based STMI.

$$\mathrm{STMI}^{\mathrm{T}} = 1 - \frac{\|T-N\|^2}{\|T\|^2} \quad (1)$$

where the distance $\|T-N\|^2$ is taken to be the shortest distance between the model outputs of the noisy token and the clean template(s).

2.2.2. Computing the STMI with ripple stimuli ($STMI^R$)

The $\mathrm{STMI}^{\mathrm{R}}$ of a channel can also be defined with respect to the fidelity of its transmission of specially designed spectro-temporally modulated test signals called *ripples*. Specifically, we use the auditory model and the ripple stimuli to measure the effects of a noisy channel on the overall MTF. Ripples combine both spectral and temporal modulations, and have been previously described in detail in (Chi et al., 1999). We first briefly describe the ripples, and then provide the procedure for using them to measure the MTF and compute the STMI^R.

2.2.2.1. The ripple stimuli. The moving ripple stimuli used in this study are broadband complex sounds consisting of 280 tones equally spaced along the logarithmic frequency axis, over a range of 5 octaves (250–8000 Hz). The spectral envelope of these stimuli forms sinusoids whose amplitude is modulated by an amount specified by ΔA (typically 100%) on a linear modulation scale. For example, $\Delta A = 0\%$ corresponds to zero modulation of the flat ripple spectrum, whereas $\Delta A = 100\%$ corresponds to 100% modulation of the flat ripple. This construction forms a drifting sinusoidally shaped spectrum along the frequency axis. The *envelope* of a moving ripple stimulus ($S(x,t)$) is fully described by the equation:

$$S(x,t) = L(1 + \Delta A \sin(2\pi(\omega t + \Omega x) + \varphi)) \quad (2)$$

where L denotes the overall level of the stimulus, t is time, and x is the tonotopic axis, defined as $x = \log_2 f/f_0$, with f_0 being the lower edge of the spectrum, and f the frequency. ω is the ripple velocity (in cyc/s), Ω is the ripple density along the x-axis (in cyc/oct), and φ is the phase of the ripple. Fig. 4A illustrates the spectrogram of such a downward sweeping ripple.

2.2.2.2. Using the ripples to measure MTF. The MTF of the auditory model (with or without an additional channel) is measured using single ripples at rate-scale $(\omega - \Omega)$ combinations over a range of $\Omega = 0.25$–8 (cyc/oct), and $\omega = -32$–32 (Hz), with negative rates denoting upward moving ripples. The input stimuli are all defined over a finite spectral range $x \in [0, X]$ (typically 5 octaves), and temporal extent $t \in [0, T]$ (typically 1 s). The MTF calculation procedure is described below, and illustrated schematically in Fig. 4B.

For each input ripple combination $\{\alpha, \beta\}$, the $\{\omega_\alpha, \Omega_\beta\}$ ripple with contrast $\Delta A = 100\%$ is given as input to the auditory model. The corresponding auditory spectrogram $y_{\alpha,\beta}(t,x;\Delta A)$ is computed and then analyzed by a bank of cortical filters $\{\mathrm{STRF}(\cdot)\}$ to generate the final integrated output pattern $\{r(\cdot)\}$ for each cortical filter (i,j):

$$r^{i,j}_{\alpha,\beta}(x;100\%) = \int_T \|y_{\alpha,\beta}(t,x;100\%) \ *_{t,x} \ \mathrm{STRF}^{i,j}(t,x)\| \, \mathrm{d}t \quad (3)$$

where $*_{t,x}$ is the convolution in time (t) and multiplication in frequency (x); STRF denotes the spectro-temporal impulse response of each filter, indexed by the best ripple scale and rate $\{\omega_i, \Omega_j\}$ of the filter, and $\|\cdot\|$ denotes the instantaneous magnitude of the final response. Note that the final output magnitude is integrated over the interval T, which may be as short as one frame of speech (8 ms) or the entire stimulus (as in the case of the test stationary ripples and static noise interference). The values of (i,j) correspond to the indices of the filter $\{\omega_i, \Omega_j\}$ in the discrete set $(\hat{w}, \widehat{W})$. The output pattern $r_{\alpha,\beta}$ is illustrated in Fig. 4B (rightmost panels) for the ripple $\omega_\alpha = 4$ Hz, $\Omega_\beta = 1$ cyc/oct.

In order to take into account the base level of the input stimuli in the transfer function calculation, we repeat the procedure for the same ripple $\{\omega_\alpha, \Omega_\beta\}$, but this time with contrast $\Delta A = 0\%$ to get the output pattern $r^{i,j}_{\alpha,\beta}(x;0\%)$ following the same calculations as in Eq. (3). This flat (0%) response pattern is then subtracted from the 100% contrast response to yield the actual response $R^{i,j}_{\alpha,\beta}(x)$ defined as:

$$R^{i,j}_{\alpha,\beta}(x) = r^{i,j}_{\alpha,\beta}(x;100\%) - r^{i,j}_{\alpha,\beta}(x;0\%) \quad (4)$$

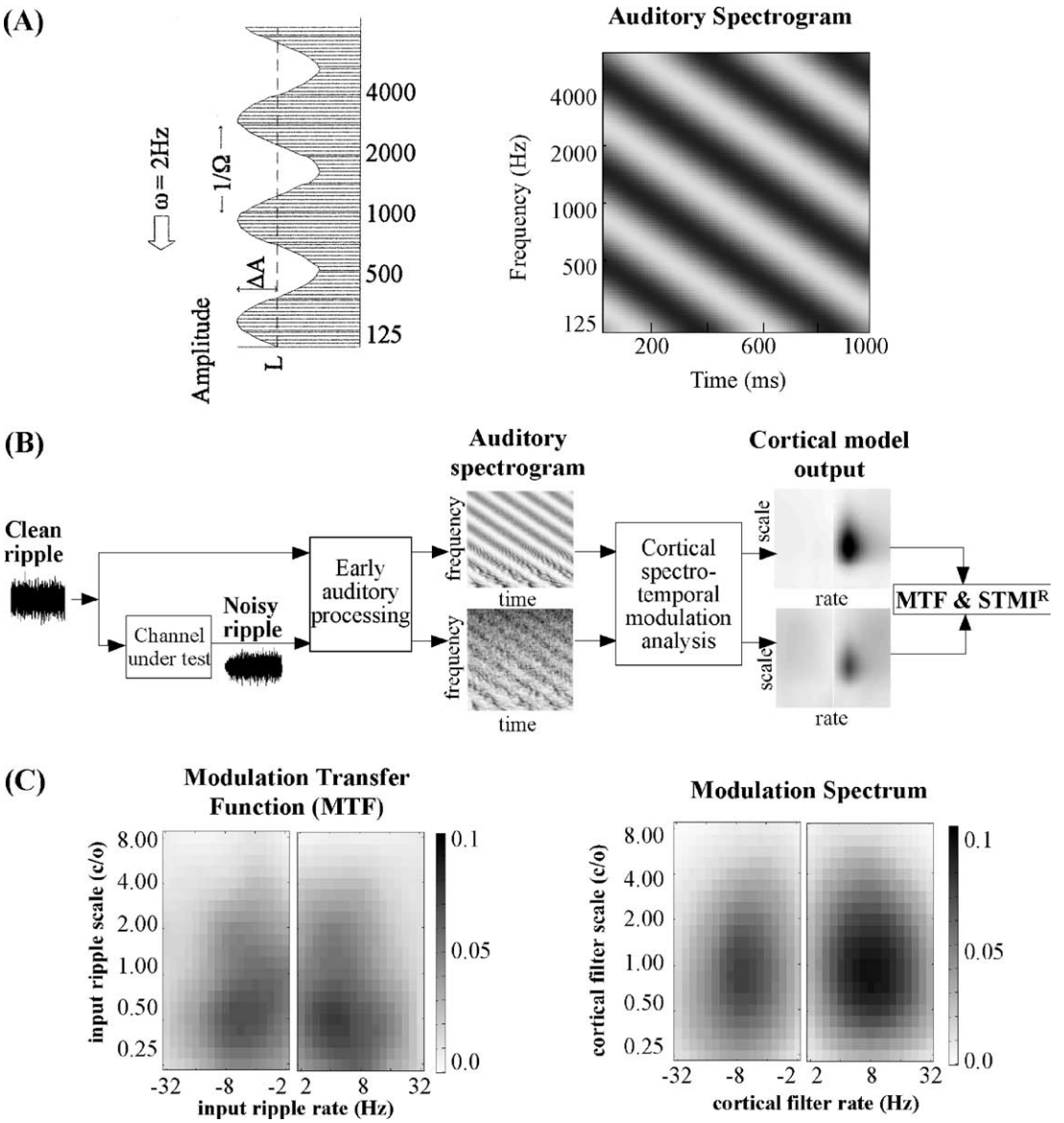

Fig. 4. *The ripple-based STMI and modulation transfer function (MTF)*. (A) The ripple spectrum at one instant in time is shown in the left panel. The ripple envelope is moving down at a rate of 2 Hz (i.e., the envelope drifts at 2 cyc/s) and has a spectral density of 0.6 cyc/oct (see text for details). The right panel illustrates the spectrogram of the envelope of the ripple. (B) To measure the MTF of a channel at a particular ripple, we first compute the auditory model outputs in response to the clean and channel-distorted ripples, and then normalize them by the output to the (flat) base of the ripple (see text for details). The panels on the right depict the cortical output collapsed on the rate-scale axes. Note the effect of the channel distortion on the clarity of the ripple representation in the auditory and cortical representations (see Chi et al., 1999 for more details). The MTF for this ripple is defined as the global gain (or attenuation) in the cortical output. This procedure is repeated at all ripple parameters to compute the total MTF. (C) The left panel depicts the global MTF obtained by averaging across cortical filters. The right panel shows the modulation spectrum computed by averaging across input ripples velocities and densities. Both graphs (left and right panels) are shown as a rate-scale plot. The frequency axis x is collapsed for display purposes.

Finally, the responses to all the ripple stimuli (all α, β) are averaged over all filters $(i, j) \in (\hat{w}, \widehat{W})$ to yield the overall transfer function of the system:

$$\mathrm{MTF}(x; \omega_\alpha, \Omega_\beta) = \frac{1}{|\hat{w}| \cdot |\widehat{W}|} \sum_i \sum_j R^{i,j}_{\alpha,\beta}(x) \qquad (5)$$

Note that this is a 3D pattern indexed by $(x, \omega_\alpha, \Omega_\beta)$, respectively, the frequency, input ripple velocity and density. This overall MTF is depicted in Fig. 4C (left panel); Note that for display purposes, the MTF here is averaged over the frequency dimension (x).

An alternative representation of the transfer characteristics would be in the form of an average *modulation spectrum* of all ripples presented at equal (unit) amplitude. This pattern results from taking the sum in Eq. (5) over all ripples $(\omega_\alpha, \Omega_\beta)$ (instead of over the channels centered at ω_i, Ω_j). Because of the band-pass nature of the filters, the resulting *modulation spectrum* (indexed by (x, ω_i, Ω_j)) is roughly similar to the MTF as illustrated in Fig. 4C (right panel). Again, this modulation spectrum is averaged over the frequency axis x for display purposes.

In the presence of any kind of noise, a similar procedure is followed to derive the noise-contaminated MTF. The input ripples used in this case are first contaminated by the noise (for example, by passing them through the channel under investigation). The resulting noisy transfer function MTF* $(x, \omega_\alpha, \Omega_\beta)$ is computed according to Eq. (5) as described above (see Fig. 4B).

2.2.2.3. Defining the ripple-based STMI^R. For a given noise condition in a channel (communication link, auditorium), we estimate the $\mathrm{STMI}^{\mathrm{R}}$ as a global measure of the attenuation of the spectro-temporal modulations in the signal when passed through the channel. This eventually translates to a measure of the expected intelligibility of a speech signal transmitted through this channel. The $\mathrm{STMI}^{\mathrm{R}}$ is defined as:

$$\mathrm{STMI}^{\mathrm{R}} = 1 - \frac{\|\mathrm{MTF} - \mathrm{MTF}^*\|^2}{\|\mathrm{MTF}\|^2} \qquad (6)$$

where

$$\|\mathrm{MTF}(x; \omega, \Omega)\| = \sqrt{\sum_k \sum_i \sum_j (\mathrm{MTF}(x_k; \omega_i, \Omega_j))^2} \qquad (7)$$

3. Results

In this section, we illustrate how the STMI is computed and used to characterize the integrity of the spectro-temporal modulations when distorted by various kinds of noise. For speech, we consider the STMI as a measure of the intelligibility of the signal in the same manner the AI and STI have been traditionally used. Since the $\mathrm{STMI}^{\mathrm{R}}$ is analogous to the traditional STI (using narrow-band carriers), we begin by comparing its estimates to those of the STI under different noise and reverberation conditions. Next, we illustrate examples of $\mathrm{STMI}^{\mathrm{T}}$ measurements for speech, and compare them to results of psychophysical tests.

3.1. The effect of noise and reverberations on STMI^R

The representation of acoustic spectro-temporal modulations in a signal is progressively degraded when noise or reverberations are added to it. The extent of the degradation is dependent on the rate and scale of the modulations, and the spectral content of the signal. This is illustrated in Fig. 5, where we plot the $\mathrm{STMI}^{\mathrm{R}}$ derived from the MTF as in Eq. (6), but for a single ripple at different rates and scales, under various white noise and reverberation conditions.

The stationary white noise condition used here is generated by adding to the original signal a random Gaussian signal whose amplitude is defined according to the signal-to-noise ratio (SNR) level. The reverberation effect is produced by convolving the signal with Gaussian white noise whose envelope is exponentially decaying.

The first observation is that increasing the level of stationary noise attenuates ripple modulations (and hence the $\mathrm{STMI}^{\mathrm{R}}$) equally regardless of rate and scale (Fig. 5A). This is not the case with increasing reverberation, where ripples with faster rates are attenuated more severely as expected

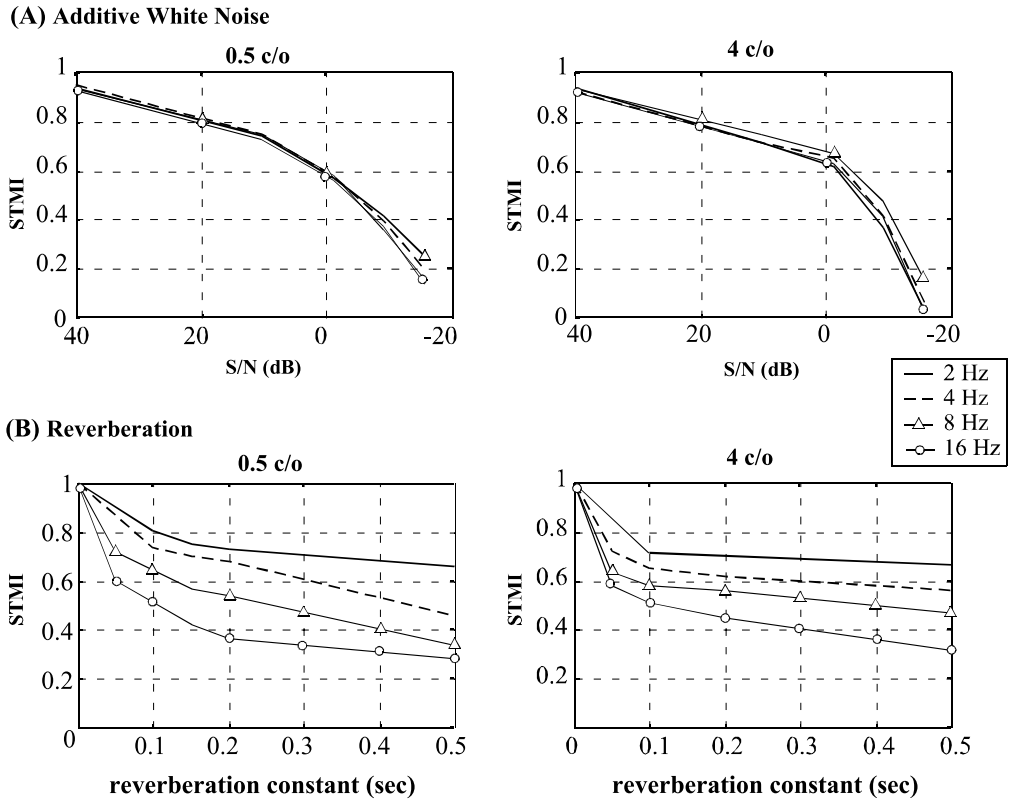

Fig. 5. *Effect of white noise and reverberation on the MTF with single ripples*. (A) White noise attenuates the output of a single ripple by the same amount regardless of ripple rates (2–16 Hz) or spectral densities (0.5, 4 cyc/oct). (B) Reverberation attenuates the responses to high rate ripples (16 Hz) significantly more than to low rate ripples (2 Hz) regardless of spectral density.

from the low-pass (smoothing) effect of the reverberation on the ripple envelope (Fig. 5B).

Fig. 6 summarizes the effect on all ripples (and the clean MTF of the auditory model in Fig. 6A) of the added white noise (Fig. 6B), of different levels of reverberation (Fig. 6C), and of the combined effect of noise and reverberation (Fig. 6D). In each case, the MTF* is plotted as a function of $\{\omega_i, \Omega_j\}$, i.e., we integrate out the frequency axis x. It is important to note here that one can apply any arbitrary noise condition and compute the resulting MTF* using exactly the same expressions presented in the previous section. These plots illustrate the effects of each of these distortions as follows. For noise (Fig. 6B), the MTF* is gradually and equally attenuated over all ripples. For increasing reverberation (Fig. 6C), higher rate ripples are more severely attenuated than lower rates. Both these trends are seen in Fig. 6D for the combined noise and reverberation conditions. Note that the "random" weak patterns seen in Fig. 6D reflect the random noise structure in a given trial, and hence are variable over different trials.

3.2. Comparing the $STMI^R$ and STI for noisy and reverberant conditions

$STMI^R$ values are computed from clean and degraded modulation transfer functions (MTF and MTF*) using Eq. (6). They are displayed in Fig. 7A for the two sets of conditions. As expected, the $STMI^R$ decreases with increasing noise and reverberation. Fig. 7B (left panel) illustrates the STI estimates for the same conditions. These were computed using commercially available software: *Lexington's Speech Transmission Index Program*,

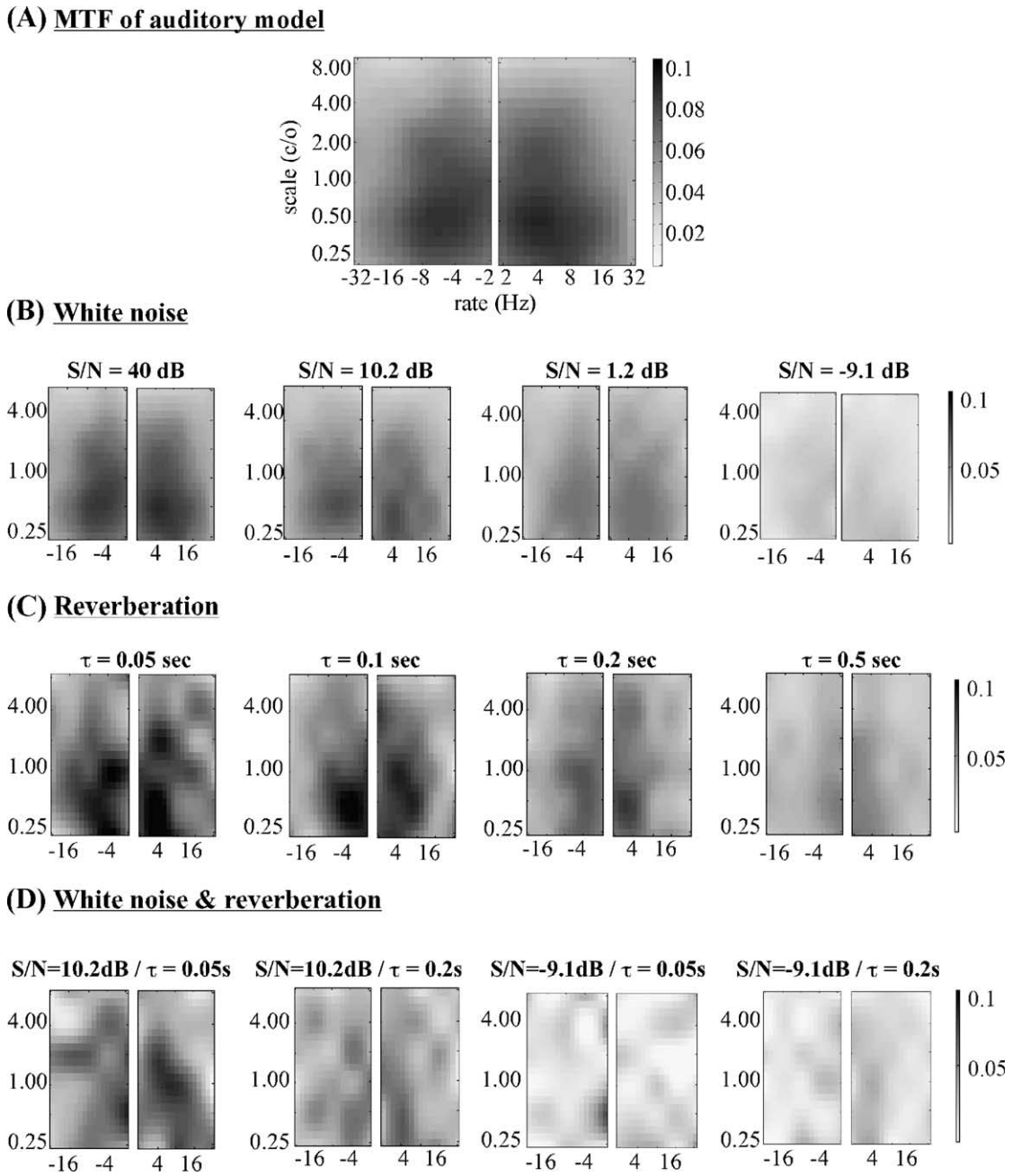

Fig. 6. *Effect of white noise and reverberation on the global MTF*. (A) The global (clean) MTF of the auditory model computed from all ripples, summarized by the rate-scale plot (i.e., collapsing the frequency axis x). (B) The attenuation of the global MTF (rate-scale plot) with increasing levels of white noise. (C) The attenuation of the global MTF at higher rates with increasing reverberation. (D) The combined effect on the global MTF of both additive white noise and reverberation.

based on the MTF method derived by Steeneken and Houtgast (1979). The software is available at http://hearingresearch.org/STI.htm. Although different in details, the $STMI^R$ and the STI measures

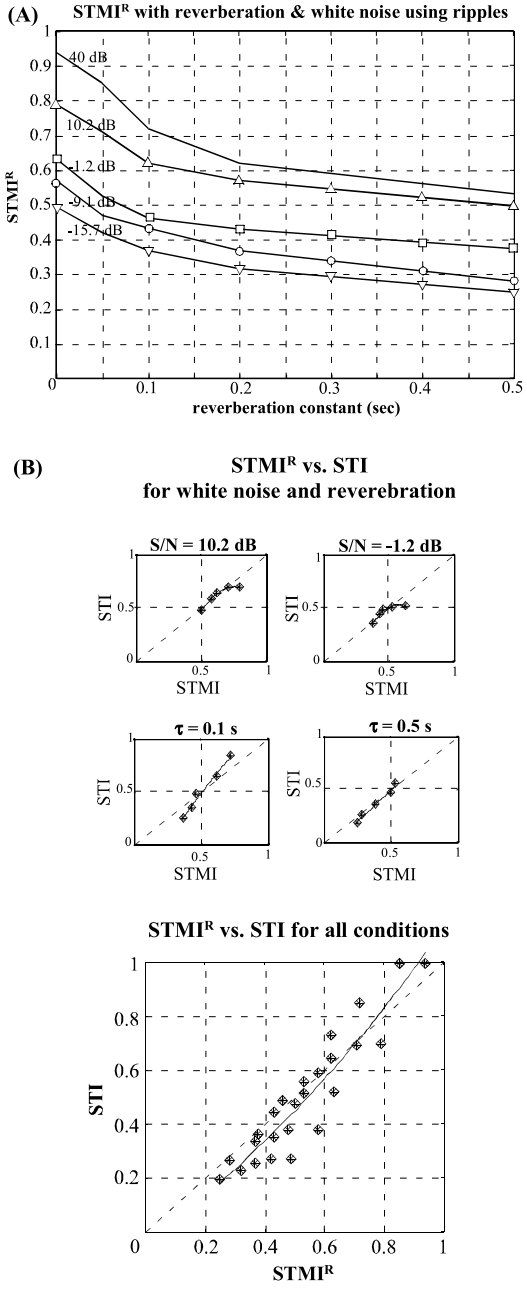

Fig. 7. *Effect of combined white noise and reverberation on STMI^R and STI.* (A) The STMI values shown in this plot are computed according to Eq. (6) for noise conditions combining stationary noise and reverberation. (B) The correspondence between the STMI and STI is given in the left panel for specific conditions of stationary white noise, and reverberation. The right panel shows the overall correspondence between STMI and STI for all conditions.

deteriorate similarly under these noise and reverberation conditions, and an approximate mapping between these two measures can be derived as shown in Fig. 7B (right panel).

3.3. Comparing STMI^T to human perception

As the STMI of a channel gradually decreases, speech transmitted through it should exhibit a concomitant loss of intelligibility that can be experimentally measured as increased phoneme recognition error rates. To relate the STMI values directly to experimental measurements of speech intelligibility, we plot in Fig. 8A the STMI^T of speech tokens (computed from Eq. (1)) with increasing additive noise and reverberation distortions. The template used in this simulation was derived by averaging the model output for each clean speech token over the entire duration of the utterance of that particular token. The noisy pattern $\{N\}$ is similarly computed by averaging the model output of the noisy speech token. While the trends in the STMI^T values are essentially similar to the STMI^R estimates in Fig. 7A, the one noticeable difference is the shallower drop of the STMI^T with reverberation, presumably due to time-averaging of the output patterns. Another source of the difference between the two measures is that they are conceptually different—the STMI^T is based on the modulation spectrum while the second, STMI^R, on the MTF. While the two are close (Fig. 4C), they are not identical, especially when the modulation filters are not very selective ($Q = 1$) leading to interactions among simultaneously applied ripples.

These results were compared to actual intelligibility scores from four subjects using speech samples contaminated by the same combined stationary noise and reverberations. Each subject was presented with 240 sets of noise-contaminated speech samples through a loudspeaker in an acoustic chamber and asked to repeat them. Each set consisted of five different words. A count of the correct phonemes reported was then averaged over all test subjects for each noise condition. The percent correct recognition scores found in these experiments are given in Fig. 8B. The good correspondence between the STMI^T and the human

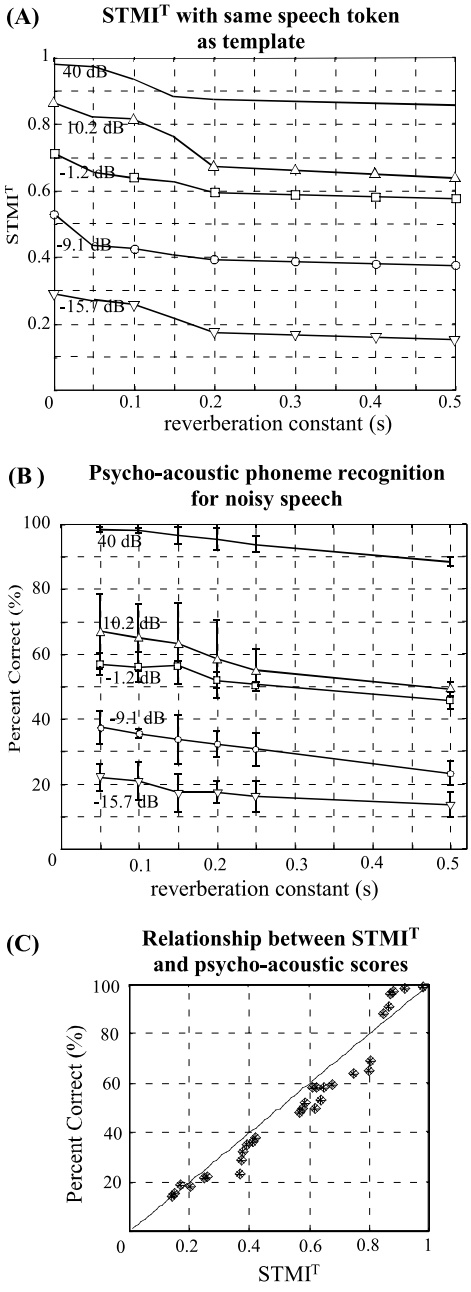

Fig. 8. *Comparing the effect of combined white noise and reverberation on the STMI^T and speech intelligibility*. (A) The STMI^T of speech signals distorted by noise and reverberation. The STMI^T is computed according to Eq. (1) using same speech as template. (B) Experimental measurements of correct phoneme recognition of human subjects in noisy and reverberant conditions. (C) The STMI^T vs. correct percentages of human psycho-acoustic experiments for the noise conditions given in (A) and (B).

scores (summarized by the data re-plotted in Fig. 8C) confirms that the STMI^T is indeed a direct measure of the intelligibility of noisy speech under conditions of combined white noise and reverberation.

Finally, for completeness, we show in Fig. 9 the STMI^T computations of speech using templates derived from clean speech samples that were *different* from the noisy speech samples. In these simulations, both clean and noisy speech samples were derived from randomly selected sentences in the TIMIT database. For each simulation, we used 10 different sentences that were sampled at 8 kHz.

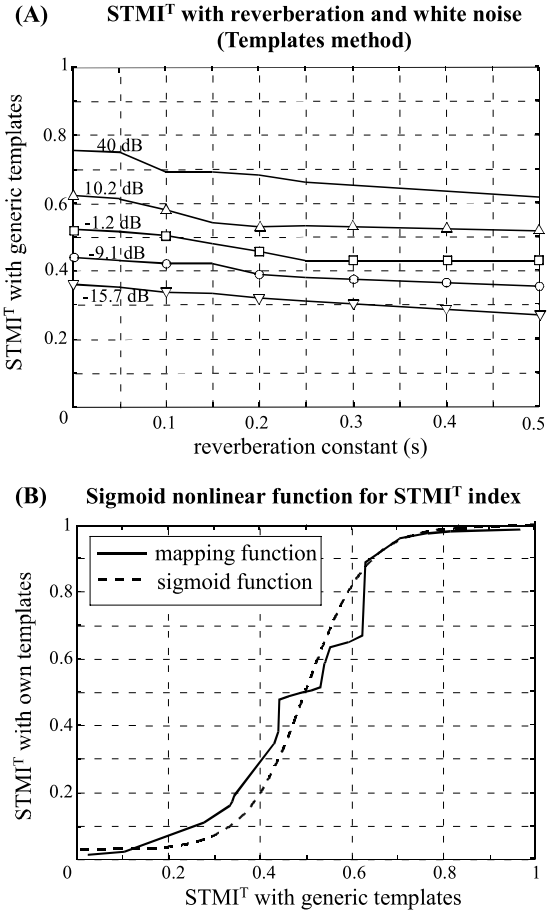

Fig. 9. *Effect of combined white noise and reverberation on speech STMI^T using generalized templates*. (A) The trends in the STMI^T decrease with noise and reverberation are similar to the case in Fig. 8. (B) Expansive nonlinear function to correct the STMI^T values to match those of Fig. 8A.

Again, the $STMI^T$ trends are essentially similar to those of Fig. 9 with one significant difference, namely the lower values of the $STMI^T$ in the clean conditions due to the inevitable mismatch between the (clean) tested speech and the templates. The $STMI^T$ derived from these two types of templates (Figs. 8A and 9A) can be related by the sigmoid non-linearity shown in Fig. 9B.

3.4. Comparing the $STMI^T$ and STI with phase jitter and shifts

The STI has been widely and successfully used in speech intelligibility assessments under noise and reverberant degradation, and has also been adapted for use with speech signals directly (Payton and Braida, 1999). Therefore, the results described above only demonstrate the correspondence between the STMI and STI, and hence the validity of the new measure. Here we compare the two measures under more difficult types of degradations: random phase-jitter and phase-shifts. These are chosen specifically because the STI clearly fails to characterize them correctly. Speech samples distorted by these conditions are available at http://www.isr.umd.edu/CAAR/pubs.html. Also included in this section are the results of psychoacoustic experiments measuring the loss of intelligibility experienced by four subjects listening to words distorted by these two conditions. All experiments were conducted exactly as described earlier in Section 3.3. The subjects were presented with 160 different distorted words. The subjects were then asked to repeat the words heard. Scores of average correct phonemes reported are presented in Figs. 10 and 11 for the two conditions.

3.4.1. Phase jitter distortions

The first condition is *phase jitter*, a condition commonly associated with telephone channels and caused by the fluctuations of the power supply voltages (Lee and Messerschmitt, 1994; Bellamy, 2000). Communication engineers report that channels cannot be defended against such degradation, but it must be taken into account in the design of the receiver (Lee and Messerschmitt, 1994). Therefore, studying the effect of this distortion on speech intelligibility is critical for improving the channel and receiver designs.

Phase jitter is commonly modeled by:

$$r(t) = Re\{s(t)e^{j\Theta(t)}\} = s(t)\cos(\Theta(t)) \tag{8}$$

where $s(t)$ is the transmitted signal, $r(t)$ is the received signal, and $\Theta(t)$ is the phase jitter function modeled as a random process uniformly distributed over $[0, 2\alpha\pi]$ $(0 < \alpha < 1)$. This jitter effectively destroys the carrier of the speech signal leaving its envelope largely intact (Fig. 10A), especially for values of α that are large enough. For $\alpha = 1$, the speech signal becomes a modulated white noise with the same envelope as before. Fig. 10B illustrates the expected loss of intelligibility as a function of jitter severity (α) as measured by the STI, $STMI^R$, and $STMI^T$ (computed as the mean of 10 different speech sentences from the TIMIT database). The STI is insensitive to such a distortion. By contrast, the STMI deteriorates with increasing α. The fundamental reason for this disparity is that the effect of phase jitter is mostly manifested in the spectral dimension (Fig. 10A), and hence does not affect the modulation amplitude of the narrowband carriers used in the STI measurement. The effect on the spectrogram of *oriented* ripples is substantial, and hence the $STMI^T$ and $STMI^R$ change accordingly. Note that by contrast the speech-based STI (Payton and Braida, 1999) will *not* sense the speech deterioration since the average modulation spectrum remains largely unaffected by the phase-jitter. Finally, we also show in Fig. 10B the results of human subject intelligibility testing which match well the predicted results from the STMI.

3.4.2. Inter-channel phase-shifts or delay scatter

The second type of channel distortion is a *linear phase-shifting* of signal frequencies over limited ranges as demonstrated in Fig. 11A. Specifically, the effects of this distortion are seen in the spectrogram as an inter-channel delay scatter or a de-synchronization of the channel outputs, with minimal change in the envelope modulation patterns on any given channel, as illustrated by the distorted spectrogram of Fig. 11A. The phase-shift function here is given by $\Phi = \omega\tau_i$, applied over

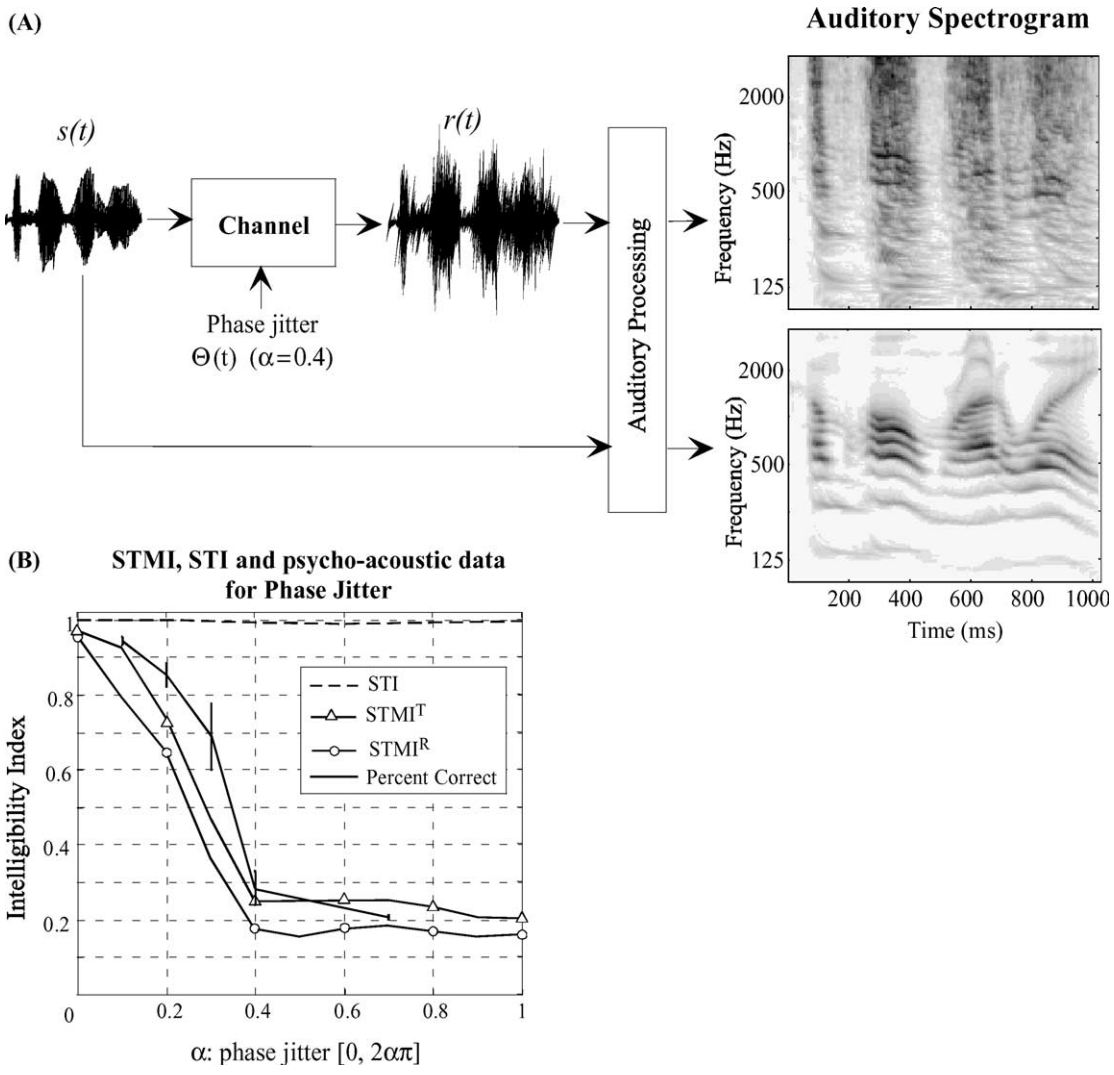

Fig. 10. *Effect of phase jitter on STMI and STI.* (A) The input signal $s(t)$ is sent through the channel, and received as $r(t)$. The channel has a phase jitter function $\Theta(t)$. In this figure, $\Theta(t)$ is uniformly distributed over the range $[0, 2\pi^{*}0.4]$. The right panel shows the spectrogram of the sentence /come home right away/ with and without the effect of the phase jitter. The spectrograms illustrate that the time dynamics of the signal are maintained while the spectral modulations are strongly affected by this type of noise. (B) The STMI^{T} and STMI drop as the jitter increases; STI fails to capture the presence of noise in this channel.

300 Hz frequency bands (each indexed by i) over the range 400–1900 Hz $(i = 1, \ldots, 5)$, where $\omega = 2\pi f$ is the frequency at which the phase-shift is applied, and τ_i is a parameter which controls the slope of the phase function (and hence the delay imposed) in the ith band. Fig. 11B illustrates the decrease in STMI^{R} and STMI^{T} with increasing delay scatter (over a range of τ values), consistent with the increasing channel distortion of the spectrogram of the ripple and speech signals. The STMI^{T} drops faster because of the specific arbitrary choice of frequency bands and shifts; and the drop (while it always occurs) is variable in steepness depending on the exact sentence. As with the previous phase-jitter distortion, STI measures (noise or speech-based) are expected to be insensi-

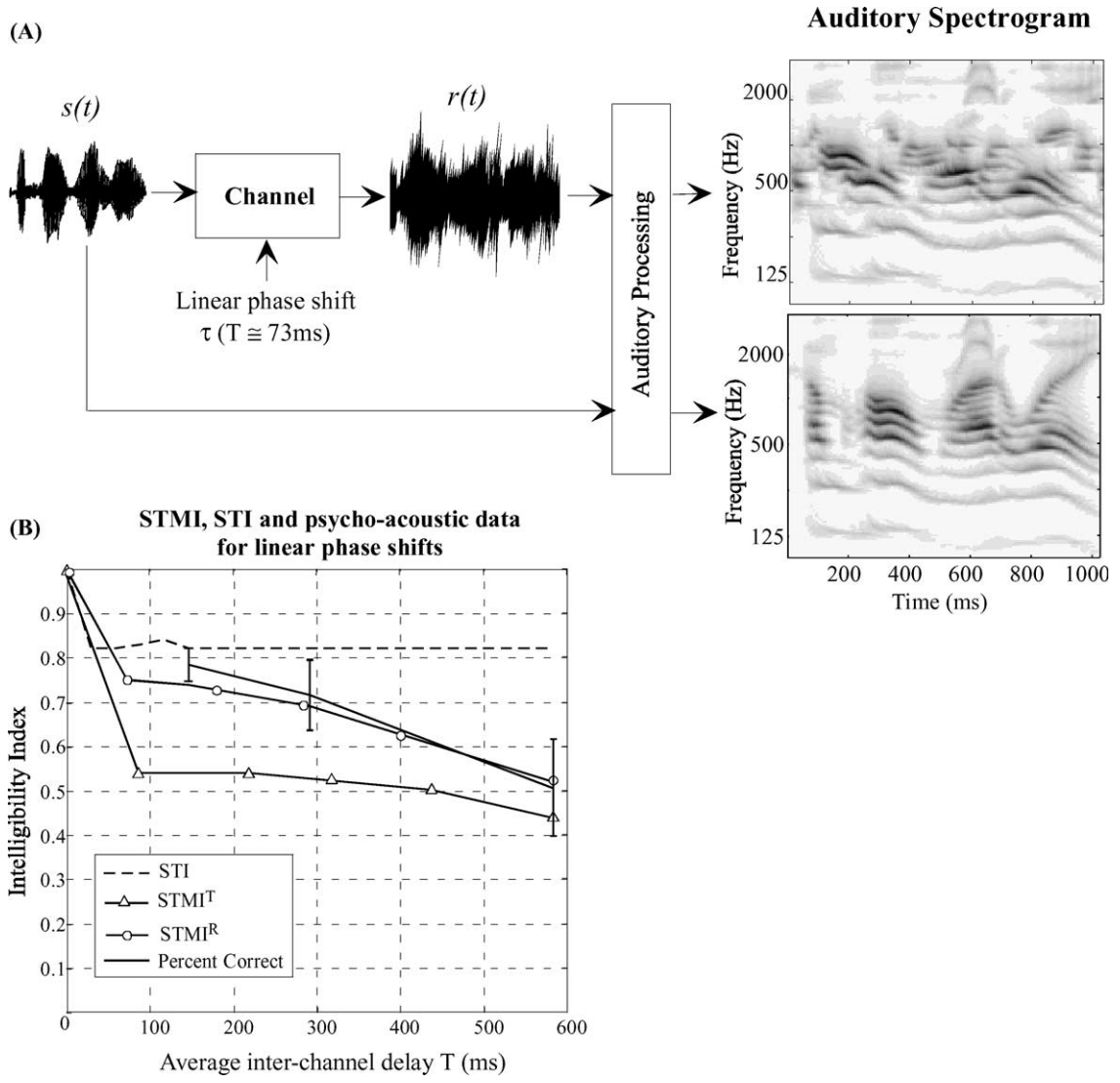

Fig. 11. *Effect of linear phase shift on STMI and STI.* (A) The input speech signal $s(t)$ (/come home right away/) is distorted by linear phase shifts on different frequency bands. Five frequency bands (of uniform 300 Hz ranges going from 400 to 1900 Hz) are phase-shifted according to the vector $[\tau, 2\tau, -3\tau, 4\tau, 5\tau]$ (a different phase shift per frequency band) where τ is the parameter that controls the amount of shift per band. The result is a de-synchronization of the different frequency bands relative to each other. This effect can be seen in the spectrograms of the clean and noisy sentences depicted in the right panel. The shift parameter used in this case is $\tau = 0.5$ (average time shift across channels of $T \cong 73$ ms). (B) The effect of the shift parameter τ (or T) on the STMI, STMI^{T} and STI. Since for each value of the shift parameter (τ) different frequency bands are time-shifted with various amounts relative to each other, the x-axis of the graph gives an average estimate (T) of the time shifts across the different frequency bands (where $T \cong 146.\tau$ (ms)). The values of STMI^{T} shown in this plot are computed for the sentence /come home right away/, whose spectrogram is given in the upper panel.

tive (Fig. 11B) to such phase-shift because this distortion does not significantly affect the modulated envelope of the narrow-band carrier test signals used in standard STI computations, nor does it affect the envelope modulations of the speech spectrogram. Human intelligibility exhibits the same deterioration as that predicted from the STMI as illustrated in Fig. 11B. Our results are

comparable to those of Greenberg and Arai (1998) who studied the intelligibility of a similarly (but not identically) distorted speech and concluded that scores dropped below 50% only after the channel jitter exceeds 200 ms.

4. Summary and conclusions

We have argued in this report that a multi-scale analysis of the spectro-temporal modulations can be effectively used to quantify the intelligibility of speech signals and the ability of a channel to transmit intelligible speech. The model parameters of the modulation analysis are based on physiological findings in the primary auditory cortex and on psychoacoustical measurements of human sensitivity to spectral and temporal modulations. The model was used to derive an intelligibility index, the STMI, which simply reflects the deterioration in the spectro-temporal modulation content of ripples or speech due to any added noise or reverberation. The STMI was validated by demonstrating that its predictions match those of the classical STI and also match error rates of human listeners in the case of speech contaminated with combined noise and reverberation.

However, a fundamental advantage of the STMI over the STI is its sensitivity to *joint* spectro-temporal modulations, and hence its detection of distortions that are inseparable along the temporal and spectral dimensions. For example, phase distortions as in Figs. 10 and 11 severely degrade intelligibility, but do not affect substantially temporal modulations on a single channel and hence are undetectable by the STI which does not look *across* channels. We conjecture that the opposite situation occurs with a special kind of distortion called "deterministic noise" by Noordhoek and Drullman (1997). A specific example of such a distortion is the clipping of the temporal envelope of the spectrogram ("BLK NOISE"). STI tends to overestimate the detrimental effects of such a manipulation compared to human perception (Noordhoek and Drullman, 1997), presumably because it distorts the modulation waveform rather severely. However, this manipulation does not change the relative phase of the temporal modulations on different spectral channels, and hence does not substantially distort the overall shape of the spectrogram. Consequently, we hypothesize that the STMI measured from the same sentences will be less sensitive compared to the STI, in line with human results.

We have defined two versions of the STMI: a ripple-based (Eq. (6)) and a speech-based (Eq. (1)) version. While the trends in the two measures are very similar under different noise conditions, the absolute values differ somewhat especially under reverberant conditions (Figs. 7A and 8A). The speech-based STMI^{T} are analogous to STI measures based directly on speech signals (Payton and Braida, 1999). Our definition however is derived completely from the auditory model (just as for the ripple-based STMI), and interestingly, it incorporates (implicitly) some of the "weighting functions" that are introduced in the STI to remove significant artifacts. For example, one is a higher emphasis placed on modulations in the higher frequency octave bands; another is a limit on the maximum modulation frequencies considered in the STI computation. The STMI incorporates both of these weighting functions. The first arises from a spectral pre-emphasis due to the lateral inhibition stage in the auditory spectrum computation (see discussion and eq. (19) in Wang and Shamma, 1994). The second is implicit in the band-pass nature of the MTF, emphasizing mostly intermediate modulation frequencies (4–12 Hz). Most artifacts are in higher modulation rates which are de-emphasized in the overall auditory model for reasons explained in detail in (Chi et al., 1999).

Finally, we note that two versions of the STMI^{T} were defined in Eq. (1) depending on the nature of the clean templates: that of the speech signal under investigation, or generalized average template(s). It is evident from the results in Fig. 8B that testing human subjects with *meaningful words* is best modeled by the first version of the STMI^{T} (Fig. 8A versus Fig. 9A). This is understandable since templates of these (English) words are available to the subjects. We conjecture that the generalized template method (second version) is a better model of performance when using nonsense words, because no templates are available and hence the subjects would likely use ("generalized") phonemic

templates. Therefore, we hypothesize that human subject intelligibility tests with nonsense words will yield results more like those shown in Fig. 9A rather than 8A.

There are several recent technological developments that are consistent with this "spectro-temporal modulations" view of the speech signal and its intelligibility. One concerns the possibility that encoding speech in terms of its modulation rate representation may prove to be a highly efficient and robust means for encoding speech in both ultra-low bit-rate and high-fidelity communication (Atlas, 2001). Another application area concerns the utility of the scale-rate representation for filtering noise to enhance robustness of speech recognition systems. For example, by removing spectral or temporal modulations that are beyond the normal range found in speech, one may clean up and stabilize the input from a microphone or a telephone channel into a speech recognizer. This has been well demonstrated by the widely used RASTA algorithm in modern recognition systems (Hermansky and Morgan, 1994). Furthermore, recent experiments have demonstrated the remarkable perceptual robustness of highly impoverished speech, such as speech with few independent spectral bands (Shannon et al., 1995), highly reverberant conditions (Greenberg et al., 1998; Arai et al., 1996), temporally desynchronized spectral bands (Greenberg and Arai, 1998), or severely distorted acoustic waveforms (Saberi and Perrott, 1999). These results are entirely consistent with the notion that any manipulation of speech that does not disrupt significantly the integrity of its spectro-temporal modulations (in the critical range shown in Fig. 6A) is harmless to its intelligibility.

Acknowledgements

This work is supported in contract with the Southwest Research Institute, and partially by the Office of Naval Research through an MURI grant (Center for Auditory and Acoustic Research). We are grateful to Dr. Brian Zook (at Southwest Research Institute) for extensive discussions, and supplying the speech intelligibility data. We are also grateful to two extremely insightful and helpful reviews of this manuscript by Dr. Brian J.C. Moore and an anonymous reviewer.

References

ANSI, 1969. ANSI S3.5-1969, American national standard methods for calculation of the speech intelligibility index. American National Standards Institute, New York (replaced by ANSI S3.5-1997).

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1996. Intelligibility of speech with filtered time trajectories of spectral envelopes. In: Proc. ICSLP, pp. 2490–2492.

Atlas, L., 2001. Efficient sound coding using coefficients on the ambiguity plane, ICASSP 2001.

Bellamy, J.C., 2000. Digital Telephony, Wiley Series in Telecommunications and Signal Processing, third ed John Wiley & Sons, New York.

Bradley, J.S., 1986. Predictors of speech intelligibility in rooms. J. Acoust. Soc. Am. 80 (3), 837–845.

Chi, T., Gao, M., Guyton, M.C., Ru, P., Shamma, S.A., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. J. Acoust. Soc. Am. 106 (5), 2719–2732.

Dau, T., Püschel, D., Kohlrausch, A., 1996. A quantitative model of the effective signal processing in the auditory system I. Model structure. J. Acoust. Soc. Am. 99 (6), 3615–3622.

Depireux, D.A., Simon, J.Z., Klein, D.J., Shamma, S.A., 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. 85, 1220–1234.

Drullman, R., Festen, J., Plomp, R., 1994. Effect of envelope smearing on speech perception. J. Acoust. Soc. Am. 95 (2), 1053–1064.

Greenberg, S., Arai, T., 1998. Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. In: Proc. Joint Meeting of the Acoust. Soc. Am. Internat. Congr. Acoust., Seattle, pp. 2677–2678.

Greenberg, S., Arai, T., Silipo, R., 1998. Speech Intelligibility derived from exceedingly sparse spectral information. In: Proc. Internat. Conf. on Spoken Lang. Process., Sydney, December 1–4.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Houtgast, T., Steeneken, H.J.M., 1980. Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics. Acoustica 46, 60–72.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77 (3), 1069–1077.

Kowalski, N., Depireux, D.A., Shamma, S.A., 1996. Analysis of dynamic spectra in ferret primary auditory cortex I. Characteristics of single-unit responses to moving ripple spectra. J. Neurophysiol. 76 (5), 3503–3523.

Kryter, K.D., 1962. Methods for the calculation and use of the articulation index. J. Acoust. Soc. Am. 34 (11), 1689–1697.

Lee, E.A., Messerschmitt, D.G., 1994. Digital Communication, second ed Kluwer Academic Publishers, Boston.

Lyon, R., Shamma, S.A., 1996. Auditory representations of timbre and pitch. In: Auditory Computation, Volume 6 of Springer Handbook of Auditory Research, Springer-Verlag, New York, pp. 221–270.

Noordhoek, I.M., Drullman, R., 1997. Effect of reducing temporal intensity modulations on sentence intelligibility. J. Acoust. Soc. Am. 101 (1), 498–502.

Payton, K., Braida, L.D., 1999. A method to determine the speech transmission index from speech waveforms. J. Acoust. Soc. Am. 106 (6), 3637–3648.

Saberi, K, Perrott, D.R., 1999. Cognitive restoration of reversed speech. Nature 398, 760.

Shamma, S.A., 1998. Methods of Neuronal modeling. In: Spatial and Temporal Processing in the Auditory System, second ed. MIT Press, Cambridge, MA, pp. 411–460.

Shamma, S.A., Chadwick, R., Wilbur, J., Morrish, K., Rinzel, J., 1986. A biophysical model of cochlear processing: intensity dependence of pure tone responses. J. Acoust. Soc. Amer. 80 (1), 133–144.

Shannon, R., Zeng, F.G., Wygonski, J., Kamath, V., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Steeneken, H.J.M., Houtgast, T., 1979. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Amer. 67 (1), 318–326.

Wang, K., Shamma, S.A., 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Trans. Speech Audio Process. 2 (3), 421–435.

Wang, K., Shamma, S.A., 1995. Spectral shape analysis in the central auditory system. IEEE Trans. Speech Audio Process. 3 (5), 382–395.

Yang, X., Wang, K., Shamma, S.A., 1992. Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38 (2), 824–839 (Special issue on Wavelet Transforms and Multiresolution Signal Analysis).