Technical University of Denmark

Alexandre Chabot-Leclerc

ANALYZING THE ROLE OF SPECTRO-TEMPORAL MODULATIONS FOR SPEECH PERCEPTION

Master's Thesis, August 2011

DTU Electrical Engineering
Department of Electrical Engineering

ALEXANDRE CHABOT-LECLERC

ANALYZING THE ROLE OF SPECTRO-TEMPORAL MODULATIONS FOR SPEECH PERCEPTION

ANALYZING THE ROLE OF SPECTRO-TEMPORAL MODULATIONS FOR SPEECH PERCEPTION

THIS THESIS WAS PREPARED BY
Alexandre Chabot-Leclerc

SUPERVISORS
Torsten Dau
Søren Jørgensen

Department of Electrical Engineering
Centre for Applied Hearing Research (CAHR)
Technical University of Denmark
Ørsteds plads building 352
DK-2800 Kgs. Lyngby
Denmark

http://www.dtu.dk/centre/cahr/
Tel: (+45) 45 25 39 32
E-mail: cahrinfo@elektro.dtu.dk

Project period:	January 17, 2011 – August 15, 2011
Category:	1 (public)
Edition:	First
Comments:	This report is part of the requirements to achieve the Master of Science in Engineering (M.Sc.Eng.) at the Technical University of Denmark. This report represents 35 ECTS points.
Rights:	© Alexandre Chabot-Leclerc, 2011

ABSTRACT

The speech-based spectro-temporal modulation index (STMI^{T}; Elhilali *et al.*, 2003) is analyzed to study the necessity of using spectro-temporal modulation filters to predict intelligibility. The performance of the STMI^{T} is investigated by comparing predictions to data for three distortions applied to noisy speech: reverberation, phase jitter, and spectral subtraction.

The predictions exhibit decent agreement with the data in the noisy reverberant condition, but the differences to similar experiments suggest that different prediction-to-intelligibility mapping functions might be required depending on the speech material. Data gathered for the phase jitter condition with a high resolution of the distortion parameter show that phase jitter affects intelligibility in a non-monotonic way. The model accounts well for phase jitter and predicts all tendencies in the data. The STMI^{T} fails in the spectral subtraction condition, similarly to the speech transmission index (STI), predicting an increase in intelligibility with increased over-subtraction factor. The analysis of the internal representation of the auditory process model shows that all distortions affect both the spectral and the temporal modulation domains.

It is showed that, in the framework of the STMI^{T}, some degree of spectral modulation selectivity is necessary, but temporal modulation frequency selectivity is not. It is also demonstrated that the spacing between modulation filters can be increased without affecting the predictions. Finally, it is argued that spectro-temporal modulation filters might not be crucial to predict intelligibility and that the STMI^{T} metrics is not suited to predict intelligibility for speech processed by spectral subtraction, as well as for other distortions for which modulation levels increase with the distortion parameter.

ACKNOWLEDGMENTS

Copenhagen, August 2011,

I would like to thank my supervisors, Torsten and Søren, for helping me put the pieces together and for sending me into deeper space when I had been orbiting around the same problems for too long. Innumerable sincere "thank yous" to all my friends from the Acoustics department, who have made this thesis, and master's degree, much more than an academic experience; without them, this thesis would likely have been possible, but nowhere as enjoyable. I also would like to mention a redheaded girl, who would probably rather not be mentioned: thank you. And a final thanks to my family, who would probably have wanted me to stay closer to them but who, nonetheless, helped me get this far.

This thesis has been supported by the Natural Sciences and Engineering Research Council of Canada and the *Fonds québécois de la recherche sur la nature et les technologies*.

Alexandre Chabot-Leclerc,

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	articulation index
ANOVA	analysis of variance
CLUE	conversational language understanding evaluation
DRT	Dynamic Rhyme Test
EPSM	envelope power spectrum model
HI	hearing-impaired
LIN	lateral inhibitory network
mAFC	m-alternative forced choice
MTF	modulation transfer function
NH	normal-hearing
RT	reverberation time
sEPSM	speech-based envelope power spectrum model
SII	speech intelligibility index
SNR	signal-to-noise ratio
SNR_{env}	envelope power signal-to-noise ratio
SSN	speech-shaped noise
STFT	short-term Fourier transform
SRT	speech reception threshold

STI	speech transmission index
sSTI	speech-based speech transmission index
STMI	spectro-temporal modulation index
STMI^{R}	ripple-based spectro-temporal modulation index
STMI^{T}	speech-based spectro-temporal modulation index
STRF	spectro-temporal response field
ΔSRT	Difference between a given SRT and a reference SRT

1 INTRODUCTION

Humans are incredibly adept at understanding speech, even in the most adverse conditions. Large amounts of research have been done to identify the key aspects responsible for our performance, in order to better understand the auditory system but to make more accurate predictions. Early models of speech intelligibility, like the articulation index (AI; ANSI-S3.5, 1969) or the speech intelligibility index (SII; ANSI-S3.5, 1997), considered that the important features of speech were found in the spectral domain. If these features were masked, intelligibility would decrease. These models account well for static distortions, like additive noise (French and Steinberg, 1947).

A later model, the speech transmission index (STI; IEC, 2003), considered that preservation of the temporal information was crucial. The STI looks at the temporal modulation transfer function (MTF), i.e. how much the modulation depth is kept intact after the distortion. This model accounts well for static noises, as well as for reverberation (Houtgast *et al.*, 1980; Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985), but has trouble with nonlinear processing that affects both the temporal *and* the spectral domain, like envelope compression (Rhebergen *et al.*, 2009), phase jitter (Elhilali *et al.*, 2003), or spectral subtraction (Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007). To extend the relevance of the speech transmission index (STI) to nonlinear processing, Payton and Braida (1999) and Goldsworthy and Greenberg (2004) have attempted to replace the wide-band noise probe of the original STI by a speech probe signal, those alternative models are called speech-based speech transmission index (sSTI) models.

A recent model, proposed by Jørgensen and Dau (2011), called the speech-based envelope power spectrum model (sEPSM), considered the signal-to-noise ratio (SNR) in the temporal modulation domain, rather than the temporal MTF, as the relevant metric to predict intelligibility. In addition to additive noise and reverberation, the model was shown to predict the effect of spectral subtraction on intelligibility.

The spectro-temporal modulation index (STMI; Elhilali *et al.*, 2003) builds on ideas of the previous two model and regards the preservation of the *joint* spectro-temporal modulations as the key factor. The STMI builds on top of a cortical process model, by Chi *et al.* (1999), which extracts the spectro-temporal modulation content from the auditory spectrogram. The STMI has two variants, one based on wide-band spectro-temporally modulated signals, the STMI^{R}, analogous to the traditional STI, and the other speech-based, the STMI^{T}. Both versions have been shown to account for additive noise, reverberation, phase jitter and phase shifts (Elhilali *et al.*, 2003).

In this thesis, STMI^{T} predictions for reverberation, phase jitter and spectral subtraction will be studied. Modifications to the underlying model of the STMI will be made in order to evaluate the necessity of using a two-dimensional, rather than single-dimensional, modulation filter bank. Observations will also be made on the importance of the metric noise when predicting intelligibility.

Chapter 2 covers the concepts of modulation in speech and describes the distortions used in this study. In Chapter 3, the STI, the STMI and sEPSM are detailed. The experimental details and results are presented in Chapters 4 and 5, respectively. In Chapter 6, an analysis of the internal representation of the model, as well as of the impact of spectro-temporal modulation selectivity on predictions, is done. Chapter 7 presents a discussion on the performance of the speech-based spectro-temporal modulation index (STMI^{T}) and the necessity of using spectral and temporal modulation filters. Finally, Chapter 8 summarizes the main findings of this study.

2 CHARACTERISTICS OF SPEECH

2.1 SPEECH INTELLIGIBILITY

Speech intelligibility is defined as how well a message is understood once it has passed through a communication channel. Most intelligibility tests express "how well" the speech is understood by the receiver by scoring the recognition of some units of speech: phonemes, syllables, words or sentences. Reasons for choosing a scoring unit rather than the other depend on the material, the aspect of speech under testing, the test duration, etc.; scoring for phonemes provides more data for a similar test length but might not reflect realistic situations for speech communication.

The intelligibility results can be presented as function of a signal-to-noise ratio (SNR) between the clean speech and a masker, or as a function of the parameter of a given distortion. When plotted as a function of the SNR, the intelligibility curves usually have a sigmoid function shape. Such a curve, describing the relationship between a parameter of a physical stimulus and a human response to that stimulus, is called a *psychometric function*. Figure 2.1 shows two of these functions: the left one tells us that the condition under test was easier than for the right one, since for similar SNRs, the intelligibility score (proportion, or percent, of correct answer) is higher. The point where the psychometric function crosses the 50 % point is called the speech reception threshold (SRT). One can compare two different conditions by subtracting the SRT of a particular condition with the reference SRT, giving a ΔSRT: a positive ΔSRT means the condition was more difficult than the reference, a negative ΔSRT means the condition was easier. The ΔSRT, however, does not give information on the slope of the psychometric function.

2.2 MODULATIONS IN SPEECH

2.2.1 *Temporal modulations*

Modulations in speech are easily observed in a spectrogram. In Fig. 2.2, temporal modulations are the variations in energy, in a given frequency band, across time. Temporal modulations arise from the onsets/offsets of speech as well as from the amplitude fluctuations between syllables, words and sentences.

These temporal variations can be quantified by extracting the temporal envelope of the speech, typically defined as the modulus of the

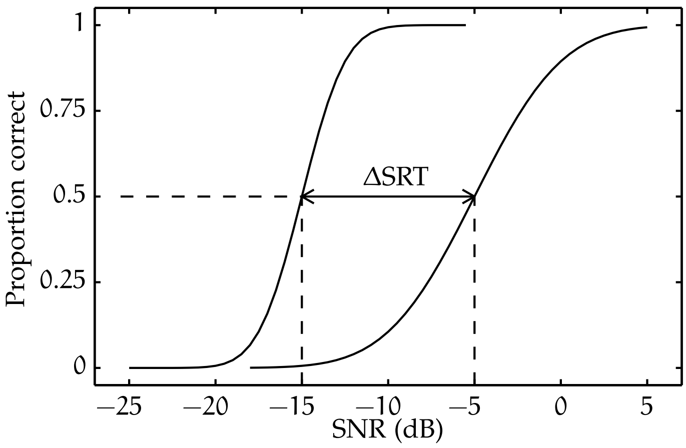

Figure 2.1: Two psychometric functions with different slopes and SRTs. The ΔSRT is shown as the difference in SNRs yielding a proportion correct of 0.5.

corresponding analytic signal (see Appendix A). Figure 2.3 shows the temporal envelope of a clean and a noisy speech signal, notice how the depth of modulations is reduced in the noisy signal. One can apply the Fourier transform to the temporal envelope to obtain the temporal modulation spectrum. It is also possible to apply a similar process to filter ed audio-frequency bands. In speech, most of the energy of this spectrum is concentrated in the modulations frequencies between 0.25–10 Hz, with a peak between 3–4 Hz (Payton and Braida, 1999). Temporal modulations are thus essential for speech intelligibility (Houtgast *et al.*, 1980; Drullman *et al.*, 1994).

2.2.2 *Spectral modulations*

Spectral modulations represent the variations in energy across the audio-frequency dimension (see Fig. 2.2). They could be called the "envelope of the audio-frequency spectrum". Low frequency spectral modulations emerge mainly from formant peaks and dips; high frequencies arise from the harmonics.

To compute the spectral modulation spectrum of a signal, three steps are needed: first, the frequency spectrum of the signal is obtained, second, the envelope of the frequency spectrum is extracted and third, the spectrum of the envelope is computed using, for example, the Fourier transform. It is also possible to apply a similar process to the short-term frequency spectrum of a signal, making it possible to see how the spectral modulation spectrum changes over time. The spectral modulation spectrum of a signal is a function of *scales*, or densities, expressed in cycles per octave (cyc/oct). Most of the spectral modulation energy of speech is concentrated in the scales below

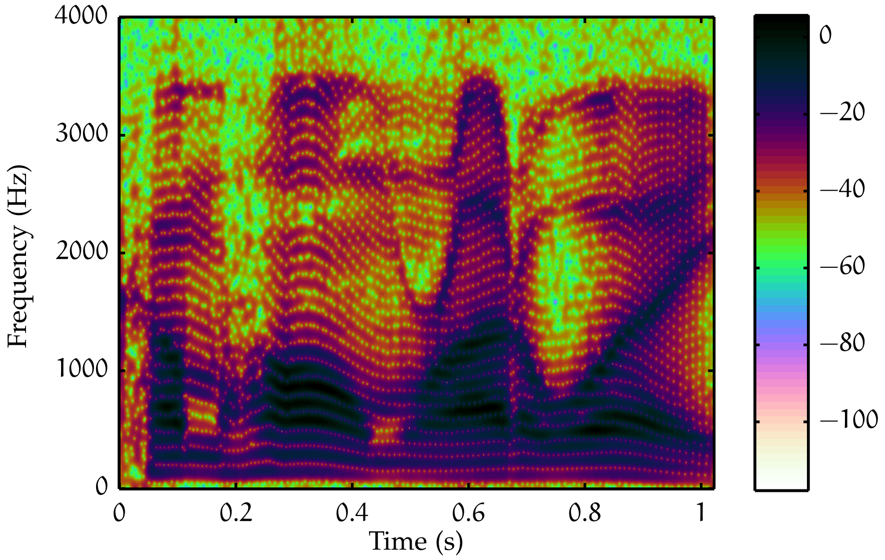

Figure 2.2: Spectrogram of the sentence /come home right away/.

4 cyc/oct (Chi *et al.*, 1999). Sensitivity to spectral modulation is highest in the scales between 0.25 and 2 cyc/oct (Chi *et al.*, 1999).

2.3 DISTORTIONS AFFECTING SPEECH INTELLIGIBILITY

A number of processes can affect speech intelligibility, additive noise and reverberation being two of the most common ones. These distortions can be linear, like the two just mentioned, or nonlinear, like peak-clipping, center clipping, compression, or two of the distortions studied in this thesis: phase jitter and spectral subtraction.

2.3.1 *Phase jitter*

Phase jitter is a common problem affecting regular telephone channels, due to fluctuations in power supply voltages (Lee and Messerschmitt, 1994, p. 164). This distortion has, nowadays, less impact on daily communications than it had a number of years ago, but it does provide valuable information on the features of speech important for intelligibility, because the effect of phase jitter is to completely destroy the carrier of the signal but to keep the temporal envelope mostly intact. It is expressed as:

$$r(t) = \Re\left\{s(t)e^{j\Theta(t)}\right\} = s(t)\cos(\Theta(t)), \tag{2.1}$$

where $s(t)$ is the clean signal, $r(t)$ is the received signal and $\Theta(t)$ is the phase jitter function, which is a random process uniformly distributed over $[0, 2\alpha\pi] (0 < \alpha < 1)$.

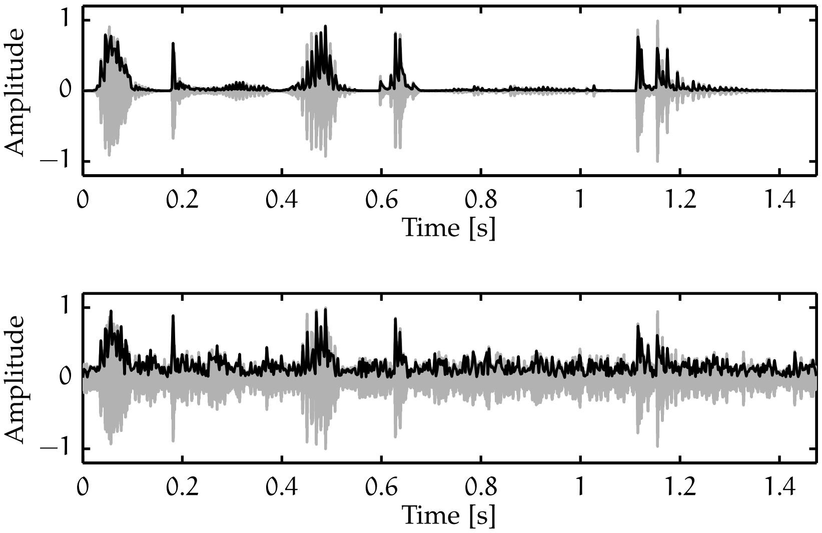

Figure 2.3: Temporal envelope of clean (top) and noisy speech (bottom).

EFFECT OF PHASE JITTER ON A SIGNAL The impact of phase jitter on the frequency spectrum does not follow a monotonic function; it becomes more severe with α going from 0 to 0.5, then has a local minimum at $\alpha = 0.75$ and is as severe at $\alpha = 1$ as at $\alpha = 0.5$. Values of $\alpha = 0.5$ and 1 let the random variable $\Theta(t)$ go between 0–π and 0–2π, respectively. With these values of α, the signal becomes an amplitude modulated white noise. Figure 2.4 shows this effect on the spectrum of a 50 Hz sine wave distorted by phase jitter.

With other α values, the jitter has an asymmetrical effect on the transmitted signal values; values below 0.25 do not produce a sign change, whereas values above 0.25 do but with a weighting toward negative or positive values. This behavior can be clearly seen in Fig. 2.5, where the phase jitter is applied to a constant signal with an amplitude of 1. Figure 2.6 illustrates this effect as well, but on the waveform of a 50 Hz sinusoidal signal.

Applying phase jitter on a more complex signal, like speech, shows how it affects the temporal and spectral properties. Figure 2.7 shows the spectrogram of a sentence distorted by phase jitter with values of α going from 0 to 1. In the clean signal, most of the energy is concentrated at low frequencies. As α goes from 0 to 0.5, the energy content at all frequencies becomes comparable while still being modulated in the time domain as the clean signal (Figure 2.7 (center)). When the signal is completely distorted in this way, the spectral modulations, due to formant structure and harmonics in clean speech, are flattened out. This can be observed in Fig. 2.8, showing the 1/3-octave long-term average audio spectra of speech distorted by phase jitter. The spectrum becomes flatter as α increases from 0 to 0.5 where it becomes

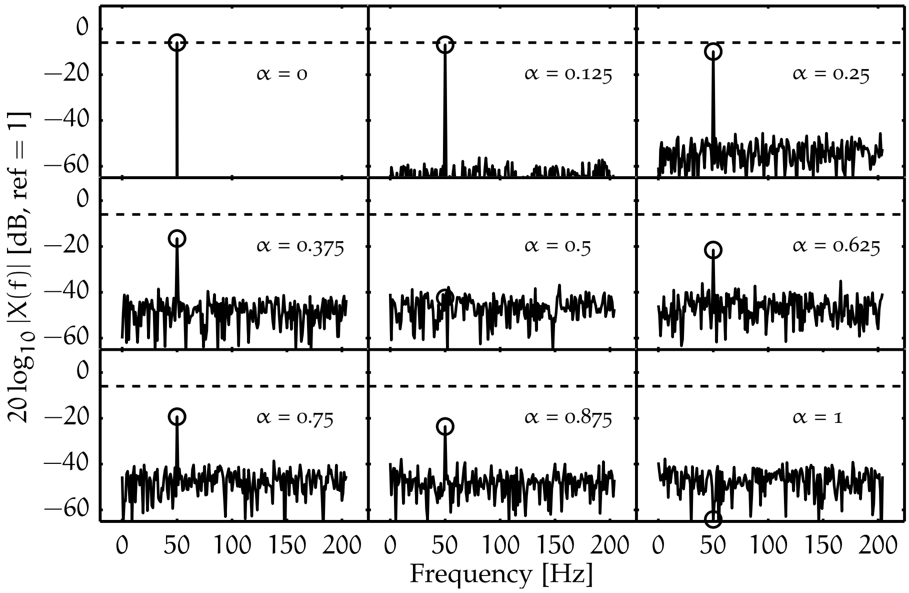

Figure 2.4: Spectra of a 50 Hz sine wave to which phase jitter has been applied. The dashed line represents the level of the uncorrupted sinusoid.

completely flat. The same thing happens at $\alpha = 1$. There is a local "maximum" of variation at $\alpha = 0.75$, where the spectrum is less flat than at 0.625 and 0.875, and very similar to the spectrum at $\alpha = 0.375$.

Fig. 2.9 shows the effect of phase jitter on the temporal modulation envelope of a 15 s sample of speech. The solid line represents the 1/3-octave temporal modulation spectra of the complete signal and the various dashed lines represent the modulation spectra of three 1/3-octave bands, centered at 0.1, 1 and 3 kHz. All spectra are normalized to their maximum value to show how the spectra of all 1/3-octave bands become similar to the envelope spectrum of the complete signal when $\alpha = \{0.5, 1\}$.

2.3.2 *Spectral subtraction*

Spectral subtraction is a method used to enhance a speech signal, $s(k)$, corrupted by uncorrelated additive noise, $n(k)$ (Berouti and Schwartz, 1979); the idea is to subtract short-term estimates of the noise, typically 10–50 ms (Berouti and Schwartz, 1979; Lim, 1978) from the spectra of the noisy signal, leaving a "clean" signal.

Considering the contaminated signal sample, $y(k)$, described as:

$$y(k) = s(k) + n(k), \tag{2.2}$$

taking the Fourier transform on both sides yields:

$$Y(j\omega) = S(j\omega) + N(j\omega), \tag{2.3}$$

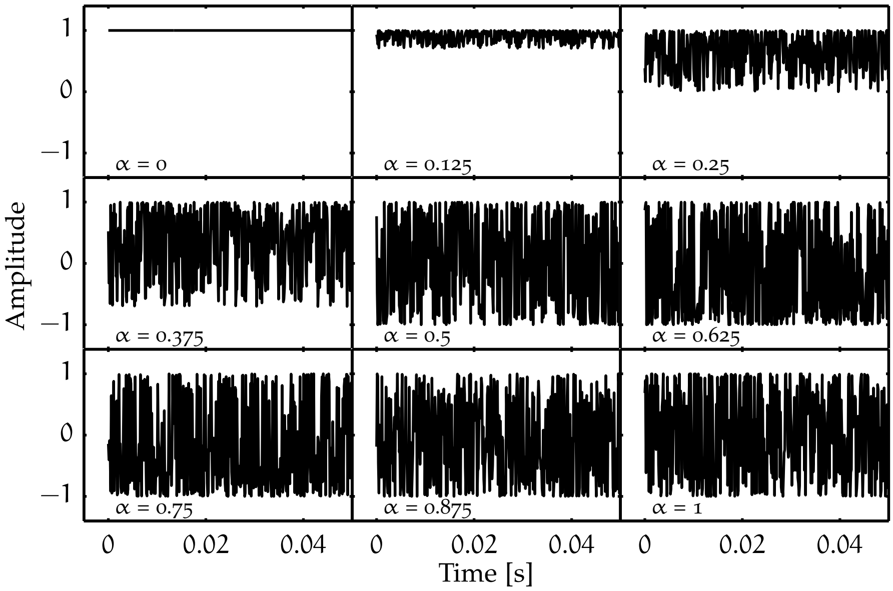

Figure 2.5: Temporal signal of constant amplitude equal to 1 with different values of phase jitter applied to it.

where $Y(j\omega)$, $S(j\omega)$ and $N(j\omega)$ represent the spectra of the noisy speech, the clean speech and the noise, respectively. In the spectral subtraction, the estimate of the noise magnitude spectrum, $|\hat{N}(j\omega)|$, is subtracted from the magnitude spectrum of the noisy speech, $|Y(j\omega)|$, and the result is combined with the original noisy-speech phase, $\theta_x(j\omega)$. The noise estimate can be obtained from the known statistical characteristics of the noise or as the expected value, $E[|N(j\omega)|] = \mu(j\omega)$, calculated for non-speech sections of the noisy sample. The estimate of the clean speech spectrum, $\hat{S}(j\omega)$, is calculated as:

$$|\hat{S}(j\omega)| = [|Y(j\omega)|^{\beta} - \mu(j\omega)^{\beta}]^{1/\beta} \cdot e^{j\theta_x(j\omega)}, \tag{2.4}$$

where β is a constant (Lim, 1978). Typical values of β are $\beta = 1$ (Boll, 1979), yielding a subtraction done in the magnitude spectrum domain and $\beta = 2$ (Berouti and Schwartz, 1979), yielding a subtraction performed in the power spectrum domain. The enhanced time signal $\hat{s}(k)$ is obtained by applying the inverse Fourier transform of the estimated clean speech spectrum:

$$\hat{s}(k) = \mathcal{F}^{-1}\{\hat{S}(j\omega)\}. \tag{2.5}$$

The spectral errors, resulting from the subtraction of the estimate of the noise:

$$e(j\omega) = N(j\omega) - \mu(j\omega)e^{j\theta_x(j\omega)}, \tag{2.6}$$

can be positive of negative. Negative values of the magnitude or power spectrum are set to zero, because they do not carry any physical

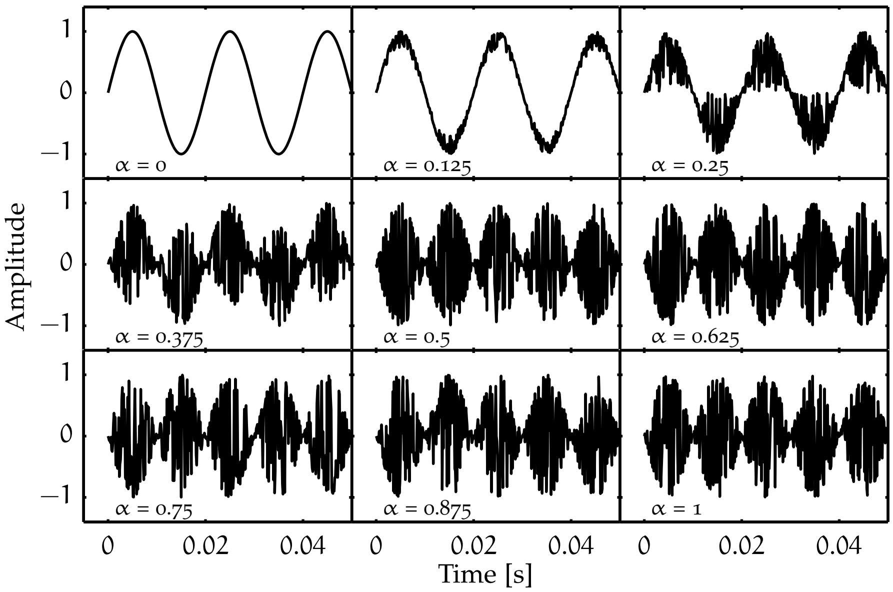

Figure 2.6: Waveforms of a 50 Hz sinusoid distorted by different values of phase jitter.

meaning. This residual error causes spectral artifacts often referred to as "musical noise". One of the solutions to this residual noise problem is proposed by Berouti and Schwartz (1979). The inclusion of the *over-subtraction* factor κ in Eq.(2.4) yields:

$$|\hat{S}(j\omega)| = [|Y(j\omega)|^{\beta} - \kappa\mu(j\omega)^{\beta}]^{1/\beta} \cdot e^{j\theta_x(j\omega)}, \tag{2.7}$$

where κ allows for the reduction of the error by subtracting a larger estimate of the noise. If the over-subtraction factor is too large, however, some of the speech spectral content is removed, resulting in distortion.

Previous studies on spectral subtraction have shown that, although the perceived quality of the processed speech is increased (Boll, 1979), speech intelligibility is not necessarily increased as well. Itoh and Mizushima (1997) found an increase in intelligibility for hearing-impaired (HI) subjects for car-cabin noise and telephone noise. Similarly, Tsoukalas *et al.* (1997) noticed improvements up to 40 % for normal-hearing (NH) listeners. In both cases, increases in intelligibility due to spectral subtraction were larger for low SNR than for large ones. On the contrary, Lim and Oppenheim (1979) and Ludvigsen *et al.* (1993) found no improvement in intelligibility. Likewise, Boll (1979) reported no increase in intelligibility when using the Dynamic Rhyme Test (DRT) in helicopter noise.

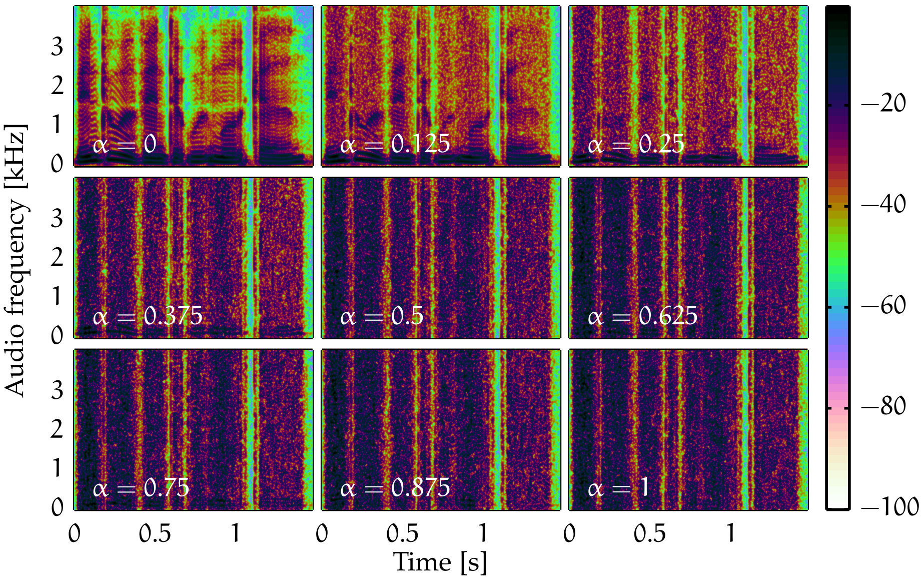

Figure 2.7: Spectrograms of a sentence distorted by phase jitter, with 9 different values of α.

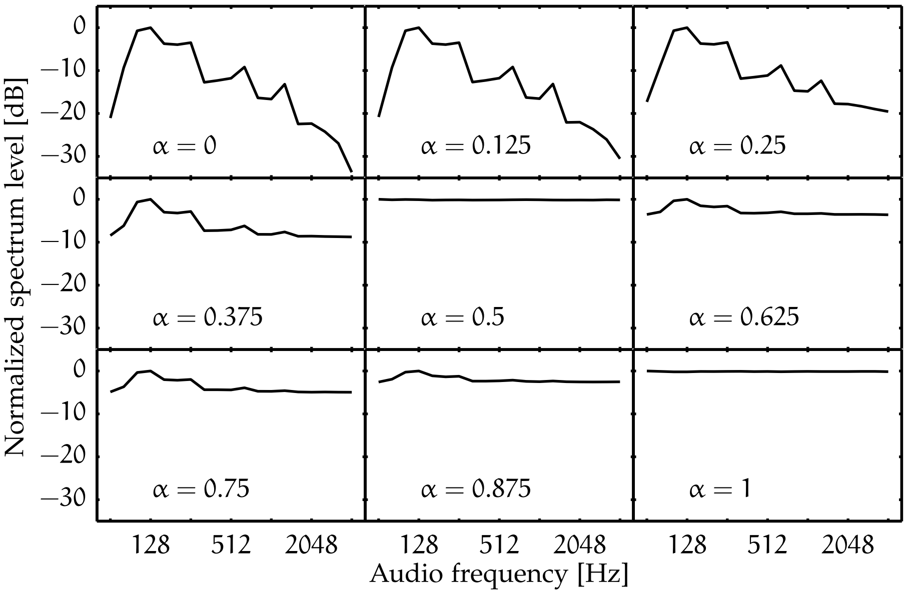

Figure 2.8: Third-octave long-term average audio spectra of speech distorted by phase jitter. The levels are normalized to the maximum value of each spectrum.

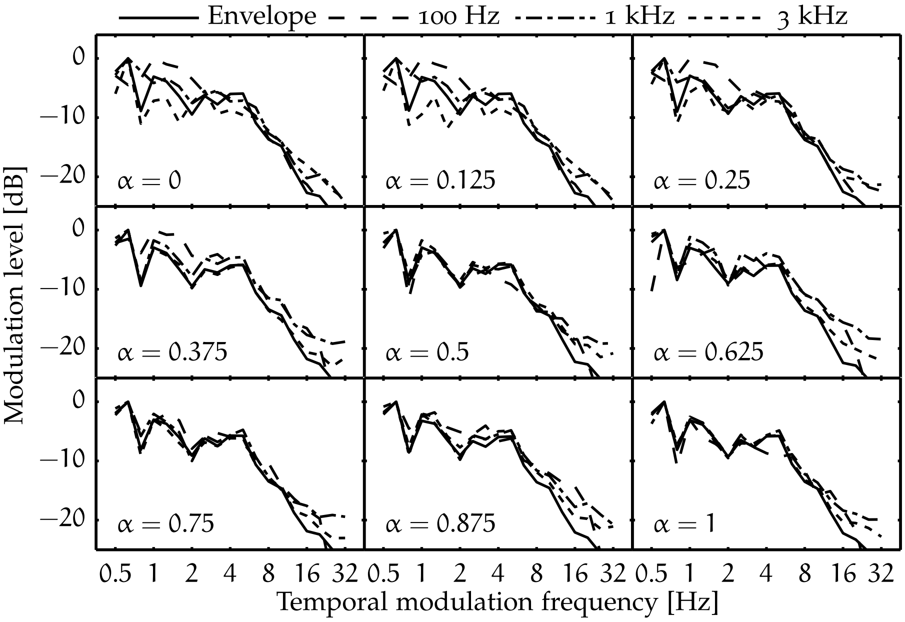

Figure 2.9: One-third octave envelope spectra of the complete signal (solid) and of filtered 1/3-octave bands of a 15 s sample of speech distorted by phase jitter (center frequencies of 0.1, 1 and 3 kHZ, various dashed lines). All spectra are normalized to their maximum value to show how the spectra of all 1/3-octave bands become similar to the envelope spectrum of the complete signal when $\alpha = \{0.5, 1\}$. The 1/3-octave band audio-frequency filtering is applied *after* the distortion.

3 MODELS OF SPEECH INTELLIGIBILITY

Various methods have been considered to analyze speech intelligibility. The earlier models considered mostly the spectral properties of speech, where it was hypothesized that the main factor affecting intelligibility was the presence of spectral noise masking the speech. One of them, the articulation index (AI; French and Steinberg, 1947; ANSI-S3.5, 1969) considers a weighted sum of signal-to-noise ratios—between a long-term average spectrum of speech and that of a noise—across a number of frequency bands; if equal-width one-third octave-band auditory filters are considered, the largest weights are given to the 1.6 and 2 kHz bands.

The speech intelligibility index (SII; ANSI-S3.5, 1997) is based on the AI but includes a number of modifications, including corrections for upward spread of masking and high presentation levels. These two models, however, failed to make accurate predictions for temporal distortions like reverberation.

3.1 THE SPEECH TRANSMISSION INDEX (STI)

Houtgast and Steeneken (1973, 1985); Houtgast *et al.* (1980) used a different approach and considered the temporal modulations as a key factor affecting speech intelligibility; they defined the speech transmission index (STI). The STI is based on the concept of the (temporal) modulation transfer function (MTF) (Houtgast and Steeneken, 1973), which is the measure of how much of the temporal envelope of the signal, at any audio frequency, is preserved after the speech has been distorted.

3.1.1 *The traditional STI*

In the traditional version of the STI (Houtgast *et al.*, 1980), the reference signal is a noise with a long-term average spectrum similar to that of speech; it is divided in 7 octave-wide bands, equally spaced on a logarithmic scale between 125 Hz and 8 kHz. The envelope of each band is, in turn, modulated at 14 different modulation frequencies between 0.63 and 12.5 Hz, with a modulation depth of 1; yielding 98 combinations of audio and modulation frequencies. The MTF is obtained for each combination of frequency band k and modulation

Table 3.1: Weight given to each audio frequency band in the STI, from (Houtgast *et al.*, 1980).

Freq. [Hz]	125	250	500	1 k	2 k	4 k	8 k
Weight, w	0.1129	0.143	0.114	0.114	0.186	0.171	0.143

frequency F by first measuring the modulation index $m(F, k)$ and then converting it to a signal-to-noise ratio:

$$\mathrm{SNR}_{F,k} = 10 \log_{10} \left(\frac{m}{1-m} \right) \text{ dB}. \tag{3.1}$$

Second, the SNR is truncated to values in the [-15, 15] dB range and the 14 SNRs for a given modulation frequency are averaged:

$$\overline{\mathrm{SNR}_k} = \frac{1}{14} \sum^{\mathrm{SNR}} \mathrm{SNR}_{F,k}. \tag{3.2}$$

Third, the $\overline{\mathrm{SNR}_k}$ are summed with the corresponding weights of Table. 3.1:

$$\overline{\mathrm{SNR}} = \sum w_k \overline{\mathrm{SNR}_k}. \tag{3.3}$$

The weights from Houtgast *et al.* (1980) emphasize the important frequency bands of speech, in an analogous way as in the AI. Finally, the $\overline{\mathrm{SNR}}$ is normalized to a value between 0 and 1, which is the STI:

$$\mathrm{STI} = \frac{\overline{\mathrm{SNR}} + 15}{30}. \tag{3.4}$$

Houtgast *et al.* (1980) mentioned that the STI was not very different when computed from a smaller number of modulation-frequency bands—e.g. only six 1/3-octave bands in one-octave intervals between 0.5 and 16 Hz—while keeping all 7 audio bands, as long as the boundaries of the modulation frequencies were shifted symmetrically relative to the ones suggested originally. This suggests that, although modulation frequency *selectivity* is important, the actual *number of filters* might not to be crucial in order to predict intelligibility, as long as the filters cover the appropriate range of frequencies.

3.1.2 *The speech-based STI (sSTI)*

Houtgast and Steeneken (1985) first suggested that one could use speech as the test signal from which to derive the MTF. It would, however, have the drawback of making it difficult, in the case of non-stationary noise, to separate the desirable modulation of speech from fluctuations of the background noise (Houtgast and Steeneken, 1985). Using speech as the probe would allow to predict the effect of different speaking styles on speech intelligibility or to make predictions when applying nonlinear signal processing. For example, Goldsworthy and Greenberg (2004) showed that it might be possible to use a speech-based, modified version of the STI, generally known as speech-based speech transmission index (sSTI), to predict intelligibility for envelope thresholding and spectral subtraction. Payton and Braida (1999) showed that it was indeed possible to use speech as the probe for reverberant, noise, and reverberant-noisy conditions.

In the speech-based STI, the basic idea is to obtain the modulation indices by comparing the original probe modulation spectrum, $X(f)$, and the noisy modulation spectrum, $Y(f)$, as expressed by Houtgast and Steeneken (1985):

$$m(f) = \frac{|Y(f)|}{|X(f)|}. \tag{3.5}$$

This expression is to be used in Eq. (3.1) and the rest of the procedure follows with Eqs. (3.2) to (3.4). Goldsworthy and Greenberg (2004) provided a review of other methods that produced qualitatively appropriate modulation indices for nonlinear operations (envelope thresholding and spectral subtraction), but did not validate, through comparison with data, that these methods could accurately predict intelligibility for these, and other, nonlinear operations.

3.2 THE SPECTRO-TEMPORAL MODULATION INDEX (STMI)

The spectro-temporal modulation index (STMI; Elhilali *et al.*, 2003) can be considered an extension of the STI. Instead of considering the temporal MTF only, the STMI uses a spectro-temporal MTF to predict intelligibility. The STMI has two variants: the speech-based spectro-temporal modulation index (STMI^T) (the *T* superscript stands for *template*) uses clean speech as the reference; the STMI^R is analogous to the traditional STI, as it uses narrow-band carriers with specific spectro-temporal modulation frequencies (called *ripples*, hence the “R”) to compute the spectro-temporal MTF. In this report, only the speech-based STMI will be considered because it can be more easily compared to other speech-based speech intelligibility models like the speech-based envelope power spectrum model (sEPSM) (Jørgensen and Dau, 2011). Furthermore, this index is much faster to compute.

The STMI^T computation employs an auditory model in two stages and an integration stage which computes the STMI^T from the auditory model output. The first stage models the early auditory system and transforms the acoustic signal into a representation called the *auditory spectrogram*. The second stage analyses the auditory spectrogram in order to extract the spectral and temporal modulation content using a bank of spectro-temporally selective modulation filters (Chi *et al.*, 1999). The inspiration for the spectro-temporal filter comes from the response characteristics of neurons observed in the mammalian primary auditory cortex (Depireux *et al.*, 2001; Kowalski *et al.*, 1996), which will be discussed later.

3.2.1 *Simluation of the early auditory processing*

The early stage of the auditory model consists of a sequence of three operations, depicted in Fig. 3.1. In the first operation, the audio signal is filtered through a filter bank consisting of highly asymmetric bandpass constant-Q filters $(Q = 4)$, equally spaced on a logarithmic axis. The filter bank covers 5 octaves and has a density of 24 filters/oct. This filtering corresponds to "place filtering" of the basilar membrane. It is an affine wavelet transform of the acoustic signal.

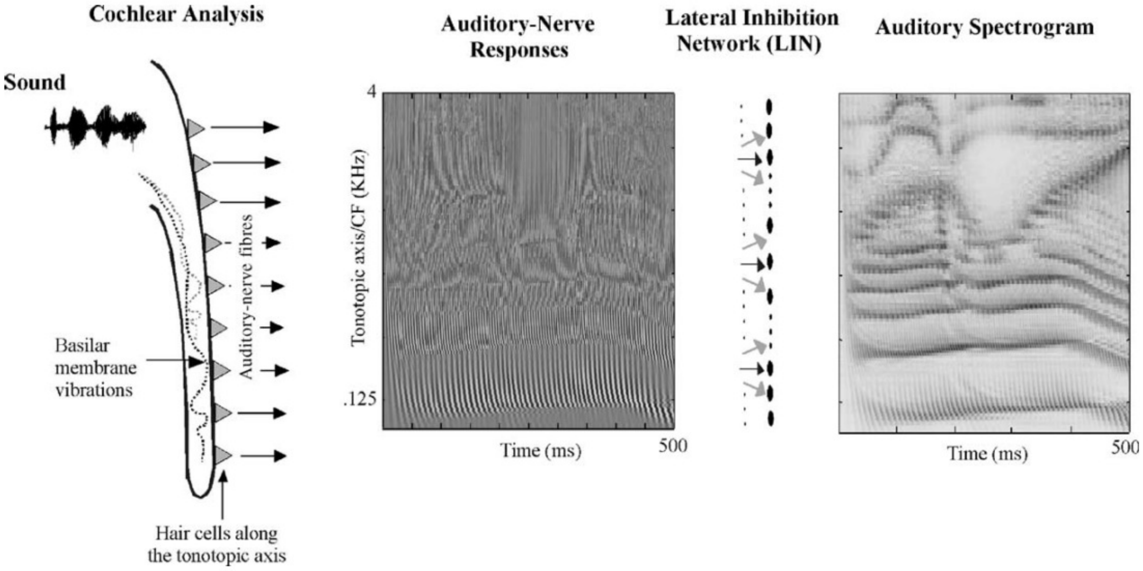

Figure 3.1: Schematics of the early auditory model. The incoming sound is analyzed by a model of the cochlea, consisting of a bank of bandpass filters (left panel). Each filter output is half-wave rectified and low-pass filtered by an inner hair cell model to produce the auditory-nerve pattern (center panel). A spacial first-difference operation is then applied to the auditory-nerve response representation, performing the function of a lateral inhibitory network (LIN). It sharpens the spectral representation of the signal and extracts its formants and harmonics. Finally, the responses of each channel is smoothed by a short-term integrator. The output is named the *auditory spectrogram* (right). The figure is from Elhilali *et al.* (2003).

In the second operation, each filter output is converted to innerhaircell outputs (intra-cellular potentials) with a three-step process: high-pass filtering (fluid-cilia coupling), instantaneous nonlinear compression (gated ionic channels) and low-pass filtering (haircell membrane leakage). Details on the mechanisms involved in each step can be found in Lyon and Shamma (1996); Shamma *et al.* (1986)

In the third operation, the auditory-nerve response goes through a lateral inhibitory network (LIN); it detects discontinuities in the response along the tonotopic axis, in a similar way as a LIN in the retina, which increases contrast and thus facilitate edge detection (Marr and Hildreth, 1980). The LIN is modeled as a three-step process: a first difference operation across channels, a half-wave rectifier and a short-term integrator. The LIN effectively sharpens the auditory filters, going from $Q = 4$ to $Q = 12$ (about 10 % of the center frequency). The advantage of the LIN is that it provides a good frequency resolution, without sacrificing the temporal resolution. Please refer to Wang and Shamma (1994) for more details on the effect of the LIN.

This three-operation process of the early auditory system *effectively* produces a spectrogram of the speech. Note, however, that the temporal modulations are to some extent affected by this process, since the output of each audio frequency filter encodes the temporal modulations due to interactions between spectral components inside the filter's transfer range (beating). The frequencies of these modulations are limited by the bandwidth of the cochlear filters.

Chi *et al.* (2005) provided a more detailed description of the auditory processing model as well as a number of examples of how different signals (3-tones, noise, harmonics complexes and ripples) are represented on the "auditory spectrogram" of this model framework.

3.2.2 *Simulation of the central auditory processing*

In the central auditory system model, further analysis of the auditory spectrogram is performed to construct a more elaborate auditory representation. The analysis extracts the spectral and temporal modulation content of the auditory spectrogram, as illustrated in Fig. 3.2. It uses a bank of modulation-selective filters, each of them tuned to a specific spectral modulation frequency, Ω (in cyc/oct), and temporal modulation frequency, ω (in Hz) (Chi *et al.*, 1999); the spectro-temporal impulse response of these filters is called the spectro-temporal response field (STRF). An example of an STRF is shown in Fig. 3.2B, together with the result of its convolution with the auditory spectrogram to the left.

THE SPECTRO-TEMPORAL RESPONSE FIELDS (STRF) Kowalski *et al.* (1996) and Depireux *et al.* (2001) measured the response of neurons in the primary auditory cortex of ferrets to what they call moving *ripples*. In its simplest form, a ripple is a signal with a single temporal (ω) and a single spectral (Ω) modulation frequency. It has an upward (negative ω) or a downward (positive ω) sweeping spectrum. The mathematical expression of a ripple is:

$$S(x, t) = L(1 + \Delta A \sin(2\pi(\omega t + \Omega x) + \varphi)), \tag{3.6}$$

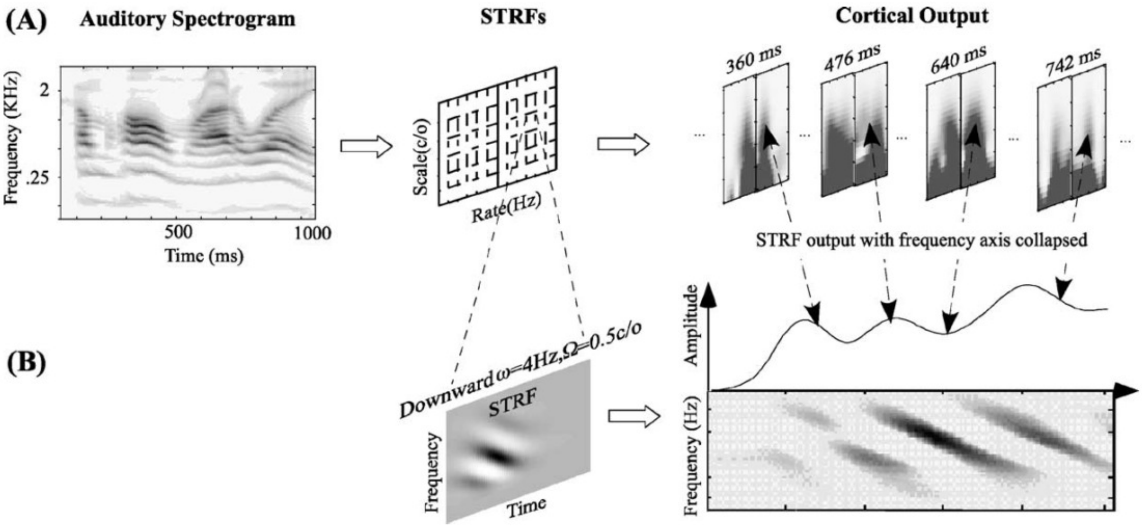

Figure 3.2: The auditory spectrogram of the sentence /come home right away/ (A, left panel) is analyzed by a spectro-temporal modulation filter bank (A, center). The STRF of one filter is shown (B, left) beside the result of the its convolution with the auditory spectrogram (B, right panel). The output is a function of time, indexed by scale-Ω, rate-ω and frequency-x. For display proposes, the frequency axis is collapsed (integrated over) to a one-dimensional time-function, as display on the top of the spectrogram in (B, right panel). The total output of the filtering by the STRFs is a series of two-dimension scale–rate plots, varying over time, as shown in (A, right panel). from Elhilali *et al.* (2003)

where L represents the stimuli presentation level, ΔA is the modulation depth, t is time and x is the tonotopic axis, defined as $x = \log_2(f/f_0)$, with f_0 being the lower bound of the spectrum, f the frequency, and φ is the phase of the ripple. Figure 3.3 shows two ripples: the left one is moving downward with a velocity of 3 Hz and a density of 0.5 cyc/oct; the right one is moving upward with a velocity of -1 Hz and a density of 2 cyc/oct.

By changing rates and scales independently, they measured the spectro-temporal transfer function of neurons, which can be transformed to the spectro-temporal response field by the means of the inverse 2D Fourier transform (see details in Depireux *et al.*, 2001). The STRF is a spectro-temporal function $\mathrm{STRF}(t, x)$. Qualitatively, the spectral axis of the STRF reflects the range of frequencies that affect the firing rate of the neuron studied and the temporal axis reflects how the firing rate changes over time. The STRF can be understood conceptually as a time-varying spectral response field, or a collection of frequency-dependent temporal impulses. The STRF is modeled using a spectro-temporal Gabor function (Chi *et al.*, 1999). Figure 3.2B (right panel) shows the result of the convolution of the auditory spectrogram with such a Gabor function.

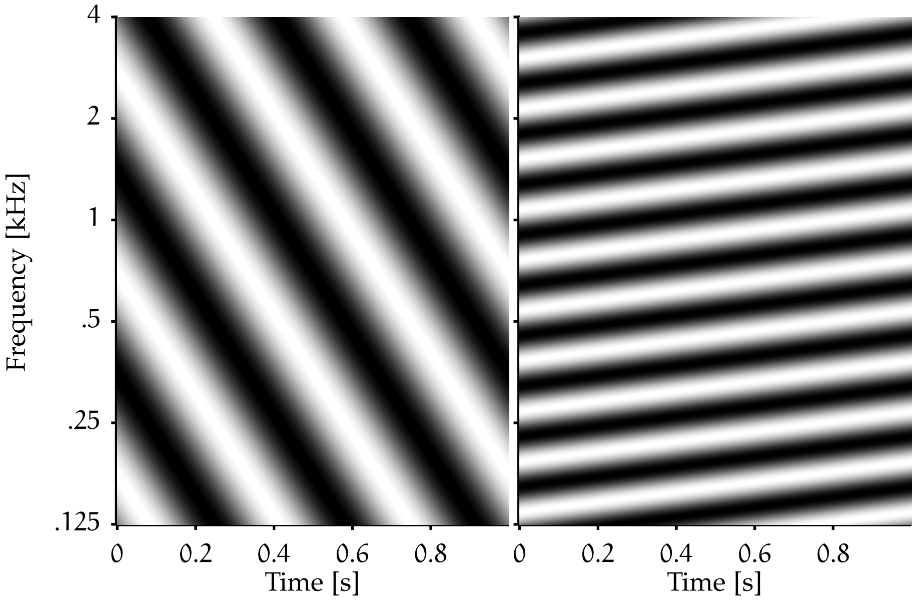

Figure 3.3: Spectrogram of spectro-temporal ripples moving downward with a rate of 3 Hz and with a spectral density of 0.5 cyc/oct (left) and upward with a rate of -1 Hz and with a spectral density of 2 cyc/oct (right). The color scale goes from black for null values, to white for maximum values.

Depireux *et al.* (2001) found that most of the measured STRFs were "quadrant separable", i.e. that *for a given direction*, they could be factorized into the product of a purely spectral and a purely temporal function. By contrast, a fully separable STRF could be factorized into two functions, independently of direction, suggesting that it is constituted of independent temporal and spectral processing stages. Quadrant separability, or more severe inseparability, implies temporally and spectrally intertwined stages of processing (Depireux *et al.*, 2001).

CORTICAL FILTERS What is referred to as the cortical representation is the output of the central auditory processing model. It is obtained by convolving the $\text{STRF}_{i,j}(t,x)$ for each cortical filter (i,j) with the auditory spectrogram $y(t,x)$

$$r(t,x,\omega,\Omega) = \int_T \|y(t) *_{t,x} \text{STRF}(t,x)\| \, dt, \tag{3.7}$$

where $*_{t,x}$ is the convolution in time (t) and multiplication in frequency (x). The resulting representation is a "four dimensional" array indexed by time, cochlear frequency axis, rate and scale $r(t,x,\omega,\Omega)$. Detailed information on the mathematical formulation of the cortical stage can be found in Chi *et al.* (2005). Figure 3.2 (A, right panel) shows "time slices", also called rate–scale plots, of the cortical representation, where the audio frequency axis has been collapsed.

Similar to the STRF, the cortical representation can be understood in a number of ways because of its four dimensions: it is possible to illustrate how two dimensions change by fixing the other two. One example would be to show the modulation content at any time t and audio frequency x for a fixed rate–scale pair, $(\omega_{\text{fix}}, \Omega_{\text{fix}})$. Such a representation can be easily visualized as the result of the convolution of the auditory spectrogram and the STRF of the $(\omega_{\text{fix}}, \Omega_{\text{fix}})$ pair (illustrated in Fig. 3.2 (B,right)). A second example would be to display the modulation content at every rate ω and scale Ω for a fixed time–audio frequency pair, $(t_{\text{fix}}, x_{\text{fix}})$, as illustrated (four times) in Fig. 3.2 (A,right).[1] The first representation is easier to understand since it results from a direct mathematical formulation: the convolution. Meanwhile, the second approach is not as intuitive since there is no "single-step" process to reach it: the auditory spectrogram must be convolved with every STRF before it is possible to have access to the spectro-temporal "slices" (the rate–scale plots) of the cortical representation.

3.2.3 *Computation of the speech-based STMI (STMI^T)*

The computation of the STMI^T (Fig. 3.4) is done individually for each sentence, assuming that their duration is short enough for the statistics of the stimuli to be considered stationary, but long enough to extract the slow temporal modulations (Elhilali *et al.*, 2003). A duration of about two seconds was used in Elhilali (2004). The stimuli are first normalized to have zero-mean and standard deviations of one before adding or applying additional distortion. The following steps are then carried out independently for the clean and the distorted speech. First the input signal, $g(t)$, is converted to the auditory spectrogram representation, $y(t, x)$. Second, the auditory spectrogram is convolved with each STRF (c.f. Eq. (3.7)), yielding the four-dimensional cortical representation, $r(t, x, \omega, \Omega)$. It is then integrated over the complete sample duration, leaving a three dimensional template of the speech token, $\{T(x, \omega, \Omega)\}$, similarly for the noisy token, $N(x, \omega, \Omega)$. Third, the "base" spectrogram of the signal, $r_0(x, \omega, \Omega)$, is subtracted from $r(t, x, \omega, \Omega)$, effectively normalizing the outputs of model. The "base" spectrogram results from the processing through the auditory model of a stationary noise with the same spectrum as the long-term average spectrum of its corresponding token (clean or noisy speech).

At this stage, the STMI^T can now be computed as:

$$\text{STMI}^{T} = 1 - \frac{||T - N||^2}{||T||^2}, \tag{3.8}$$

[1] Note, however, that the audio frequency dimension in the figure has been collapsed and not fixed, but the result is a rate–scale plot nonetheless.

where $||\cdot||$ is the euclidian distance between the cortical model's output of the noisy token and clean templates, computed as:

$$||T|| = \sqrt{\sum_k \sum_i \sum_j (T(x_k, \omega_i, \Omega_k).} \tag{3.9}$$

The algorithm of the STMI^{T} is provided in Appendix B.

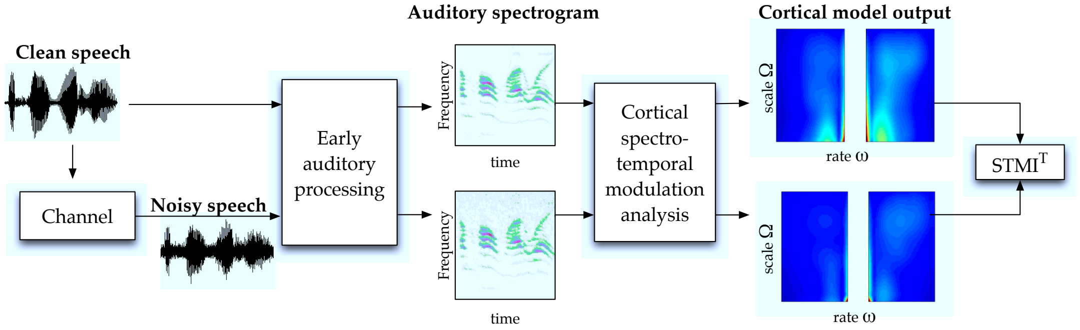

Figure 3.4: Block diagram of the STMI^{T}. The clean and noisy speech are given as inputs to the auditory model. The cortical model outputs are normalized by the base spectrum, as explained in the text. The cortical representation is then used to compute the STMI^{T}. Reproduced from Elhilali *et al.* (2003).

Elhilali *et al.* (2003) have shown that the STMI^{T} predictions are linearly linked to data, i.e. an STMI^{T} of zero equals 0 % correct and an STMI^{T} of one equals 100 % correct.

3.3 THE SPEECH-BASED ENVELOPE POWER SPECTRUM MODEL (SEPSM)

The speech-based envelope power spectrum model (sEPSM; Jørgensen and Dau, 2011) uses the envelope power signal-to-noise ratio (SNR$_{\text{env}}$) as the metric for predicting speech intelligibility; it is based on the envelope power spectrum model, originally developed for amplitude modulation detection and masking (EPSM; Dau *et al.*, 1999; Ewert and Dau, 2000). The sEPSM hypothesizes that reduction in speech intelligibility of distorted speech is due mainly to the intrinsic fluctuations in the envelope of the noisy waveform. It was shown to predict ΔSRTs accurately for speech in stationary noise, reverberant noisy speech and noisy speech processed by spectral subtraction (Jørgensen and Dau, 2011).

The first stage of the model is a band-pass filter bank made of 22 fourth order gammatone filters, equally spaced on a logarithmic scale at each 1/3-octave between 63 Hz and 8 kHz. Only filters with output energy above the hearing threshold are considered for further processing. The envelope of each filtered signal is extracted using the Hilbert transform. The resulting envelope functions are then input to a modulation filter bank consisting of a third-order low-pass filter with a cutoff frequency of 1 Hz and 6 overlapping, octave-spaced, second-order bandpass filters with center frequencies from 2 to 64 Hz, having a constant Q-factor of 1. For each modulation filter output, the ac-coupled envelope power is calculated by integrating the envelope power density in the filter's transfer range. A threshold, representing some internal noise, sets the lower limit of the envelope power to −20 dB; any filter output with a lower level is set equal to this threshold.

The sEPSM assumes that the model has access to an estimate of the noise alone (N) in addition to the noisy speech ($S+N$), thus the envelope power spectrum of the noise alone ($P_{\text{env},N}$) and noisy speech ($P_{\text{env},S+N}$) are available at the model's modulation processing stage. An estimate of the clean speech envelope power is obtained by subtracting the noise envelope power from the noisy speech envelope power:

$$\hat{P}_{\text{env},S} = P_{\text{env},S+N} - P_{\text{env},N}. \tag{3.10}$$

The SNR_{env} is calculated by taking the ratio between the estimated clean speech envelope power and the noise envelope power:

$$\text{SNR}_{\text{env}} = \frac{P_{\text{env},S+N} - P_{\text{env},N}}{P_{\text{env},N}} = \frac{\hat{P}_{\text{env},S}}{P_{\text{env},N}}, \tag{3.11}$$

where the envelope power of the noise is always lower or equal to the envelope power of the noisy speech, such that the denominator of Eq. (3.11) never becomes negative:

$$P_{\text{env},S+N} = \max\{P_{\text{env},S+N}, P_{\text{env},N}\} + \epsilon, \tag{3.12}$$

where ϵ is a small positive constant which prevents the numerator of Eq. (3.11) to be zero in the case $P_{\text{env},S+N} = P_{\text{env},N}$.

The 7 × 22 SNR_{env} values—7 modulations filters and 22 gammatone filters—are integrated across channel, with an optimal linear combination which, for n channels is expressed as

$$\text{SNR}_{\text{env}} = \left[\sum_{i=1}^{n} (\text{SNR}^{*}_{\text{env},i})^2\right]^{1/2}. \tag{3.13}$$

The seven SNR_{env} are integrated first, producing 22 single values of SNR_{env} that are combined in the same manner.

Lastly, the SNR_{env} are converted to percent correct by a process involving an "ideal observer", by first converting the SNR_{env} to a sensitivity index d':

$$d' = k \cdot (\text{SNR}_{\text{env}})^q, \tag{3.14}$$

where k and q are constants, independent of speech material and experiment conditions. The ideal observer represents an m-alternative forced choice (mAFC) decision model combined with an unequal-variance Gaussian model; it assumes a different probability for correctly recognizing a speech item, designated as the target distribution, and for failing to correctly recognized the speech, designated as the noise distribution. The sensitivity index is linked to the percent correct by

$$P_{\text{correct}}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \tag{3.15}$$

where Φ designates the cumulative normal distribution, μ_n denotes the mean of the noise distribution, and σ_S and σ_N are the standard deviation of the target and noise distributions, respectively. The values of σ_N and μ_N are determined by the number of alternatives, m, which can be adjusted. The value of σ_S, which it is inversely proportional to the slope of the ideal observer's psychometric function, can also be adjusted to account for different degrees of redundancy in the speech materials.

4 METHOD

This chapter presents the experimental details for three experiments conducted to evaluate the performance of the STMI^T for different distortions. The first experiment tests the performance for noisy reverberant conditions, the second experiment involves speech-shaped noise (SSN) and phase jitter, and the third experiment considers spectral subtraction.

4.1 EXPERIMENT 1: REVERBERATION

In this experiment, STMI^T predictions are compared to speech intelligibility data from Jørgensen and Dau (2011), considering noisy speech presented in reverberant conditions.

4.1.1 *Speech material*

The experiment uses the stimuli from the conversational language understanding evaluation (CLUE) test (Nielsen and Dau, 2009); the sentence material is in Danish and is composed of excerpts from Danish newspapers; the sample rate is 44.1 kHz. Each sentence has five words, each word less than four syllables and the number of syllables per sentence is 8–9. Sentences are grouped in lists of ten; each list has an overall SRT of -3.15 dB. The deviation in SRT between lists is less than 0.5 dB. The CLUE test allows a number of variations in the responses: (1) change in verb tenses, (2) change in article, and (3) change between singular and plural nouns.

4.1.2 *Stimuli and apparatus*

The distorted stimuli were obtained by mixing the clean sentences from the CLUE material with SSN and convolving the results with the impulse response of a reverberant room. The impulse responses were created using the ODEON room acoustics software, version 10, (Christensen, 2007), simulating the response of a rectangular 3200 m^3 room with distributed absorption, such that the reverberation time was the same in the frequency range 63–8000 Hz. Five T_{30} were considered: 0, 0.4, 0.7, 1.3 and 2.3 seconds. The speech samples were presented using the CLUE MATLAB software created by Nielsen and Dau (2009).

4.1.3 *Subjects*

The stimuli were presented to five male and three female normal hearing subjects aged between 24 and 33 years old.

4.1.4 *Simulation details*

The 180 sentences from the CLUE test were downsampled to 8192 Hz to reduce computation time. Their average length was 1.53 s, with a standard deviation of 0.16 s. Speech-shaped noise was added to the clean speech to achieve the appropriate SNR; the noise had the same duration as the clean speech and the SNRs ranged from -9 to 9 dB in steps of 2 dB. The reverberation was applied in the same way as for the measured data. Due to the addition of low-frequency temporal modulations introduced by the convolution, the final waveforms were truncated at the beginning and the end, where the low-passed envelope was less than 5 % of the envelope maximum.[1]

The cortical model used 34—17 for positive rates and 17 for negative rates—1/3-octave wide band-pass temporal modulation filters, placed equally on a logarithmic scale at every 1/4-octave, from 2 to 32 Hz. There were 21 1/3-octave wide band-pass spectral modulation filters between 0.25 and 8 cyc/oct, placed at every 1/4-octave.

The STMI^T was computed for each sentence individually and the final result was taken as the mean across all sentences. The predicted SRT was taken as the SNR producing an STMI^T of 0.5. The ΔSRT for a given reverberation time (RT) was obtained as the difference between the SRT in the processed condition and in the reference condition ($\text{RT} = 0$).

The computation of the STMI^T was based on the code-base of the NSL Tools Package accessible at http://www.isr.umd.edu/CAAR/pubs.html. It provides a MATLAB implementation of the early auditory process and the cortical process models.

4.2 EXPERIMENT 2: PHASE JITTER

Data were collected on the intelligibility of noisy speech distorted by phase jitter. STMI^T predictions are compared to the collected data.

4.2.1 *Stimuli*

The speech-shaped noise was added prior to the phase jitter, in the same way as described in Sec. 4.1. The phase jitter was then applied to the noisy signal. The α values were $\alpha = [0, 1]$ with a step size of 0.125.

[1] The envelope of the signal was extracted using the Hilbert transform and then low-pass filtered using a forth-order Butterworth filter with a cutoff frequency of 20 Hz.

4.2.2 *Subjects*

Measurements were obtained with three normal-hearing males, aged between 24 and 26 years old. Their pure-tone thresholds were of 20 dB hearing level or better in the frequency range 0.25–8 kHz, please refer to Appendix C for their audiograms. All three subjects had previous experience with psychoacoustic measurements. All of them were native Danish speakers and students at the Technical University of Denmark. None of them was remunerated for his participation.

4.2.3 *Apparatus and procedure*

The stimuli were presented through HD580 headphones, plugged into a RME DIGI96/8 sound card, in a double-walled sound-attenuating booth. The speech level was adjusted to be constant at 65 dB SPL using a Brüel & Kjær artificial ear, type 4152, and noise was added to achieve the desired SNR before applying further processing. Each sentence was presented once with the noise starting 1 second before and ending 600 ms after; the noise was ramped on and off using 400 ms cosine ramps. The presentation software was a modified version of the CLUE MATLAB software (Nielsen and Dau, 2009), which randomly presented the selection of fixed SNRs. Each subject was presented with 18 ten-sentence lists: two lists were used for each α value and two sentences per list for each SNR, resulting in 4 data points for each (SNR,α) combination per subject. The lists and SNRs were presented in random order. The training consisted of 3 lists using α values 0, 0.25 and 0.5, also presented in random order. The subjects were asked to repeat the sentence heard and were allowed to guess. No feedback was provided.

4.2.4 *Simulation details*

The down-sampled stimuli were processed in the same way as for the measurements.

4.3 EXPERIMENT 3: SPECTRAL SUBTRACTION

In this experiment, model prediction were compared to human ΔSRT data (Jørgensen and Dau, 2011) for noisy speech processed by spectral subtraction.

4.3.1 *Psychoacoustical experiment*

In the experiment by Jørgensen and Dau (2011), the spectral subtraction algorithm was implemented using a 1024-point short-term Fourier transform (STFT) with a 24 ms window (Hanning) and a 50 % overlap. The over-subtraction factors κ used were 0, 0.5, 1, 2, 4, 8, where $\kappa = 0$ is the reference condition where no spectral subtraction is applied. The presentation setup and the subjects were the same as for the reverberation experiment (Sec. 4.1).

4.3.2 *Simulation details*

Using the stimuli sampled at 8 kHz, the SSN is first added to the speech. The spectral subtraction is then applied in the same way as for the psychoacoustical experiment, but using a 256-point STFT and a window length of 32 ms.

5 RESULTS

5.1 EXPERIMENT 1: EFFECTS OF REVERBERATION ON SPEECH INTELLIGIBILITY

Figure 5.1 shows the STMI^T as a function of the SNR and reverberation time. Each curve shows an increasing STMI^T with increasing SNR, reflecting an increase in intelligibility. For a fixed SNR, the STMI^T decreases with increasing RT, reflecting a decreasing intelligibility. The presence of reverberation tends to flatten the STMI^T curves and shift them toward lower values.

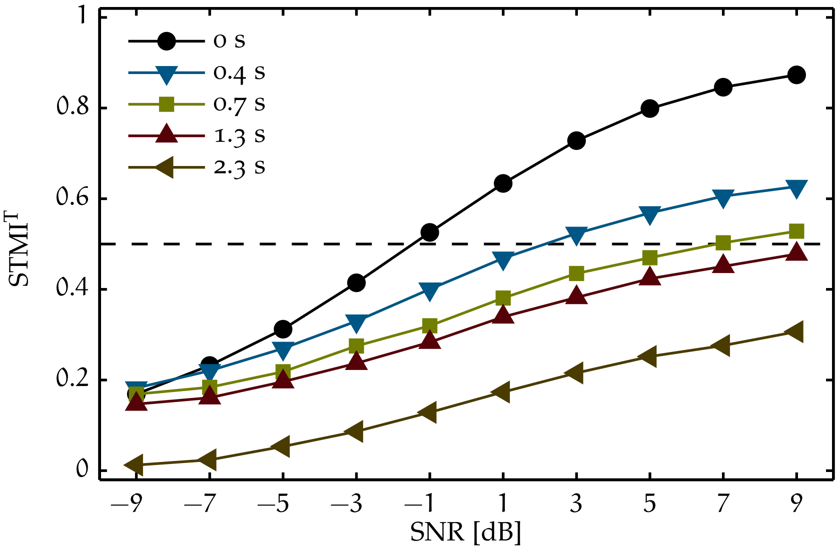

Figure 5.1: STMI^T for sentences corrupted by speech-shaped noise and reverberation. The dashed line represents an STMI^T of 0.5. Each line and symbol represent different reverberation times: ○: 0, ▽: 0.4, □: 0.7, △: 1.3 ◁: 2.3 s.

The STMI^T functions were converted to ΔSRTs in order to compare them with measured data. Figure 5.2 shows ΔSRTs as a function of the reverberation time. The model predicts the data correctly for reverberation times below 0.7 s but overestimates them for RT = 0.7 s. The STMI^T predictions could not be translated to ΔSRTs for RTs above 0.7 s because the STMI^T did not reach 0.5.

The differences between the measured and the simulated data may be due to the simulation procedure; applying the reverberation increases the duration of the signal—it inserts a short delay at the beginning and creates a long low amplitude "tail"—effectively adding low temporal modulation energy. This leads to a decrease of the clean-to-noise modulation ratio, which in turn causes a higher STMI^T (c.f. Eq (3.8)). It seems that the envelope threshold of 5 % of the maximum value was not sufficient, a more aggressive/restrictive approach might be necessary.

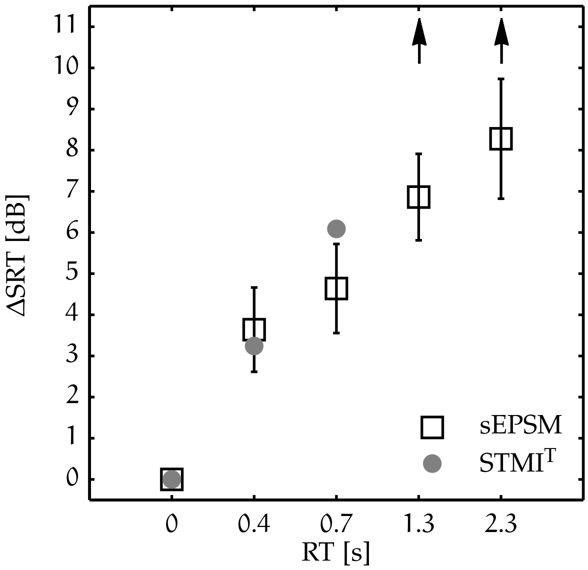

Figure 5.2: ΔSRT as a function of the reverberation time. The open squares represent the measured data and the filled circles the simulation data transformed from Fig. 5.1.

Elhilali *et al.* (2003) obtained very good agreement between measurements of phoneme recognition and simulations of the STMI^T when adding reverberation to noisy speech. The reverberation was applied by convolving the signal with Gaussian white noise having an envelope which was exponentially decaying with a reverberation constant τ (Chi *et al.*, 1999; Houtgast *et al.*, 1980).[1] On the contrary, the current simulations, using impulse responses generated with the Odeon software, provided a clearly worse agreement. This could be due to the differences between the impulse responses or to the speech material.

5.2 EXPERIMENT 2: EFFECTS OF PHASE JITTER ON SPEECH INTELLIGIBILITY

Figure 5.3 (left panel) shows the average measured data, as a function of α. The vertical bars indicate plus/minus one standard deviation. On the average data, the maximum standard deviation is 0.24. For all three subjects, intelligibility is zero for $\alpha = \{0.5, 1\}$, no matter the SNR, which is expected since for those values, phase jitter has the effect of completely replacing the carrier with white noise. The intelligibility

[1] The relation between the reverberation time and τ is RT $= 6.91\tau$ (Ratnam *et al.*, 2004).

decreases with the decreasing SNR. Maximum intelligibility is obtained for larger SNRs and when α is smaller than 0.5. A local intelligibility maximum is present at $\alpha = 0.75$ for all subjects. A three-way analysis of variance (ANOVA) of the mean data showed a significant effect of α ($F_{8,120} = 53.2$, $p < 0.001$) and SNR ($F_{4,120} = 30.0$, $p < 0.001$), but no significant difference between subjects ($F_{2,120} = 0.3$ $p = 0.73$).

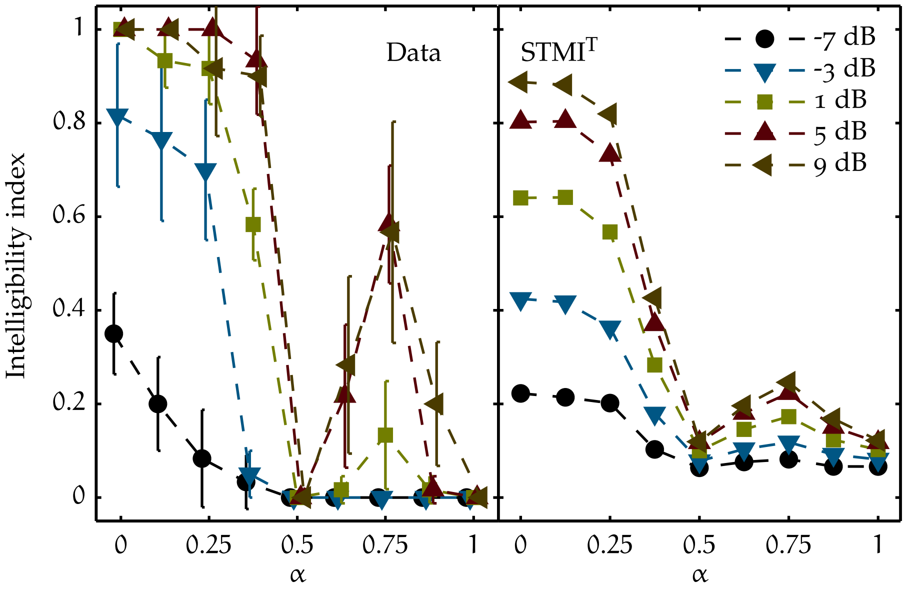

Figure 5.3: Average percent correct (left panel) for the three subjects and STMI^T predictions (right panel) for the phase jitter experiment as a function of α and with the SNR as parameter. The filled symbols are for the different SNRs, they are respectively ○: -7, ▽: -3, □: 1, △: 5 ◁: 9 dB SNR.

The right panel shows predictions from the STMI^T. They have similar characteristics than the data, with minima at $\alpha = \{0.5, 1\}$, implying that phase jitter has the strongest impact. The intelligibility maxima are when $\alpha = \{0, 0.125\}$. At these values, the intelligibility reaches zero, at all SNRs, for the subjects, whereas the predicted minima are between 0.07 and 0.13. For $\alpha \leqslant 0.25$, the intelligibility score reaches 1 for SNRs larger than zero, whereas the largest predicted value is just below 0.9. Both the data and the predictions have a local maximum when $\alpha = 0.75$, although the intelligibility is less than twice lower for the predictions than for the data when the SNR is larger or equal to 5 dB. For SNRs below 5 dB, the predictions are higher than the measured score.

In short, the STMI^T accounts reasonably well for the data. It does not consistently overestimate nor underestimate the measured speech intelligibility. This can be seen more clearly when plotting the STMI^T versus the proportion correct, as in Fig. 5.4. For values below 0.2, the STMI^T overestimates the intelligibility, where for values larger than 0.2, the STMI^T underestimates the intelligibility; this behavior is consistent for all α and SNR values.

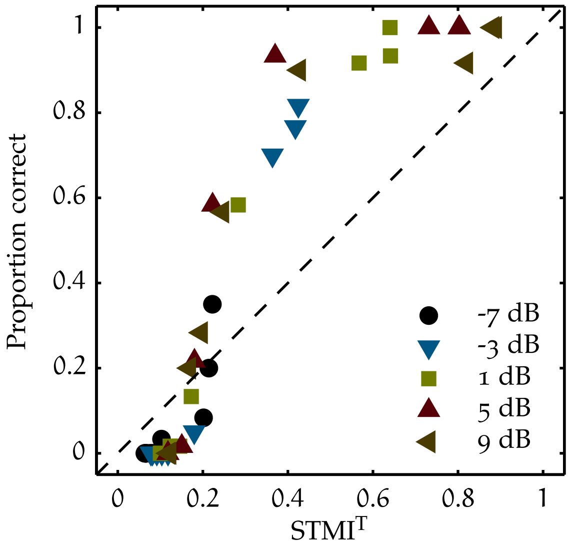

Figure 5.4: STMI^T versus proportion correct as obtained for the speech-shaped noise and phase jitter experiment. Each color (shape) is a different SNR.

5.3 EXPERIMENT 3: EFFECTS OF SPECTRAL SUBTRACTION ON SPEECH INTELLIGIBILITY

Figure 5.5 shows the predictions for noisy speech processed by spectral subtraction as a function of the SNR. The STMI^T increases with incresing value of the over-subtraction factor κ, predicting an increase in intelligibility. The STMI^T increase is more pronounced for low SNR values (increase of 0.2 at -9 dB SNR) and becomes almost zero for SNRs above 7 dB.

Converting the STMI^T values to ΔSRT provides the results shown in Fig. 5.6. Here, model predictions (circles) are compared to measured data (squares) (Jørgensen and Dau, 2011). The STMI^T fails to capture the trends in the data. The predicted ΔSRTs decrease with increasing κ, reflecting an increase in intelligibility, whereas the measured ΔSRTs increase. For $\kappa = 8$, the STMI^T predicts a ΔSRT of -3.3 dB whereas the average measured data is 2.7 dB.

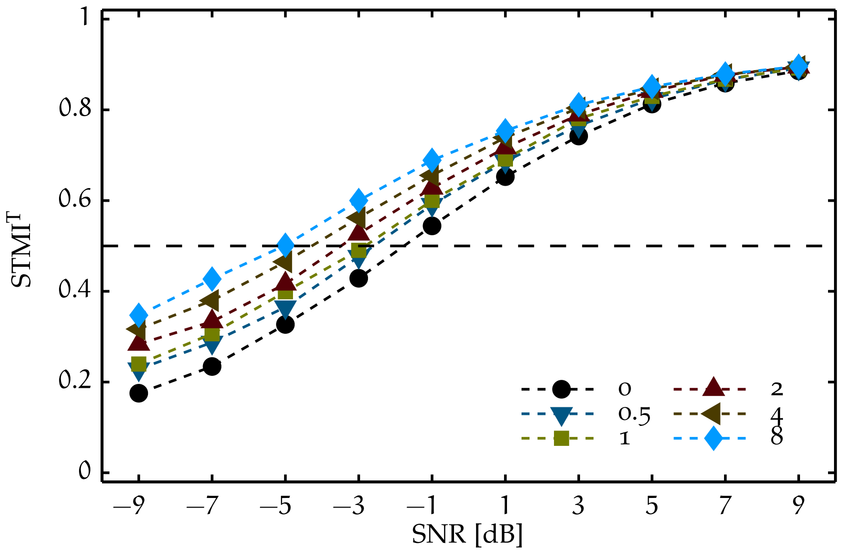

Figure 5.5: STMI^T predictions for noisy speech processed by spectral subtraction as a function of the SNR and with the over-subtraction factor, κ, as the parameter. The dashed line represents an STMI^T of 0.5. The difference symbols represent the κ values: ◯: 0; ▽: 0.5; □: 1; △: 2; ◁: 4 and ◇: 8.

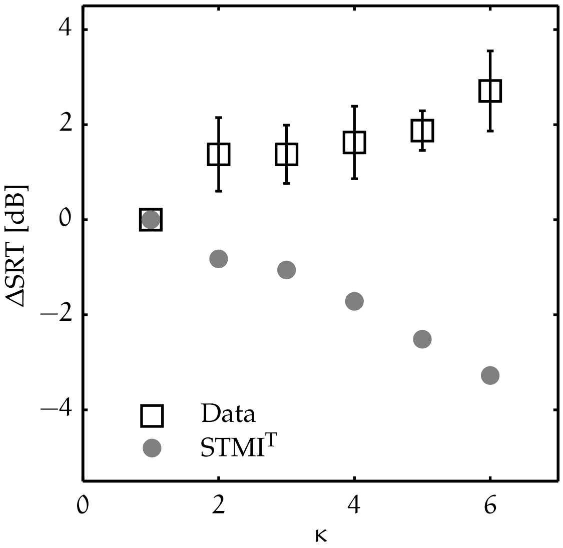

Figure 5.6: Comparison of ΔSRTs between subjects (open squares) (Jørgensen and Dau, 2011) and STMI^T (filled circles) for noisy speech processed by spectral subtraction, as a function of the over-subtraction factor κ.

6 MODEL ANALYSIS

In order to characterize and evaluate the STMI^T, two aspects of the model were investigated. The first aspect was how the various distortions affect the cortical representation. The second aspect was how temporal and spectral modulation frequency selectivity affect predictions.

6.1 REPRESENTATION OF DISTORTIONS IN THE CORTICAL MODEL

The values input to the stage calculating the STMI^T (Eq. (3.8)) are the time-integrated cortical representation, i.e. the magnitude of each spectral and temporal modulation filter output for each audio-frequency band (c.f. Sec. 3.2.3). These can therefore be called "modulation magnitudes". When the modulation magnitudes of the clean and noisy signals are similar, the predicted STMI^T is high, since it means that the spectro-temporal modulation information has been kept intact. It is possible to see exactly how modulations are affected by the various distortions by looking at this internal representation in the model.

Due to the large number of "dimensions" of the data and the shear size of the data (each speech sample generates more than 180,000 data points), only a subset of the modulation magnitudes will be presented below. Figures 6.1 to 6.4 show the temporal modulation magnitudes at rates $\omega = \{2, 4\}$ Hz and the spectral modulation magnitudes at scales $\Omega = \{0.5, 1\}$ cyc/oct. Both types of modulation magnitudes are presented at the audio-frequency band of 0.5 kHz. These rates and scales were chosen because they correspond to the highest spectro-temporal sensitibvities in the auditory model MTF (Chi *et al.*, 1999). The 0.5 kHz audio-frequency band was shown to be the one with the most energy in speech (Byrne *et al.*, 1994).

The modulation magnitude plots can be considered as a "vertical slice" of a superposition of rate–scale plots for different distortion conditions. Figure 6.1 (top) would then represent the result of superposing rate–scale plots for the 2 kHz audio-frequency band, slicing the "stack" at $\Omega = 0.5$ cyc/oct, and then looking at the "insides" of the stack.

Some plots with larger sets of rates ($\omega = \{2, 4, 8, 16, 32\}$ Hz), scales ($\Omega = \{0.25, 0.5, 1, 2, 4, 8\}$ cyc/oct) and frequencies ($f = \{0.5, 1, 2\}$ kHz) are available in Appendix D, they will prove to be useful in the comprehension of the impact of frequency selectivity, described in the next section.

6.1.1 *Effects of noise and reverberation*

It is known that the addition of noise reduces the salience of temporal modulations in speech, thus reducing intelligibility (Houtgast *et al.*, 1980). Elhilali *et al.* (2003) showed that white noise and reverberation affect the MTF of the auditory model and smear the auditory spectrogram. Figure 6.1 shows that in the internal representation of the cortical model, the temporal and spectral modulation magnitudes are reduced as the SNR decreases. Both types of modulations are affected in a similar fashion: the larger the clean speech modulation magnitude is, the larger is the reduction in magnitude due to noise. Rates (scales) with low magnitudes are less affected by the noise, since they are already small. At an SNR of -9 dB SNR, the spectro-temporal modulations are reduced to about one-quarter of the clean-speech value at the corresponding rate (scale).

Temporal modulation magnitudes of the clean speech are below zero for rates below -16 Hz and above 16 Hz, which is due to the subtraction of the base spectra (c.f. Sec. 3.2.3). It means that, for these ranges of temporal velocities, the base spectrum has more modulation energy than its corresponding token. The temporal modulation frequency at which this sign change happens is dependent on the spectral modulation frequency and audio frequency (see Appendix D for modulation magnitude plots with a larger number of rates, scales and audio-frequencies.).

The noisy temporal modulation magnitudes become higher than that of the clean speech above 15 Hz and below -15 Hz. A similar observation is made for most other distortions; the crossing point sometimes happens at a lower rate.

Asymmetry between the negative (upward) and positive rates (downward) of the temporal magnitudes are due to the accumulating phase lag of the cochlear filters (traveling wave) (Chi *et al.*, 2005). This can be understood by the fact that low-frequency content is “delayed” in the auditory spectrogram, creating a diagonal-like “pattern”, going from the upper left to the lower right corner (see Chi *et al.* (2005, their Fig. 3a) for a good example of this).

Figure 6.2 shows the spectro-temporal modulation magnitudes for noisy speech affected by reverberation. Four RTs are plotted, from 0 to 2.3 s and the SNR is set to 9 dB. In both dimensions, the magnitudes decrease as the RT becomes longer. Here, the magnitudes are not simply scaled down as the distortion increases; reverberation also affects the shape of the modulation magnitudes. Long RTs reduce the temporal modulation magnitudes at high rates more than at low ones, because reverberation acts like a low-pass filter on the temporal envelope spectrum (Payton and Braida, 1999; Dubbelboer and Houtgast, 2008). Whereas spectral modulation magnitudes are more reduced at low scales than at high scales.

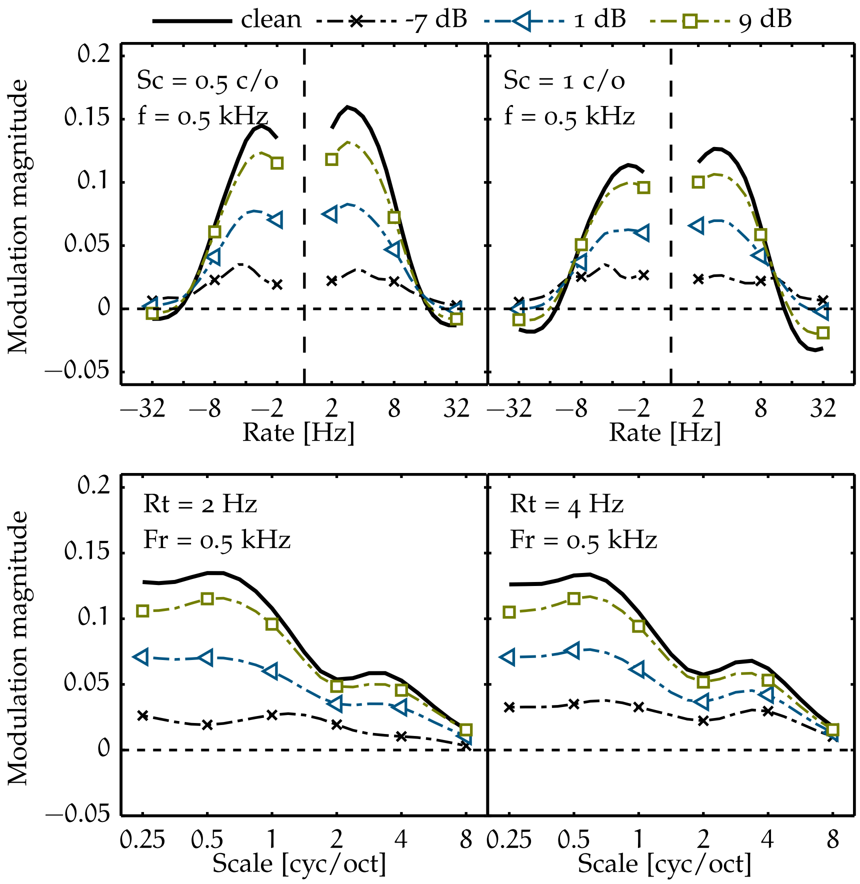

Figure 6.1: Modulation magnitudes of speech with added speech-shaped noise. (Top panels): Average temporal modulation magnitudes for clean speech and three SNRs. (Lower panels): Average spectral modulation magnitudes of positive rates for clean speech and three SNRs. Solid line: clean speech; ×: -7; ◁: 1; □: 9 db SNR.

6.1.2 *Effects of noise and phase jitter*

In order to observe the effect of phase jitter "only" on the modulation magnitude, the SNR was fixed to 9 dB.

Figure 6.3 shows the spectral and temporal modulation magnitudes for α values of 0.125 to 0.5 in steps of 0.125; values of $\alpha > 0.5$ were not shown since they produce redundant STMI^{T} values and thus redundant modulation magnitudes. Spectral and temporal modulations are affected in similar ways; they are reduced as α increases, in a similar way as the STMI^{T} versus α, i.e. the slope of the reduction is shallow for $\alpha \leqslant 0.25$ and steeper for $0.25 \leqslant \alpha \leqslant 0.5$. At $\alpha = 0.5$, there is little to no energy left. This observation does not apply for all combinations of rate, scale and frequency. There is still energy at very low scales and high audio frequencies ($\Omega = 0.25$ cyc/oct and f = 2 kHz, see Figs. D.5 and D.6)

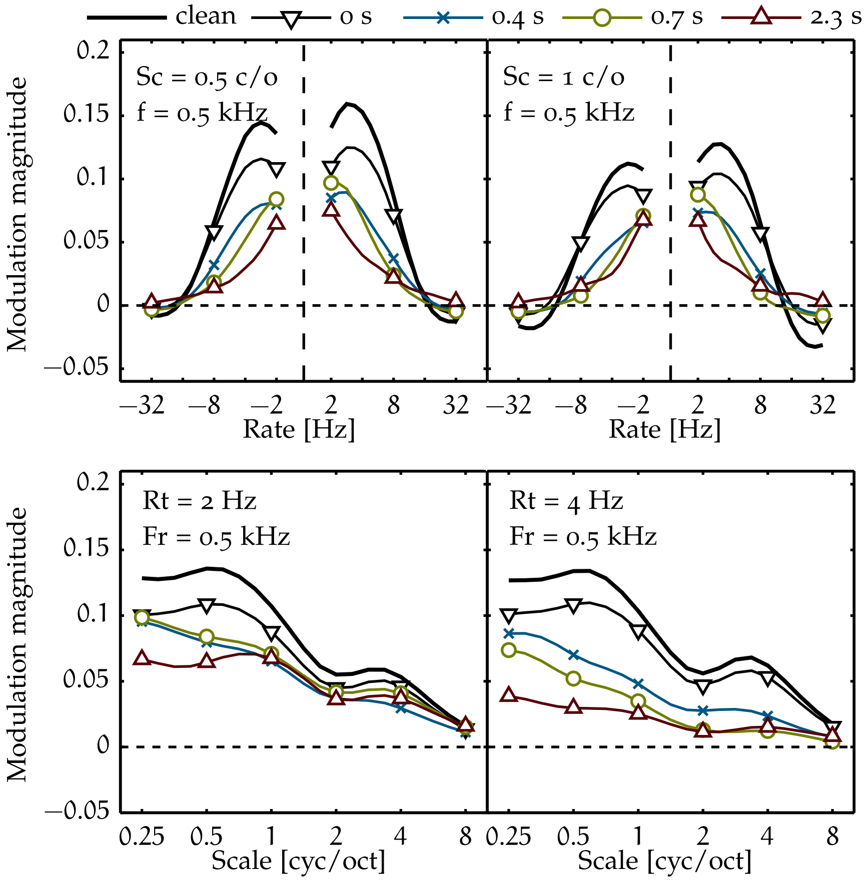

Figure 6.2: Modulation magnitudes with speech-shaped noise combined with reverberation. (Top panels): Average temporal modulation magnitudes for clean speech and four reverberation times (RTs). (Lower panels): Average spectral modulation magnitudes of positive rates for clean speech and four RTs. Solid line: clean speech; ▽: 0 s; ×: 0.4 s; ○: 0.7 s; △: 2.3 s.

The spectral modulation magnitudes are rendered flat for densities above 0.5 cyc/oct when $\alpha = 0.5$. This is consistent with the fact that applying phase jitter with $\alpha = \{0.5, 1\}$ replaces the carrier with white noise, i.e. a flat audio-frequency spectrum. This large difference between the clean and distorted values leads to a small STMI$^{\mathrm{T}}$, which is in line with data gathered with subjects.

This analysis indicates that, in fact, both the temporal and spectral modulation dimensions are affected by phase jitter. Reduction of intelligibility might thus be predicted by considering only one of these dimensions, and not necessarily by analyzing the combined modulations.

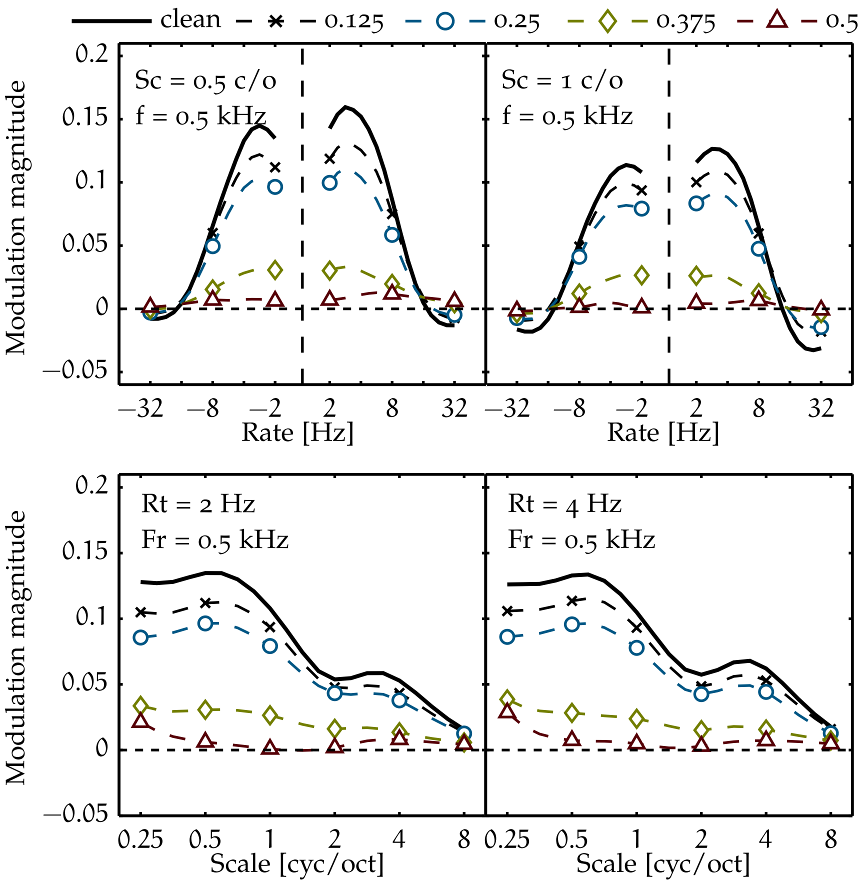

Figure 6.3: Modulation magnitudes of clean speech and speech distorted with speech-shaped noise (9 dB SNR) and phase jitter for $0.125 \leqslant \alpha \leqslant 0.5$. (Top panel) Average temporal modulation magnitudes for clean speech and four α values. (Lower panel) Average spectral modulation magnitudes of positive rates for clean speech and four α values. Solid line: clean speech; ×: $\alpha = 0.125$; ◯: $\alpha = 0.25$; ◇: $\alpha = 0.375$; △: $\alpha = 0.5$

6.1.3 *Effects of spectral subtraction*

In the case of spectral subtraction, both spectral and temporal modulations are affected. To show this, the spectro-temporal modulations were plotted for $\kappa = \{0, 1, 8\}$ at an SNR of -9 dB SNR; a small SNR value illustrates the effect of spectral subtraction in a more obvious manner, since there is "more noise" to be subtracted.

Fig. 6.4 (top) shows that, for $\kappa = 1$, there is a slight increase in the temporal modulation values compared to the situation with $\kappa = 0$ (no spectral subtraction, only noise). Processed modulation magnitudes are higher than the clean ones for rates in the range -14–10 Hz. When $\kappa = 8$, the temporal magnitudes of the processed speech are increased even further and the crossing point happens at smaller rates. Overall, spectro-temporal modulation at low audio-frequency are enhanced more than high-frequency ones (Figs. D.7 and D.8 in Appendix D).

The increase in spectro-temporal modulation magnitude with increasing over-subtraction factor κ is in line with what has been observed by Dubbelboer and Houtgast (2008). They showed that the original envelope spectrum is restored by the spectral subtraction, i.e. the temporal modulation spectrum is restored. It is incorrect, however, that the algorithm "enhances" the speech so much as to add information where there was not any originally. The STMI^T predicts an *increase* in intelligibility with increasing κ, because the spectro-temporal modulation energy increases with κ. The data, however, shows the opposite: the intelligibility *decreases* when κ increases (c.f. Fig. 5.6). Using the STMI^T metric, the cortical representation would have to show a decrease in modulation magnitude as κ increases to predict the data correctly.

6.2 EFFECTS OF MODULATION FREQUENCY SELECTIVITY ON MODEL PREDICTIONS

Houtgast *et al.* (1980, p.63) found that the STI could predict similar results with a reduced number of filters, thus reduced temporal modulation frequency selectivity, as long as the remaining filters were placed in a symmetrical fashion relative to the original recommended selection. On the contrary, Jørgensen and Dau (2011) found that frequency selectivity in both the audio- and temporal-frequency domains was essential for accurate predictions with the sEPSM.

In its original form, the STMI^T uses 21 spectral modulation filters and 34 temporal modulation filters—17 for positive rates and 17 for negative rates, only the positive rate filters are reported below although both directions were used to compute the STMI^T—where in both cases, all modulation filters are 1/3-octave band wide filters with a constant Q-factor of 1 and a spacing of 1/4-octave. A number of simulations were carried out to investigate how modulation frequency selectivity affected the STMI^T. In the first set of simulations, the spectral and temporal modulation filters were replaced by single low-pass filters with different cutoff frequencies. In the second set, the number of spectral and temporal modulation filters was reduced, one dimension at a time, while keeping the original filter bandwidth.

The simulations were done using two types of distortions in addition to the speech-shaped noise added to the speech: reverberation (because of its effect in the temporal domain) and phase jitter (because of its effect in the audio-frequency domain). Reverberation times were the same ones as for Experiment 1 (Sec. 4.1). In the case of phase jitter, the α values ranged from 0 to 0.5 in steps of 0.125. The SNRs span the range from -9 to 9 dB in 2 dB steps for both distortions. Only 45

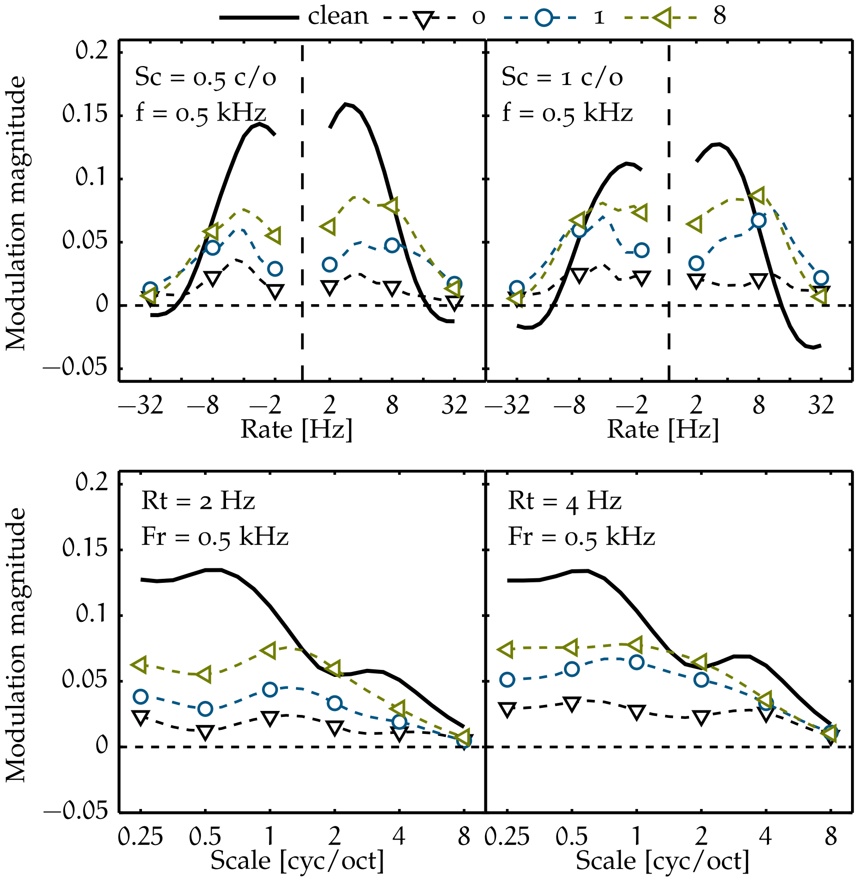

Figure 6.4: Modulation magnitudes of clean speech and speech combined with speech-shaped noise (-9 dB SNR) and spectral subtraction with over-subtraction factor $\kappa = 0$, 1, and 8. (Top panels): Average temporal modulation magnitudes for clean speech and three κ values. (Lower panels): Average spectral modulation magnitudes of positive rates for clean speech and three κ values. Solid line: clean speech, ▽: $\kappa = 0$; ○: $\kappa = 1$; ◁: $\kappa = 8$.

sentences were used to do all simulations. The reference results are replots from the results of the three experiments (Sections 5.1 to 5.3).

6.2.1 *Effects of low-pass modulation filters on STMI^T predictions copy*

This section first presents an analysis of the effects of using a single low-pass modulation filter for the reverberant noisy condition and for the phase jitter condition. The analysis when a temporal low-pass modulation filter is used is presented first and then is presented the analysis when a spectral low-pass modulation filter is used.

temporal modulation low-pass filtering Figure 6.5 shows the predictions for reverberant noisy speech (top row) and for phase jitter distorted speech (bottom row), for the unmodified model (first column) as well as for simulations where a single low-pass temporal modulation filter was used; the cutoff frequencies were 2 and 32 Hz (second and third column, respectively). The irregularity of the curves for the modified models is due to the smaller number of sentences used in the simulations.

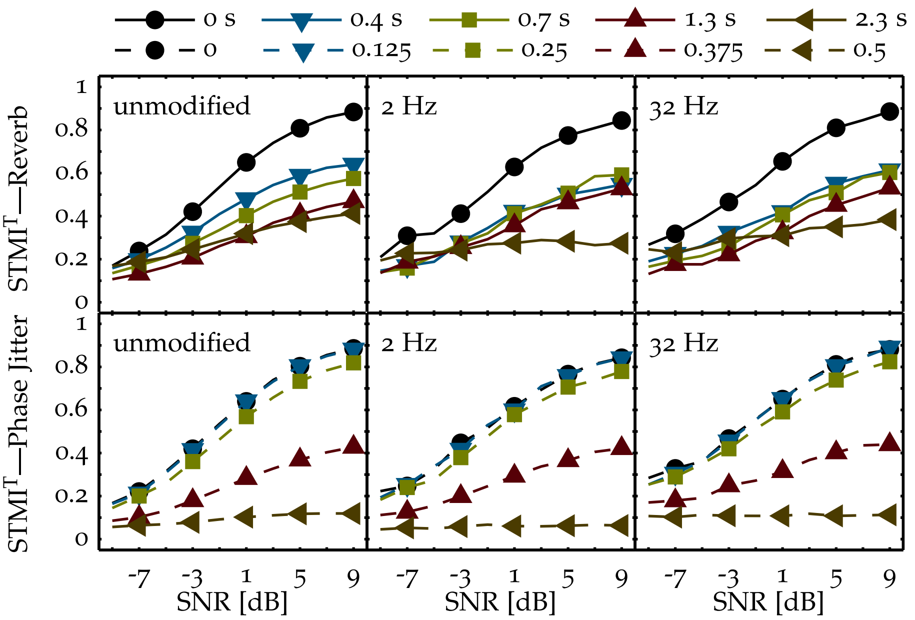

Figure 6.5: STMI^T predictions for different cutoff frequencies of the temporal modulation filters for noisy speech distorted by reverberation (top row) and phase jitter (bottom row). Unmodified model predictions with 17 (34 in total, including positive and negative rates) temporal modulation filters (first column). And predictions for single (two in total) low-pass temporal modulation filters with cutoff frequencies of 2 Hz (middle column) and 32 Hz (right-hand colum).

In the reverberant condition, the difference between predictions for short RTs are smaller when using a low cutoff frequency. When the cutoff frequency is at 32 Hz, differences are only partly restored. Then RT = 2.3 s, the predictions are slightly flatter in the modified models than in the reference one. This can be explained by the fact that reverberation acts as a low-pass temporal modulation filter, meaning that the difference between the clean and distorted modulation spectrum gets larger for increasing modulation frequency. When increasing the cutoff frequency to 32 Hz, the difference between the clean and noisy modulation energy increases, thus leading to a decrease in STMI^T in the high-SNR conditions. For low SNRs, it is the intrinsic temporal modulations of the noise that limit the STMI^T.

In the phase jitter condition (bottom row), the STMI^T for all cutoff frequencies and α are similar to the reference case. The predictions are slightly lower than the reference when the cutoff frequency is 2 Hz. A low cutoff frequency has a very small impact because the spectro-temporal modulation magnitudes of the noisy speech stay proportional to the clean speech magnitudes, independently of the temporal modulation filter frequency (see Fig. D.6), meaning that they produce the same STMI^T values.

Overall, it appears that, for both distortions, temporal modulation selectivity has only a small impact on the predictions and is not essential to predict intelligibility.

spectral modulation low-pass filtering Figure 6.6 shows the STMI^T when varying the spectral modulations filters cutoff frequency. In the reverberant conditions (top row), the modified models predictions are almost identical to the reference.

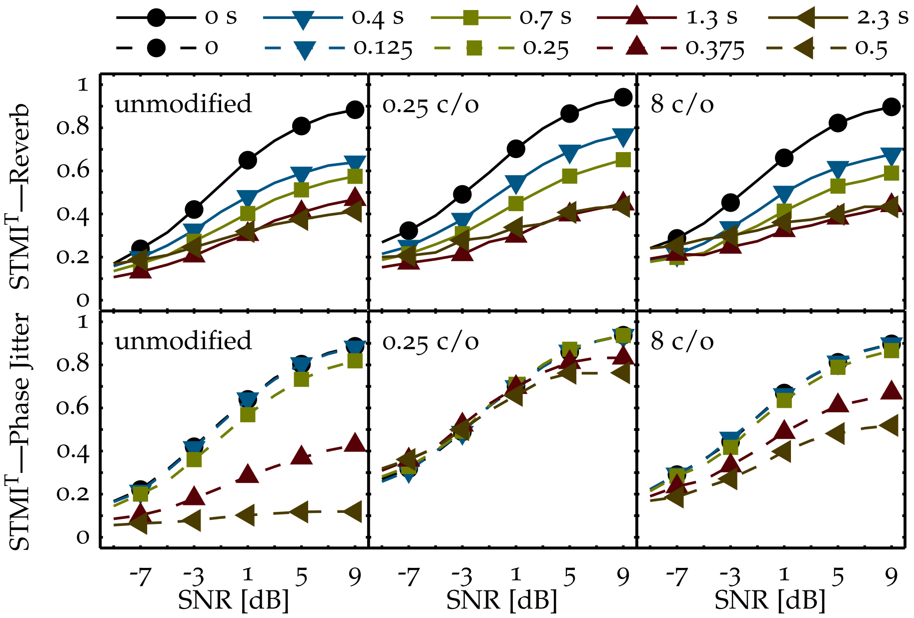

Figure 6.6: STMI^T predictions for different cutoff frequencies of the spectral modulation filters for noisy speech distorted by reverberation (top row) and phase jitter (bottom row). Unmodified model predictions with 21 modulation filters (first column). STMI^T for single low-pass spectral modulation filters with cutoff frequencies of 0.25 cyc/oct (second column) and 8 cyc/oct (third column).

In the phase jitter condition (bottom row) changing the spectral cutoff frequencies generates the same values for the reference and for the modified models for $\alpha \leqslant 0.125$. For $\alpha \geqslant 0.25$, however, a low-pass spectral modulation filter cutoff frequency of 0.25 cyc/oct renders the STMI^T almost independent of the increase in α (bottom center panel). Only at SNRs larger than -1 is the STMI^T affected by α. As the cutoff frequency is increased to 8 cyc/oct (bottom right panel), the STMI^T for $\alpha \geqslant 0.25$ decreases with increasing α but never reaches the reference values.

The insensibility of the STMI^T for low spectral cutoff frequencies can be explained by the fact that low-scale modulation magnitudes are not affected by the jitter (c.f. Sec 6.1.2 and Fig. D.5).

Considering the simulation results presented, it could be concluded that temporal modulation frequency selectivity is not critical: using a single low-pass temporal filter rather than a bank of band-pass filters does not affect the STMI^T when the distortion is phase jitter and has only a mild effect for the reverberant condition, for the largest RT. In contrast, spectral modulation frequency selectivity appears to be important, especially for the phase jitter distortion.

6.2.2 *Effects of limiting the number of band-pass modulation filters on STMI^T predictions*

This section presents an analysis of the effects on the predictions, by changing the number of spectro-temporal modulation filters, for reverberant noisy speech and phase jitter-distorted speech.

TEMPORAL MODULATION BAND-PASS FILTERING Figure 6.7 shows the effect of reducing the number of temporal modulation filters from 17 1/4-octave-spaced filters (from 2 to 32 Hz) to 2 and 5 octave-spaced filters, all from 2 Hz, for reverberant noisy speech (top row) and noisy speech distorted by phase jitter (bottom row).

In the noisy reverberant condition, for RTs below 1.3 s, the STMI^T does not change with the number of filters. For RTs of 1.3 s and above, the predictions of the model with two filters (top, second panel) are higher by about 0.1 for SNRs larger than 3 dB. When octave spaced filters are used (top, middle panel), the predictions are essentially identical to the reference ones. The variations due to the change in the number of filters are well within the variations due to the simulation process. The variations are due to the statistic differences between the sentences, to the random noise added, as well as to the base spectrum subtraction.

In the phase jitter condition, the predictions are similar, irrespective of the number of filters. This means that in the case of phase jitter, very few extra information is provided to the STMI^T stage by increasing the number temporal modulation filters above 4 Hz. It is a conclusion analogous to the one made when changing the cutoff frequency of the low-pass temporal modulation filter.

SPECTRAL MODULATION BAND-PASS FILTERING Figure. 6.8 shows the STMI^T when changing the number of band-pass spectral modulation filters, for the noisy reverberant condition (top row) and for the phase jitter condition (bottom row). In the unmodified model (first column), there are 21 1/4-octave-spaced 1/3-octave-wide spectral modulation filters centered at frequencies between 0.25 and 8 cyc/oct. In

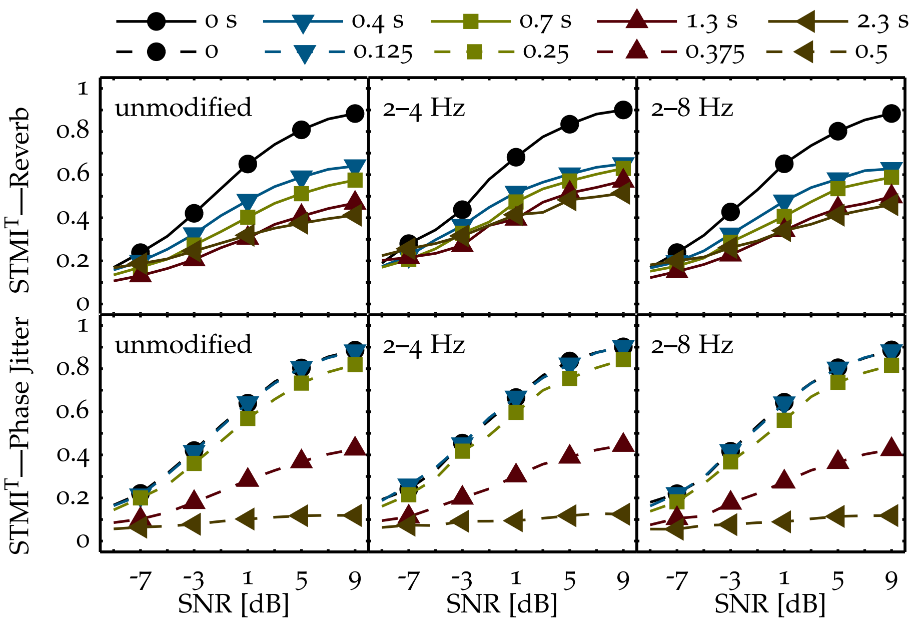

Figure 6.7: STMI$^{\mathrm{T}}$ predictions for different numbers of temporal modulation filters for noisy speech distorted by reverberation (top row) and phase jitter (bottom row). Predictions of unmodified model with 17 (34 in total, including positive and negative rates) temporal filters (first column) as well as predictions with 2 filters centered at 2 and 4 Hz (middle column) and 5 filters centered at 2, 4, 8, 16, 32 Hz (right column).

the modified models, the filters are octave-spaced and their positions are varied.

In the reverberant condition, the predictions are almost identical to the reference when using two low-scale filters, at 0.5 and 1 cyc/oct (top, second pannel). When the two filter are placed at 4 and 8 cyc/oct (top, third panel), the STMI$^{\mathrm{T}}$ has negative values for low SNRs. Negative values of STMI$^{\mathrm{T}}$ imply that the ratio $\frac{||T-N||^2}{||T||^2}$ is larger than 1 (c.f. Eq (3.8)). This indicates that the modulation magnitudes of the noisy speech are either (1) more than twice as large as the clean speech modulation magnitudes or (2) have a different sign than the clean speech modulation magnitude. When using octave-spaced filters, rather than the reference 1/4-octave-spaced filters, the predictions are the same as the reference.

Reducing the number of spectral modulation filters when making predictions on noisy speech distorted by phase jitter (bottom row) had the same effect as in the case of reverberant noisy speech: predictions are similar to the reference when the spectral modulations filter bank starts at low scales, but are lower than the reference when starting at higher scales. The STMI$^{\mathrm{T}}$ has negatives values when the filters are placed at 4 and 8 cyc/oct. The number of spectral modulation filters in the cortical model affects the STMI$^{\mathrm{T}}$ less than the position of the spectral center frequencies.

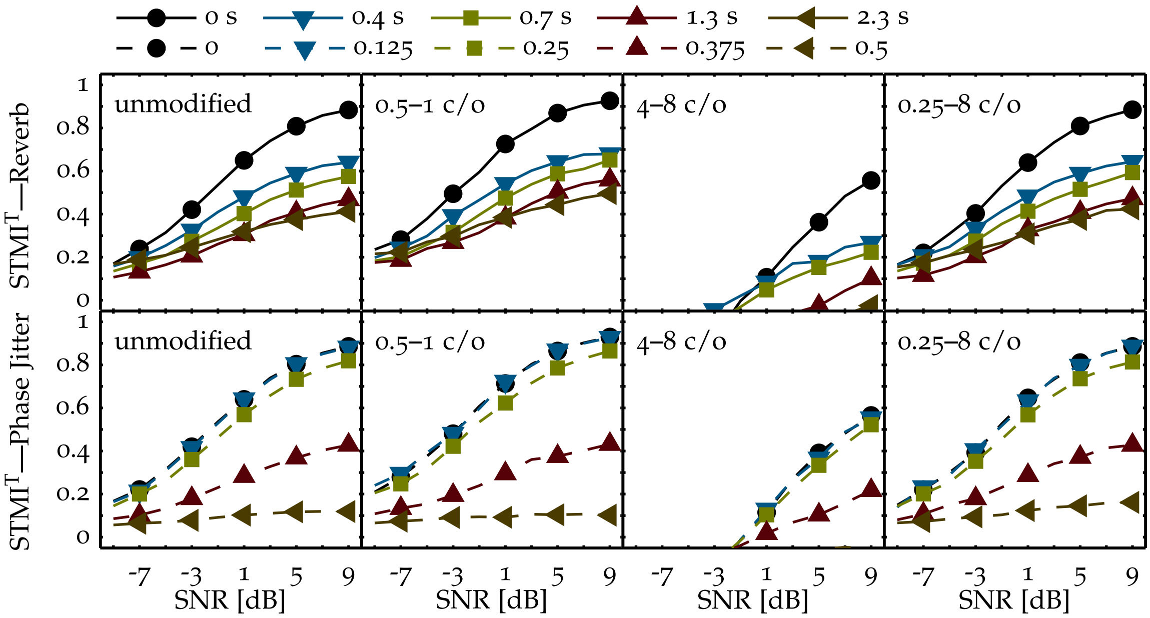

Figure 6.8: STMI^T predictions for different numbers of spectral modulation filters for noisy speech distorted by reverberation (top row) and phase jitter (bottom row). Predictions of unmodified model with 21 spectral modulation filters (first column), as well as predictions with two filters centered at 0.5 and 1 cyc/oct (second column), two filters centered at 4 and 8 cyc/oct (third column) and six filters centered at 0.25, 0.5, 1, 2, 4 and 8 cyc/oct (fourth column).

This analysis suggests that temporal modulation selectivity of the central auditory process is not crucial to predict speech intelligibility within the framework of the STMI^T; a simple temporal low-pass modulation filter would produce results similar to the reference model for the two distortions studied. The necessary cutoff frequency depends on the distortion but is in the 2–4 Hz range. Spectral modulation selectivity, however, does seem to be critical, especially when the distortion is phase jitter. When using band-pass filters, the number of filters required is quite low, one or two, as long as they cover the rates (scales) between 2 and 4 Hz (0.5 and 2 cyc/oct). It is important to remember that, while the number of modulation filters in one dimension was reduced, the other dimension still had as many filters as in the reference model.

7 DISCUSSION

In this thesis, the performance of the STMI^T was evaluated for three distortions applied to noisy speech: reverberation, phase jitter and spectral subtraction. Data from three subjects were gathered for the phase jitter distortion. The impact of each distortion on the cortical representation of the auditory model was studied.

The effect of spectro-temporal modulation selectivity on the STMI^T was investigated for two distortions: reverberation and phase jitter. First, the effect of using either a temporal or spectral low-pass modulation filter and, second, the effects of reducing the number of either spectral or temporal band-pass modulation filters, were analyzed.

7.1 EVALUATION OF THE STMI^T PERFORMANCE

7.1.1 *Reverberation*

The STMI^T prediction of the reverberation data ΔSRTs was good for RTs of 0, 0.4 and 0.7 s. For RTs larger than 0.7 s, no ΔSRT could be calculated on the STMI^T predictions, because the curves did not reach 0.5.

Elhilali *et al.* (2003, Fig. 8C) obtained a good fit between the STMI^T and the data for the noisy reverberant condition. Their conditions, however, were different: white noise was used instead of speech-shaped noise, the reverberation impulse response was a Gaussian white noise with exponentially decaying envelope rather than resulting from a room-acoustics simulation, the speech material was nonsense syllables rather than meaningful sentences, and the scoring method was phonemes rather than words. It is not possible to pinpoint exactly which of these aspects are responsible for the deviations in the predictions. The different impulse responses might be the culprit, or it might be that the STMI^T, like the STI or the AI, requires different mapping functions depending on the speech material in order to make accurate predictions (French and Steinberg, 1947; IEC, 2003). A more detailed comparison of predictions in both reverberant conditions would have been possible if enough data to produced psychometric functions had been available.

7.1.2 *Phase Jitter*

The STMI^T accounts well for the data obtained in the phase jitter distortion experiment. The predictions follow the same trends as the data, showing minima when $\alpha = \{0.5, 1\}$ and a local intelligibility maxima when $\alpha = 0.75$. Unlike the data, the predictions never reach zero nor one. This could probably be solved by adding some form of internal noise or threshold to either the auditory model or the STMI^T computation stage.

Elhilali *et al.* (2003) also obtained a good fit between the STMI^T and the data. In their experiment, the data were obtained by scoring phonemes rather than words. They did not notice the minima and local maxima in the intelligibility curve because the resolution of the α parameter was too coarse.

Considering the cortical representation of the auditory model, it was found that both the temporal and the spectral modulation dimensions where affected by the phase jitter. The impact was not of the same magnitude for all rate–scale–frequency combination, but was noticeable nonetheless.

7.1.3 *Spectral subtraction*

The STMI^T fails to predict the data when spectral subtraction is applied to noisy speech. The data, from Jørgensen and Dau (2011), as well as results by Boll (1979) and Sarampalis *et al.* (2009), showed that intelligibility *decreases* when the over-subtraction factor increases. In other words, spectral subtraction worsens intelligibility. In contrast, the STMI^T, like the sSTI (Jørgensen and Dau, 2011), predicts that the intelligibility *increases* with the over-subtraction factor, i.e. that spectral subtraction improves intelligibility.

The STMI^T fails in this condition because it looks at the reduction in spectro-temporal modulation energy. In this case, the difference between the clean and noisy modulation spectra decreases while the distortion parameter increases, as if more of the spectro-temporal modulation content was preserved through the processing.

The fact that the STMI^T fails does not imply that the spectro-temporal modulation filter processing is inappropriate. It does suggest, however, that the STMI^T metric might not be suitable to predict intelligibility in this condition, as well as probably in other conditions where the modulation content increases with an increasing distortion factor. It could be attempted to use other metrics on the cortical representation, in order to produce correct predictions for spectral subtraction. Goldsworthy and Greenberg (2004) proposed a number of alternate metrics for the speech-based STI that could possibly be adapted to the current model. All of these metrics consider that the model has as its input the clean and the noisy speech.

7.2 NECESSITY OF THE SPECTRO-TEMPORAL CORTICAL FILTERS

It can be seen in the modulation magnitude figures of Sec. 6.1, as well as in Appendix D, that each distortion affects both the spectral *and* the temporal modulation domains. The salience of the effect is dependent on the rate–scale–frequency combination, but the impact is always observable in both modulation dimensions. It is important to note that the spectral, or temporal, modulation magnitude plots are not a "one-dimension" analysis, because they still consider the *joint* spectro-temporal modulations. The distortions have thys inherently an impact on both domains, and on both types of figures.

The temporal and spectral modulation selectivity assumed in the model was reduced by replacing one of the modulation filter banks by a single low-pass filter, while keeping the selectivity in the other modulation dimension. A high cutoff frequency of the modulation filter corresponds to *no selectivity* to the corresponding modulations. For example, setting a high spectral cutoff frequency on the spectral modulation filters is equivalent to using only temporal modulation filters.

It was found that replacing the temporal filter bank by a temporal low-pass filter with a cutoff frequency as low as 2 Hz did not affect the STMI^T predictions, neither for the noisy reverberant condition nor for the phase jitter applied to noisy speech condition. Replacing the spectral modulation filter bank by a single spectral low-pass filter at 8 cyc/oct did not affect the predictions for the noisy reverberant condition, but did affect the predictions for the phase jitter condition when values of α larger than 0.25 were considered. This suggests that in the framework of the STMI^T, temporal modulation frequency selectivity is not necessary but some form of spectral modulation selectivity is required.

This observation seems contradictory to the fact that the STI, which uses temporal modulation filters, can predict intelligibility for some distortion conditions, similarly to the STMI^T. Elhilali *et al.* (2003) showed that the STI was insensitive to the phase jitter and predicted a constant 100 % intelligibility. The good performance of the STMI^T with such a distortion is an argument toward the necessity of using the joint spectro-temporal modulations to predict intelligibility. It is hypothesized, however, that the STI cannot predict intelligibility for phase jitter, not because it considers temporal modulation only, but because of the way it is computed. The STI probe signals are modulated with the same rate at all audio frequencies, such that the envelope of each individual audio-frequency band is the same as the envelope of the signal. It has been seen that phase jitter affects the temporal modulations (Fig. 2.7) but not the temporal envelope of the signal. If the probe had different modulation frequencies in all audio-frequency bands, such as is one of the speech-based STI variations, the STI would probably be sensitive to the effect of the jitter and could account for phase jitter. This suggests that having appropriate inputs to the models is crucial and that it could be possible to predict intelligibility using only temporal modulation filters.

Similar to what Houtgast *et al.* (1980, p. 63) mentioned about the possibility of reducing the number of temporal modulation filters in the STI without affecting predictions, the STMI^{T} predictions did not vary when reducing the number of modulation filters in a given dimension, as long as they covered the correct frequency range. The number of modulation filters in the other dimension was kept as in the unmodified model. For both modulation dimensions, using octave-spaced instead of 1/4-octave spaced filters did not affect the predictions. It is hypothesized that the number of modulation filters in a given dimension could be reduced without impacting the predictions because of how close to each other they were in the unmodified model. The outputs of neighboring filters are probably strongly correlated, such that reducing the number of filter simply reduces the amount of redundant information. The number of modulation filters could be reduced to two in a given dimension (at 2 and 4 Hz, and at 0.5 and 1 cyc/oct) without noticeable impact on the predictions. These rate and scale bands are where most of modulation energy is for a given audio-frequency band. It is also in these bands that the effects of the distortions are the most pronounced.

Placing two spectral filters at higher scales (4 and 8 cyc/oct) resulted in negative values of STMI^{T}. It was observed that at these scales, the modulation magnitudes of both clean and noisy speech are very small and often have opposite signs, resulting in a negative STMI^{T}. The difference in size between low and high rates or scales is a property of the speech signal, but also of the model. The decrease in the outputs at high scales is due to the finite bandwidth of the cochlear filters. When spectral peaks become close, they become unresolved by the cochlear filters. Based on this explanation, the upper limits of the temporal and spectral filters are inversely related through the effective bandwidth of the cochlear filters (Chi *et al.*, 2005). This means that the auditory process model, thus the STMI^{T}, has an implicit weighting of scale, rate and frequency.

Since the STMI^{T} failed for one of the conditions, it could be tried to apply the SNR$_{env}$ metric, from the sEPSM (Jørgensen and Dau, 2011), to the cortical representation. In the sEPSM, the model has the noisy speech and the noise at its input; it has access to an estimate of the noise, rather than access to the clean signal.

To show how the sEPSM performs with regard to the conditions studied, its predictions are shown side by side with the STMI^{T} predictions and the data in Fig 7.1. The conditions for the sEPSM simulations were as follow: 150 sentences from the CLUE material, 144 sentences from the DANTALE II material and 100 single words from the DANTALE material were used, the noise duration was set to the same duration as the speech stimuli. All distortions were applied in the same way as in this study.

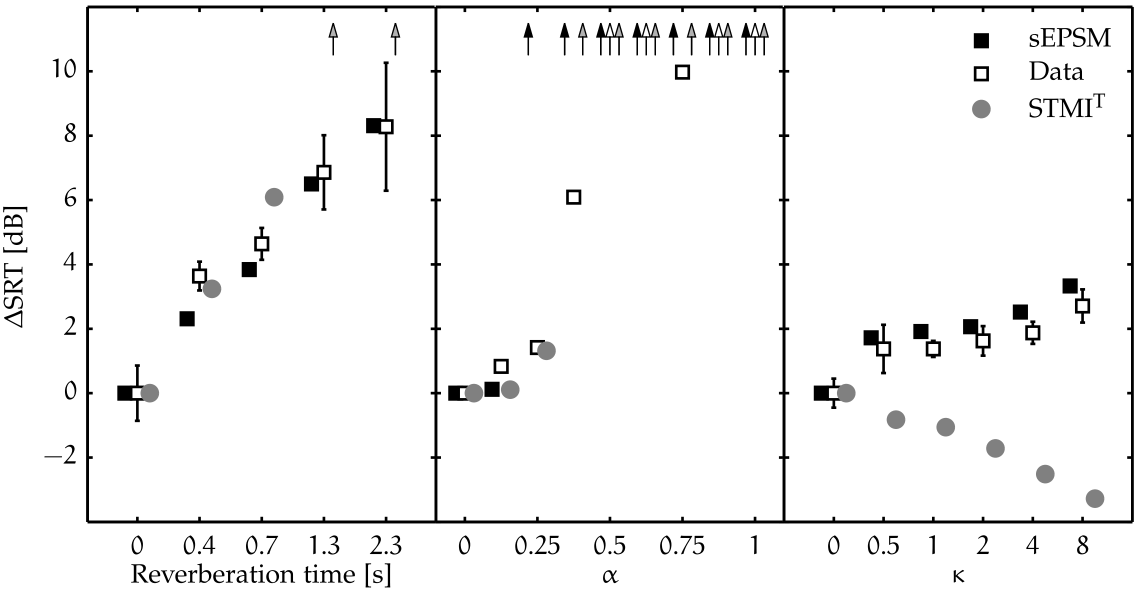

Figure 7.1: Comparison between intelligibility predictions by the sEPSM (filled squares) and the STMI^T (gray circles), and data (empty squares). The figure shows the ΔSRTs for the noisy speech in the reverberant condition (left), distorted by phase jitter (center) and processed by spectral subtraction (right). The vertical bars represent one standard deviation. The data and the sEPSM predictions for the reverberant and spectral subtraction conditions are from (Jørgensen and Dau, 2011). The sEPSM predictions for the phase jitter conditions are unpublished.

The left-hand panel shows ΔSRTs the reverberant condition. The sEPSM predicts the data almost exactly for all RTs. The STMI^T makes good predictions for RTs below 1.3 s but could not produce a ΔSRT for longer RTs.

The center panel shows the ΔSRTs for the phase jitter condition. The sEPSM could produce ΔSRTs for $\alpha = \{0, 0.125\}$ only. The STMI^T could produce good fitting ΔSRT predictions up to $\alpha = 0.25$. Both model fail to predict a ΔSRT for larger values of α.

The right panel shows the ΔSRTs for the spectral subtraction condition. The sEPSM makes accurate predictions and follows the trend of the data: ΔSRTs increase with the over-subtraction factor κ, i.e. intelligibility decreases with increasing values κ. The STMI^T predicts the opposite trend, it predicts that intelligibility *increases* with κ.

Figure 7.2 shows the predictions for the phase jitter condition in more details. It can be seen that although no ΔSRTs could be produced for $\alpha \geqslant 0.25$, the sEPSM accounts fairly well the data. The reasons why no minima occur at $\alpha = \{0.5, 1\}$ will be further investigated.

For the first two distortions, both model have good or excellent agreement with the data. Yet, the sEPSM only considers a subset of the modulations the STMI^T considers. This comparison suggests that

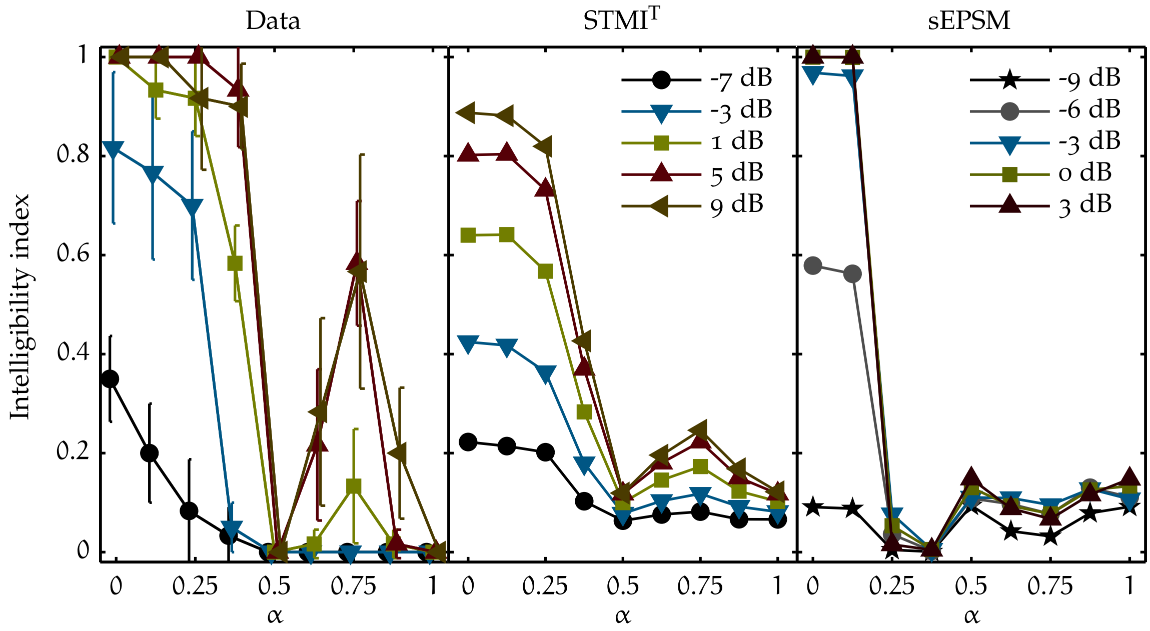

Figure 7.2: Comparison of the sEPSM, the STMI^T and data for the phase jitter experiment, as a function of α, with the SNR as the parameter.

the combination of the spectro-temporal modulation might not be necessary to predict intelligibility. As for the third distortion, it does not lead to conclusions about the spectro-temporal modulation filters, but rather about the fact the STMI^T metric is not suited to predict intelligibility for speech processed by spectral subtraction.

8 SUMMARY AND CONCLUSIONS

The speech-based spectro-temporal modulation index (STMI^T) was evaluated by comparing predictions with data for three conditions: (i) reverberant noisy speech, (ii) noisy speech distorted by phase jitter, and (iii) noisy speech processed by spectral subtraction. In the first and third conditions, predictions were compared to ΔSRT data. In the second conditions, psychometric functions were compared.

The STMI^T predictions exhibited decent agreement with the data in the noisy reverberant condition, but the differences to similar experiment suggest that the STMI^T might require different prediction-to-intelligibility mapping functions depending on the speech material. The model accounted well for phase jitter and predicted all tendencies in the data. The STMI^T failed in the spectral subtraction condition, predicting an increase in intelligibility with increased over-subtraction factor. The analysis of the model revealed that the spectral subtraction process increased the spectro-temporal modulation energy, which lead to an increase in predicted intelligibility.

Examination of the internal representation of the auditory process model showed that phase jitter affected both spectral and temporal modulation domains. Predictions by the sEPSM showed that it was possible to account for such a distortion by considering only the temporal modulation domain, which lead to the hypothesis that the speech-based speech transmission index (sSTI) could be able to account for phase jitter as well.

It was showed that, in the framework of the STMI^T, some degree of spectral modulation selectivity was necessary, but temporal modulation frequency selectivity was not. It was also demonstrated that the spacing between modulation filters could be increased to one octave without affecting the predictions.

It was shown that spectro-temporal modulation filters might not be crucial to predict intelligibility and that the STMI^T metric was not suited to predict intelligibility for speech processed by spectral subtraction.

BIBLIOGRAPHY

ANSI-S3.5 (**1969**), *American National Standard methods for the calculation of the Articulation Index* (American National Standards Institute, Inc., New York).

ANSI-S3.5 (**1997**), *American National Standard methods for calculation of the Speech Intelligibility Index* (American National Standards Institute, Inc., New York).

Berouti, M. and Schwartz, R. (**1979**), "Enhancement of speech corrupted by acoustic noise," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79. (1), 208–211, URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1170788.

Boll, S. (**Apr. 1979**), "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech, and Signal Processing **27**(2), 113–120, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1163209.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., and Others (**1994**), "An international comparison of long-term average speech spectra," Journal of the Acoustical Society of America **96**(4), 2108–2120, URL http://www.dydaktyka.cba.pl/AvSpeech.pdf.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**), "Spectro-temporal modulation transfer functions and speech intelligibility," Journal of the Acoustical Society of America **106**, 2719, URL http://asadl.org/jasa/resource/1/jasman/v106/i5/p2719_s1.

Chi, T., Ru, P., and Shamma, S. A. (**2005**), "Multiresolution spectrotemporal analysis of complex sounds," The Journal of the Acoustical Society of America **118**(2), 887, URL http://link.aip.org/link/JASMAN/v118/i2/p887/s1&Agg=doi.

Christensen, C. L. (**2007**), *Odeon Room Acoustics Program, Version 10.1, User Manual, Industrial, Auditorium and Combined Editions* (Odeon A/S, Kgs. Lyngby), URL http://ww.odeon.dk/.

Dau, T., Verhey, J., and Kohlrausch, A. (**Nov. 1999**), "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers." The Journal of the Acoustical Society of America **106**(5), 2752–60, URL http://www.ncbi.nlm.nih.gov/pubmed/10573891.

Depireux, D., Simon, J. Z., Klein, D. J., and Shamma, S. (**Mar. 2001**), "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex." Journal of neurophysiology **85**(3), 1220–34, URL http://www.ncbi.nlm.nih.gov/pubmed/11247991.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**), "Effect of temporal envelope smearing on speech reception," Journal of the Acoustical Society of America **95**(2), 1053–1064.

Dubbelboer, F. and Houtgast, T. (**Nov. 2007**), "A detailed study on the effects of noise on speech intelligibility." The Journal of the Acoustical Society of America **122**(5), 2865–71, URL http://www.ncbi.nlm.nih.gov/pubmed/18189576.

Dubbelboer, F. and Houtgast, T. (**Dec. 2008**), "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility." The Journal of the Acoustical Society of America **124**(6), 3937–46, URL http://www.ncbi.nlm.nih.gov/pubmed/19206818.

Elhilali, M. (**2004**), "Neural Basis and Computational Strategies for Auditory Processing," Ph. d. thesis, University of Maryland, College Park, URL http://129.2.17.93/drum/handle/1903/2084.

Elhilali, M., Chi, T., and Shamma, S. A. (**Oct. 2003**), "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Communication **41**(2-3), 331–348, URL http://linkinghub.elsevier.com/retrieve/pii/S0167639302001346.

Ewert, S. D. and Dau, T. (**Sep. 2000**), "Characterizing frequency selectivity for envelope fluctuations." The Journal of the Acoustical Society of America **108**(3 Pt 1), 1181–96, URL http://www.ncbi.nlm.nih.gov/pubmed/11008819.

French, N. R. and Steinberg, J. C. (**1947**), "Factors governing the intelligibility of speech sounds," Journal of the Acoustical Society of America **19**(1), 90–119, URL http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=JASMAN000019000001000090000001&idtype=cvips&gifs=yes.

Goldsworthy, R. L. and Greenberg, J. E. (**Dec. 2004**), "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," The Journal of the Acoustical Society of America **116**(6), 3679–3689, URL http://link.aip.org/link/?JAS/116/3679/1http://scitation.aip.org/journals/doc/JASMAN-ft/vol_116/iss_6/3679_1.html.

Houtgast, T. and Steeneken, H. (**1973**), "The modulation transfer function in room acoustics as a predictor of speech intelligibility," The Journal of the Acoustical Society of America **54**, 557, URL http://link.aip.org/link/?JASMAN/54/557/1.

Houtgast, T. and Steeneken, H. (**1985**), "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," The Journal of the Acoustical Society of America **77**(March 1985), 1069, URL http://link.aip.org/link/?JASMAN/77/1069/1.

Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**), "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**(1), 60–72.

IEC (**2003**), *60268–16-2003 Sound system equipment — Part 16: Objective rating of speech intelligibility by speech transmission index* (International Electrotechnical Commision, Geneva, Switzerland), 3rd ed.

Itoh, K. and Mizushima, M. (**1997**), "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE Comput. Soc. Press), vol. 1, pp. 419–422, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=599662.

Jørgensen, S. and Dau, T. (**2011**), "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," Journal of the Acoustical Society of America , in press.

Kowalski, N., Depireux, D., and Shamma, S. (**1996**), "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses tomoving ripple spectra," Journal of Neurophysiology **76**(5), 3503, URL http://jn.physiology.org/content/76/5/3503.abstract?sid=9274c826-96c7-43ba-957a-304ab29e0cc4.

Lee, E. A. and Messerschmitt, D. G. (**1994**), *Digital Communication* (Kluwer Academic Publishing, Boston), 2nd ed.

Lim, J. (**Oct. 1978**), "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Transactions on Acoustics, Speech, and Signal Processing **26**(5), 471–472, URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1163129http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1163129.

Lim, J. and Oppenheim, A. (**1979**), "Enhancement and bandwidth compression of noisy speech," Proceedings of the IEEE **67**(12), 1586–1604, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1455809.

Ludvigsen, C., Elberling, C., and Keidser, G. (**Jan. 1993**), "Evaluation of a noise reduction method–comparison between observed scores and scores predicted from STI." Scandinavian audiology. Supplementum **38**, 50–5, URL http://www.ncbi.nlm.nih.gov/pubmed/8153564.

Lyon, R. and Shamma, S. (**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, URL http://www.springerlink.com/content/k21w7783281p5167/.

Marr, D. and Hildreth, E. (**Feb. 1980**), "Theory of Edge Detection," Proceedings of the Royal Society B: Biological Sciences **207**(1167), 187–217, URL http://rspb.royalsocietypublishing.org/cgi/content/abstract/207/1167/187http://cirl.lcsr.jhu.edu/wiki/images/7/77/MarrHildreth.pdf.

Nielsen, J. B. and Dau, T. (**Jan. 2009**), "Development of a Danish speech intelligibility test." International journal of audiology **48**(10), 729–41, URL http://www.ncbi.nlm.nih.gov/pubmed/19626512.

Payton, K. L. and Braida, L. D. (**Dec. 1999**), "A method to determine the speech transmission index from speech waveforms." The Journal of the Acoustical Society of America **106**(6), 3637–48, URL http://www.ncbi.nlm.nih.gov/pubmed/10615702.

Ratnam, R., Jones, D., and O'Brien, W. (**Jun. 2004**), "Fast Algorithms for Blind Estimation of Reverberation Time," IEEE Signal Processing Letters **11**(6), 537–540, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1300603.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**Dec. 2009**), "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise." The Journal of the Acoustical Society of America **126**(6), 3236–45, URL http://link.aip.org/link/doi/10.1121/1.3257225/html.

Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (**Oct. 2009**), "Objective measures of listening effort: effects of background noise and noise reduction." Journal of speech, language, and hearing research : JSLHR **52**(5), 1230–40, URL http://www.ncbi.nlm.nih.gov/pubmed/19380604.

Shamma, S., Chadwik, R., Wilbur, W., Morrish, K., and Rinzel, J. (**1986**), "A Biophysical Model of Cochlear Processing: Intensity Dependence of Pure Tone Response," Journal of the Acoustical Society of America **80**(1), 133–145, URL http://129.2.17.93/drum/handle/1903/4459.

Steeneken, H. J. and Houtgast, T. (**Jan. 1980**), "A physical method for measuring speech-transmission quality." The Journal of the Acoustical Society of America **67**(1), 318–26, URL http://www.ncbi.nlm.nih.gov/pubmed/7354199.

Tsoukalas, D., Mourjopoulos, J., and Kokkinakis, G. (**1997**), "Speech enhancement based on audible noise suppression," IEEE Transactions on Speech and Audio Processing **5**(6), 497–514, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=641296.

Wang, K. and Shamma, S. (**Jul. 1994**), "Self-normalization and noise-robustness in early auditory representations," IEEE Transactions on Speech and Audio Processing **2**(3), 421–435, URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=294356.

A

THE HILBERT TRANSFORM AND ENVELOPE EXTRACTION

The Hilbert transform is a linear operator which transforms function $f(t)$ into $\hat{f}(t)$ within the same domain. The Hilbert transform is defined as:

$$\hat{f}(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{f(\tau)}{t - \tau} \, d\tau. \tag{A.1}$$

It can be used to produce the so-called *analytical signal*, $z(t)$, i.e. a signal with no negative spectrum. The analytical signal of signal $s(t)$ is

$$z(t) = s(t) + j\hat{s}(t), \tag{A.2}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$.

The *Hilbert envelope*, $s_{\text{env}}(t)$, of signal $s(t)$, is the magnitude of the analytical signal:

$$s_{\text{env}}(t) = |z(t)| = \sqrt{s^2(t) + \hat{s}^2(t)}. \tag{A.3}$$

B

THE STMI^T ALGORITHM

Algorithm 1 exposes the process to compute the STMI^T from a series of sentences stored in audio files. It is a slightly modified version of the algorithm from Elhilali (2004), where the speech is in a long stream which is split in smaller two-second long pieces.

Algorithm 1 Algorithm for the computation of the STMI^T (Elhilali, 2004).

```
for file in files do
    $x \leftarrow$ read(file)
    $x \leftarrow (x - \mu_x)/\sigma_x$    (normalize signal)
    $x_0 \leftarrow$ make-base-signal($x$)
    $y \leftarrow$ auditory-spectrogram($x$)
    $y_0 \leftarrow$ auditory-spectrogram($x_0$)
    $r \leftarrow$ cortical-filtering($y$)
    $r_0 \leftarrow$ cortical-filtering($y_0$)
    $rsf \leftarrow$ average-over-time($r$)
    $rsf_0 \leftarrow$ average-over-time($r_0$)
    $T \leftarrow rsf - rsf_0$
    for distortion-condition in all-distortion-conditions do
        $x_n \leftarrow$ apply-distortion($x$, distortion-condition)
        $x_n \leftarrow (x_n - \mu_{xn})/\sigma_{xn}$
        $y_{n0} \leftarrow$ auditory-spectrogram($x_{n0}$)
        $r_n \leftarrow$ cortical-filtering($y_n$)
        $r_{n0} \leftarrow$ cortical-filtering($y_{n0}$)
        $rsf_n \leftarrow$ average-over-time($r_n$)
        $rsf_{n0} \leftarrow$ average-over-time($r_{n0}$)
        $N \leftarrow rsf_n - rsf_{n0}$
        $\mathrm{STMI}^T \leftarrow 1 - ||T - N||^2/||T||^2$
    end for
end for
```

C

AUDIOGRAMS OF SUBJECTS FOR PHASE JITTER EXPERIMENT

The audiograms for the three test subjects who participated to this thesis are show in Fig. C.1. All subjects are considered to have normal hearing.

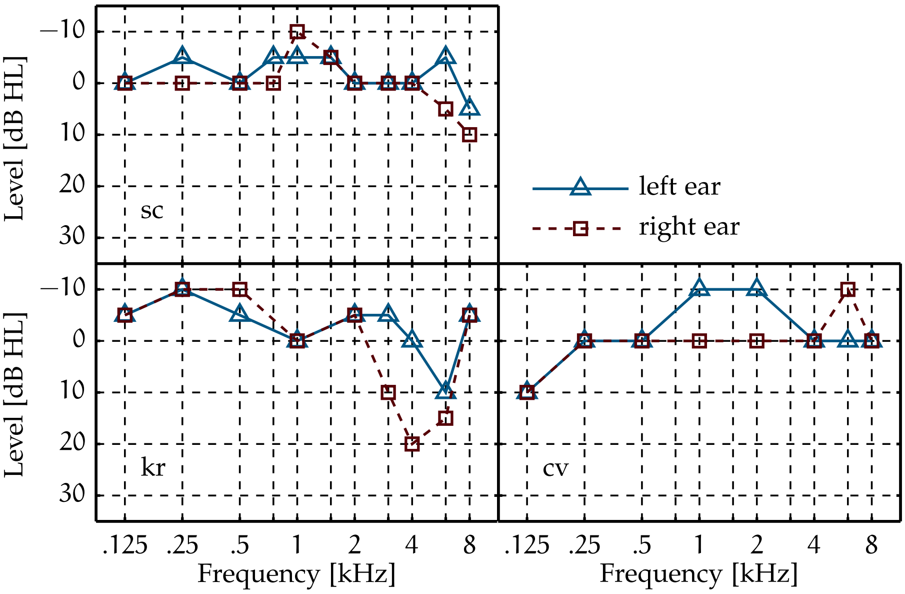

Figure C.1: Pure tone audiograms of the three test-subjects who participated in the phase jitter experiment.

D MODULATION MAGNITUDE FIGURES

Figures D.1 to D.8 show temporal and spectral modulation magnitudes for the distortion studied in this thesis. The rate–scales–frequency combinations are:

- rates: $\omega = \{2, 4, 8, 16, 32\}$ Hz;
- scales: $\Omega = \{0.25, 0.5, 1, 2, 4, 8\}$ cyc/oct;
- and frequencies: $f = \{0.5, 1, 2\}$ kHz.

The distortions are:

- speech-shaped noise only: Figs. D.1 (temporal) and D.2 (spectral);
- reverberation in combination with speech-shaped noise (9 dB SNR): Figs. D.3 (temporal) and D.4 (spectral);
- phase jitter and speech-shaped noise: Figs. D.5 (temporal) and D.6 (spectral);
- spectral subtraction applied on speech-shaped noise corrupted speech: Figs. D.7 (temporal) and D.8 (spectral).

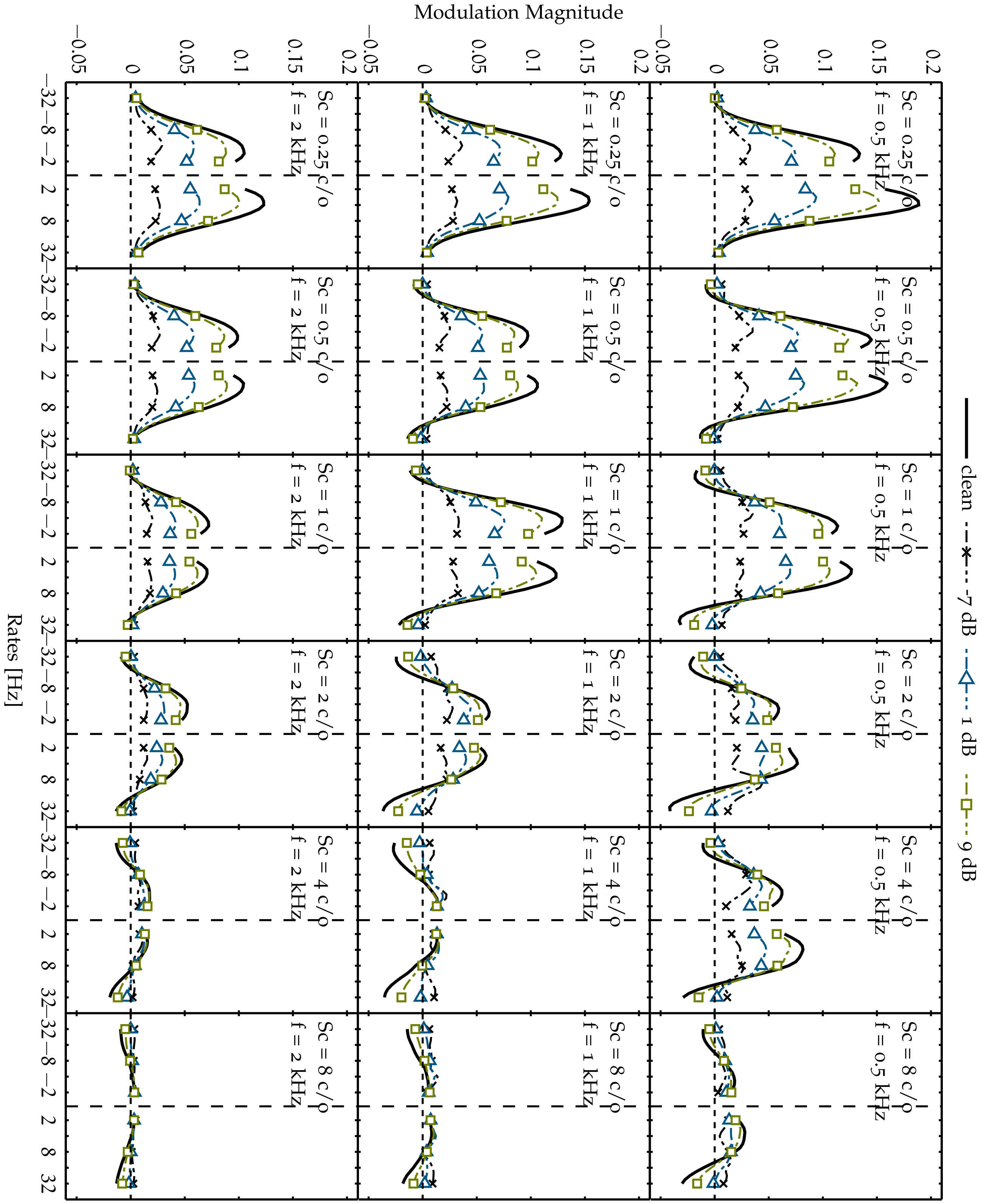

Figure D.1: *Temporal modulation magnitudes: speech-shaped noise only.* Temporal modulation magnitudes for combinations of octave-spaced scales, in the range $\Omega = [0.25, 8]$ cyc/oct, and for audio-frequency bands 0.5, 1 and 2 kHz.

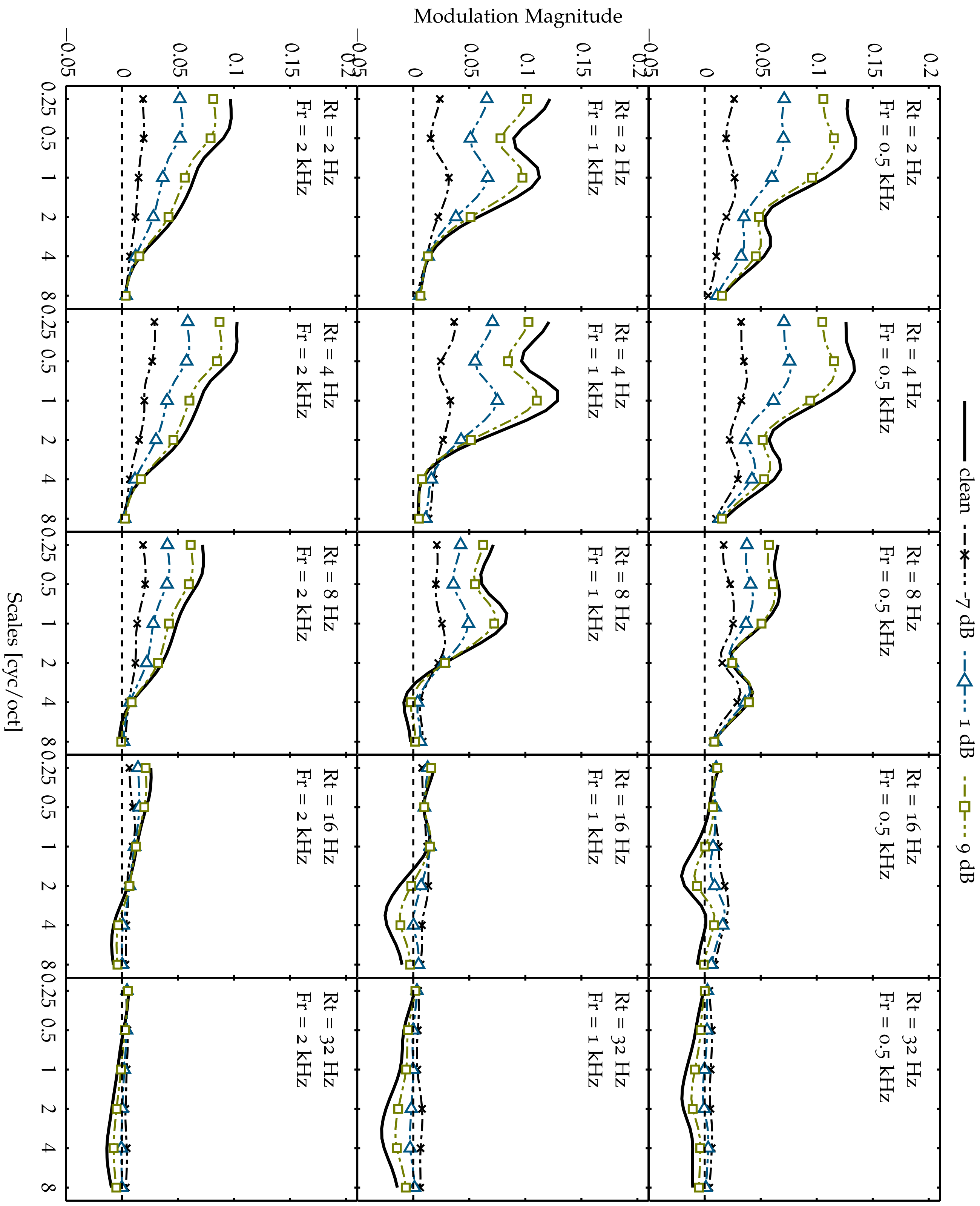

Figure D.2: *Spectral modulation magnitudes: speech-shaped noise only.* Spectral modulation magnitudes for combinations of octave-spaced positive rates, in the range $\omega = [2, 32]$ Hz, and for audio-frequency bands 0.5, 1 and 2 kHz.

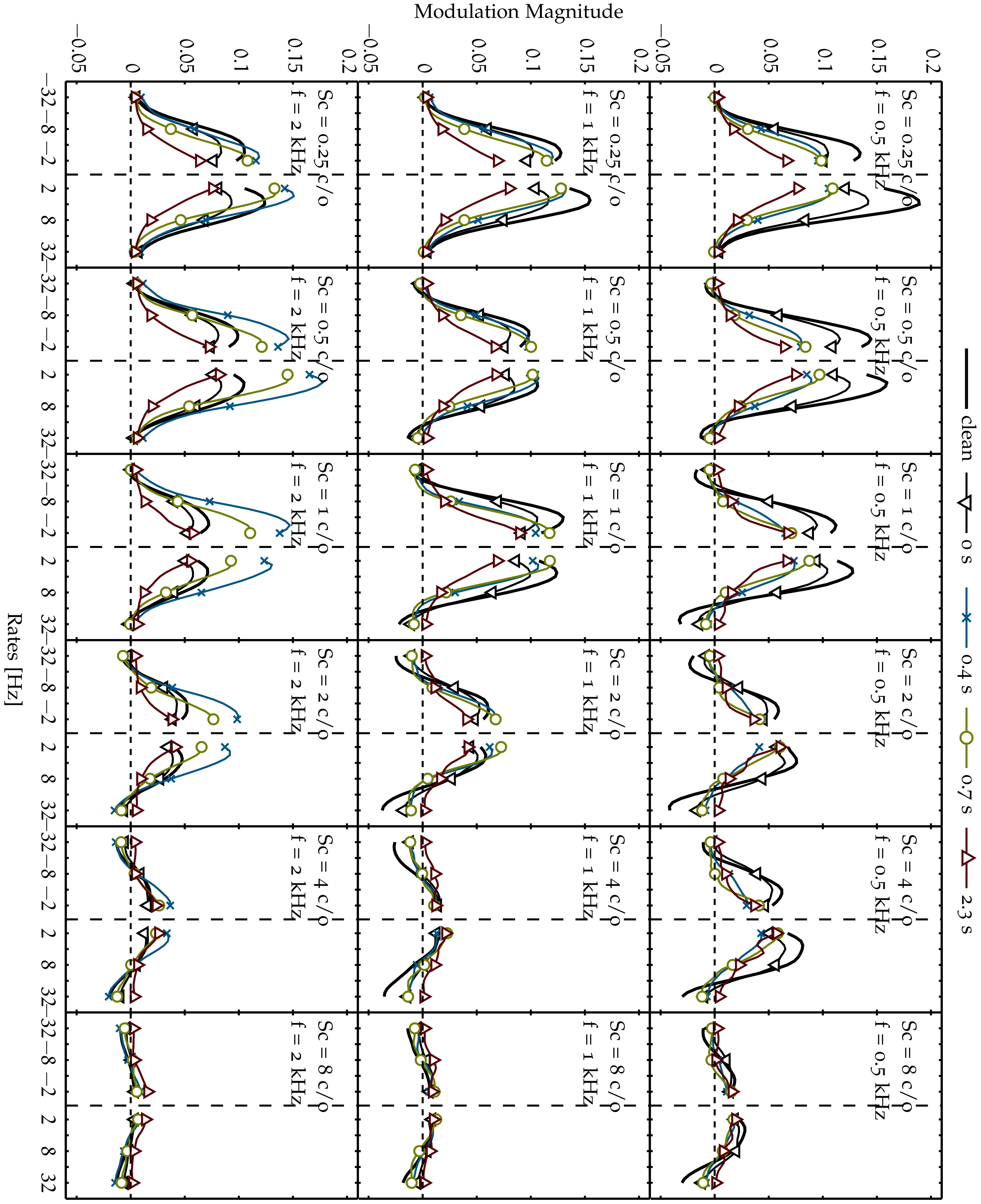

Figure D.3: *Temporal modulation magnitudes: reverberation.* Temporal modulation magnitudes for combinations of octave-spaced scales, in the range $\Omega = [0.25, 8]$ cyc/oct, and for audio-frequency bands 0.5, 1 and 2 kHz.

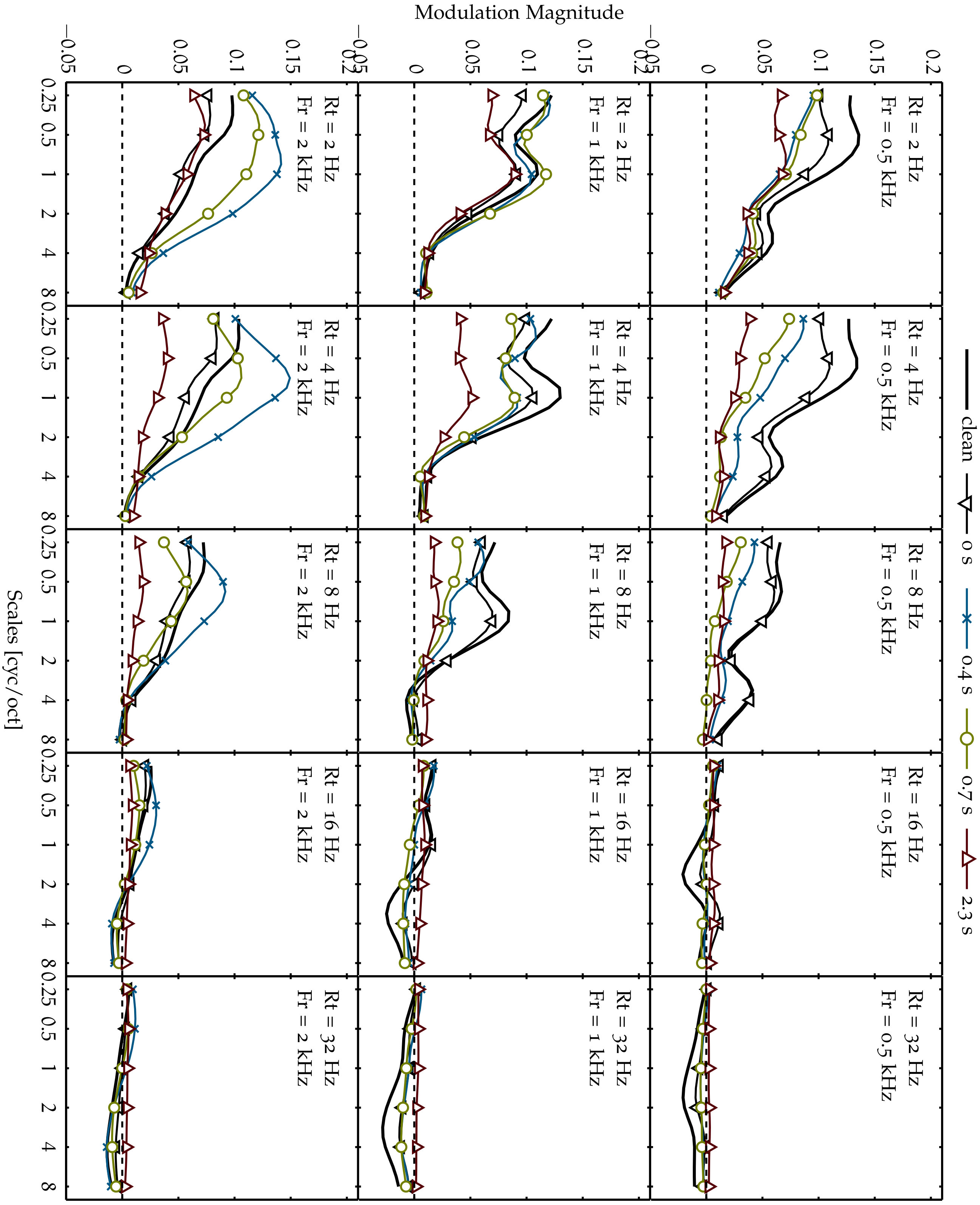

Figure D.4: *Spectral modulation magnitudes: reverberation.* Spectral modulation magnitudes for combinations of octave-spaced positive rates, in the range $\omega = [2, 32]$ Hz, and for audio-frequency bands 0.5, 1 and 2 kHz.

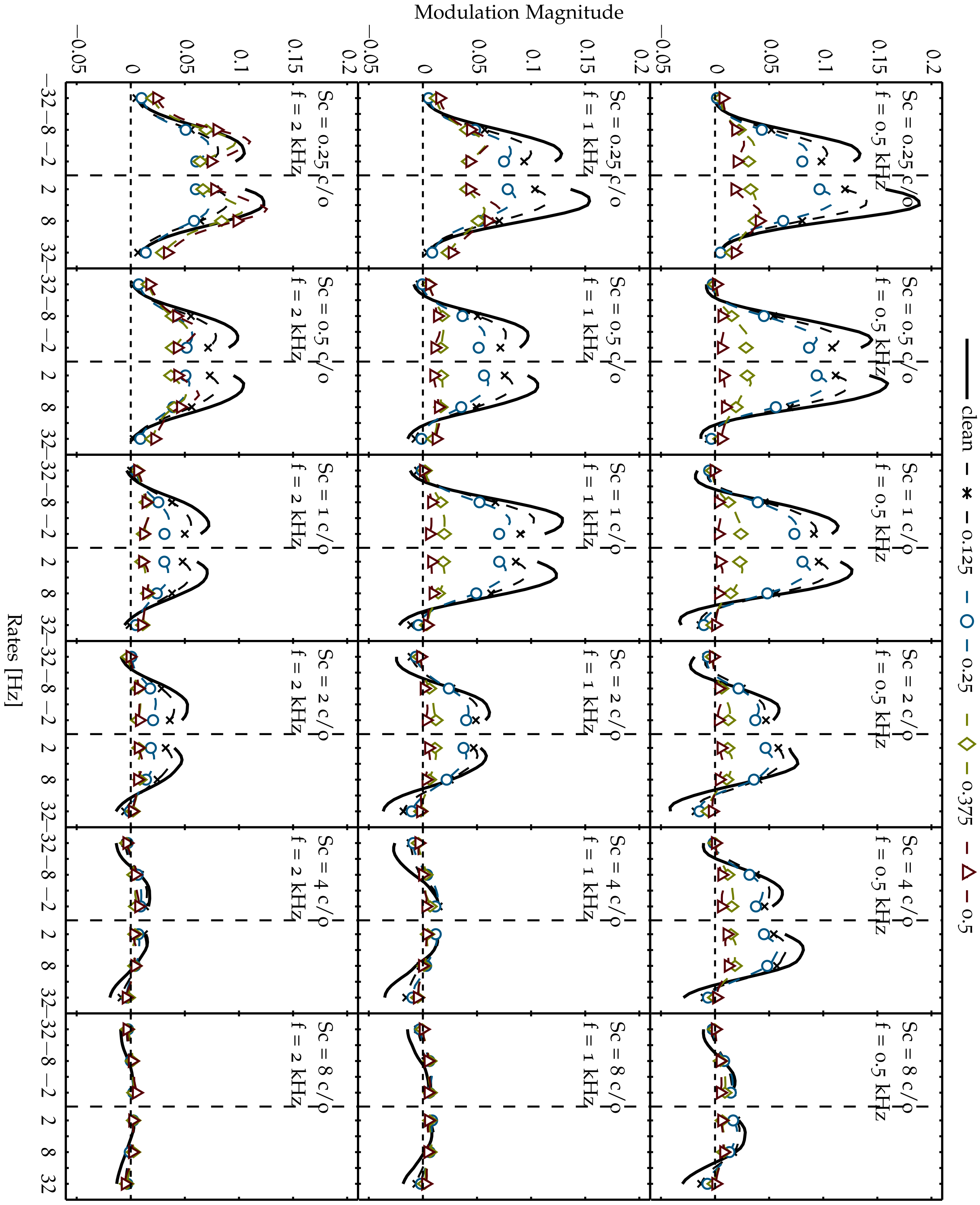

Figure D.5: *Temporal modulation magnitudes: phase jitter.* Temporal modulation magnitudes for combinations of octave-spaced scales, for the range $\Omega = [0.25, 8]$ cyc/oct, and for audio-frequency bands 0.5, 1 and 2 kHz.

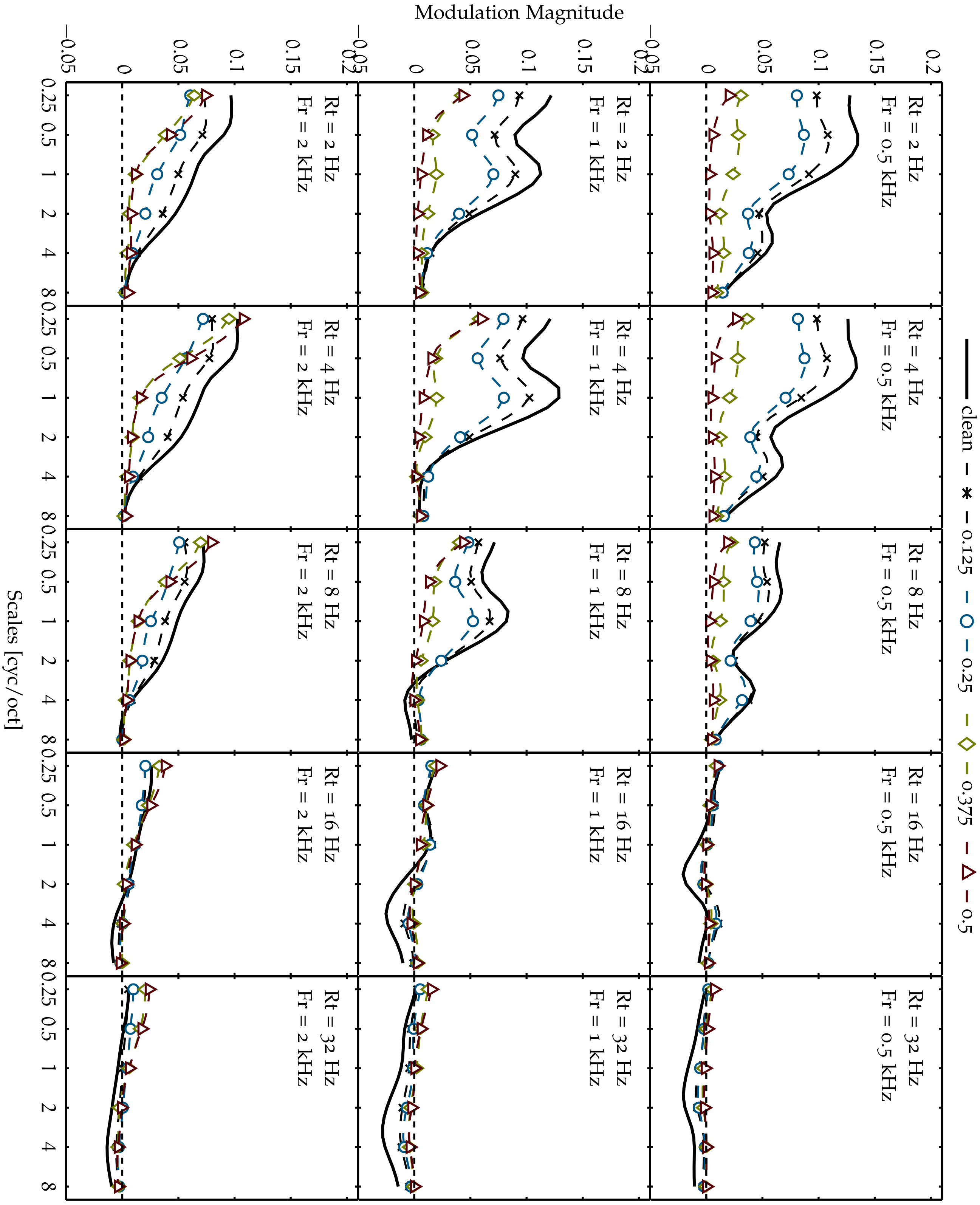

Figure D.6: *Spectral modulation magnitudes: phase jitter.* Spectral modulation magnitudes for combinations of octave-spaced positive rates, in the range $\omega = [2, 32]$ Hz, and for audio-frequency bands 0.5, 1 and 2 kHz.

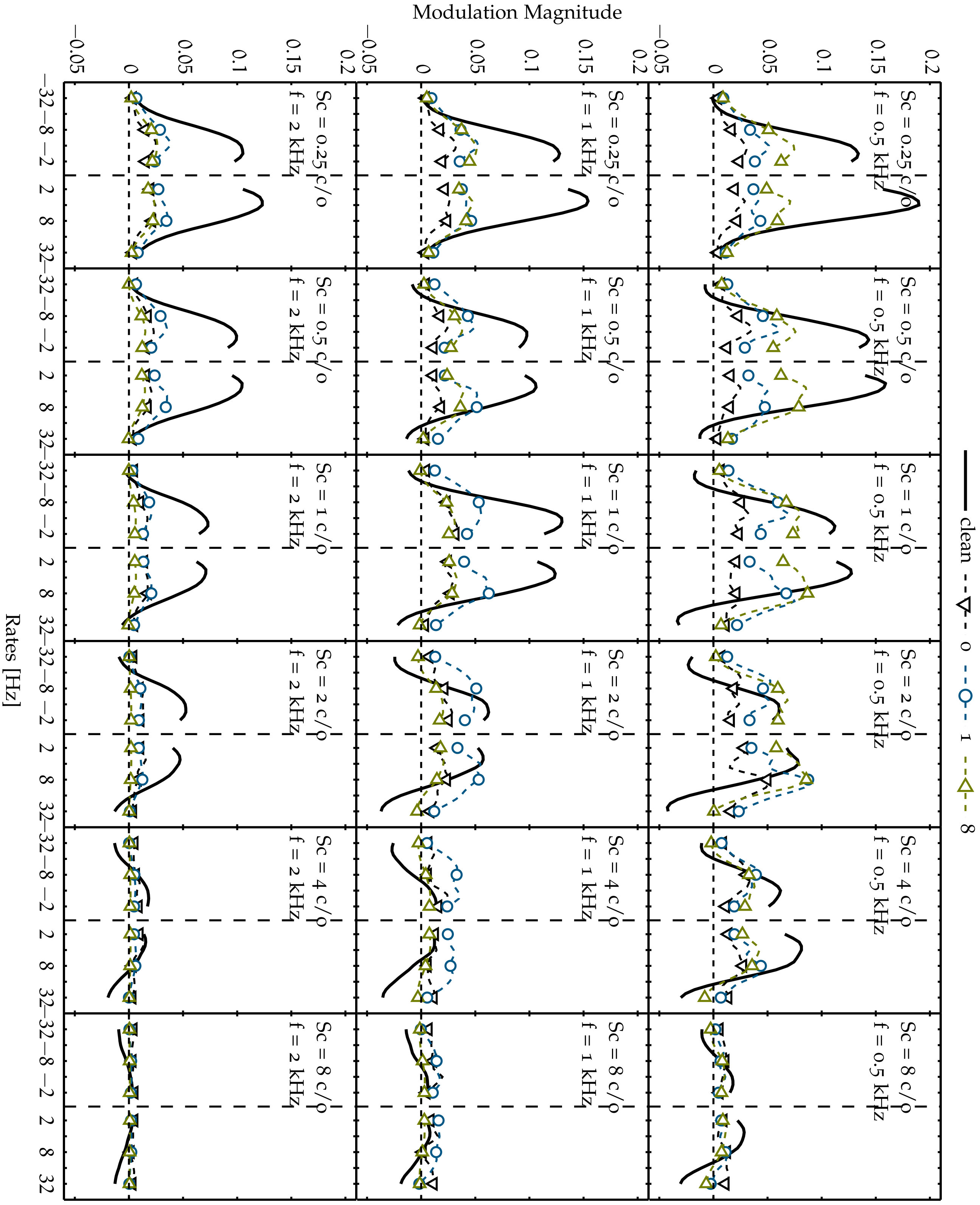

Figure D.7: *Temporal modulation magnitudes: spectral subtraction.* Temporal modulation magnitudes for combinations of octave-spaced scales, in the range $\Omega = [0.25, 8]$ cyc/oct, and for audio-frequency bands 0.5, 1 and 2 kHz.

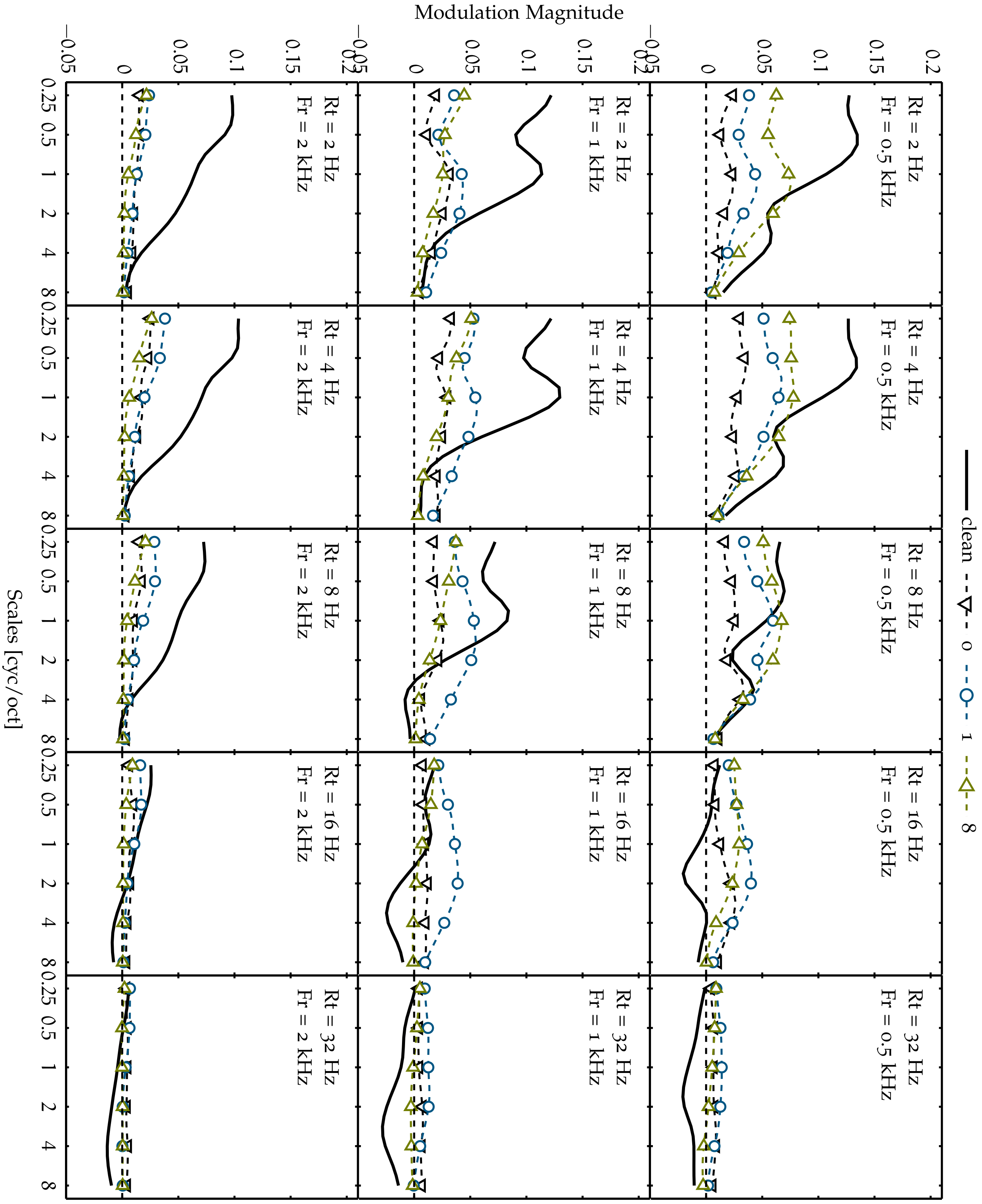

Figure D.8: *Spectral modulation magnitudes: spectral subtraction.* Spectral modulation magnitudes for combinations of octave-spaced positive rates, in the range $\omega = [2, 32]$ Hz, and for audio-frequency bands 0.5, 1 and 2 kHz.