Chapter 40
Effects of Peripheral Tuning on the Auditory Nerve's Representation of Speech Envelope and Temporal Fine Structure Cues

Rasha A. Ibrahim and Ian C. Bruce

Abstract A number of studies have explored how speech envelope and temporal fine structure (TFS) cues contribute to speech perception. Some recent investigations have attempted to process speech signals to remove envelope cues and leave only TFS cues, but the results are confounded by the fact that envelope cues may be partially reconstructed when TFS signals pass through the narrowband filters of the cochlea. To minimize this reconstruction, investigators have utilized large numbers of narrowband filters in their speech processing algorithms and introduced competing envelope cues. However, it has been argued that human peripheral tuning may be two or more times sharper than previously estimated, such that envelope restoration may be stronger than originally thought. In this study, we utilize a computational model of the auditory periphery to investigate how cochlear tuning affects the restoration of envelope cues in auditory nerve responses to "TFS speech." Both the envelope-normalization algorithm of Lorenzi et al. (Proc Natl Acad Sci USA 103:18866–18869, 2006) and the speech-noise chimaeras of Smith et al. (Nature 416:87–90, 2002) were evaluated. The results for the two processing algorithms indicate that the competing noise envelope of the chimaeras better reduces speech envelope restoration but does not totally eliminate it. Moreover, envelope restoration is greater if the cochlear tuning is adjusted to match Shera and colleagues' (Proc Natl Acad Sci USA 99:3318–3323, 2002) estimates of human tuning.

Keywords Speech Perception • Auditory Models • Cochlear Tuning • Envelope Cues • Temporal Fine Structure • Auditory Chimaeras • Spectro • Temporal Modulation Index

I.C. Bruce (✉)
Department of Electrical and Computer Engineering, McMaster University, Room ITB-A213,
1280 Main Street, West Hamilton, ONT L8S 4K1, Canada
e-mail: ibruce@ieee.org

E.A. Lopez-Poveda et al. (eds.), *The Neurophysiological Bases of Auditory Perception*,
DOI 10.1007/978-1-4419-5686-6_40, © Springer Science+Business Media, LLC 2010

40.1 Introduction

Speech signals can be characterized by two types of coding: temporal envelope (E) coding and temporal fine structure (TFS) coding (Smith et al. 2002; Lorenzi et al. 2006). Temporal envelope coding is the relatively slow variation in amplitude over time, while the fine structure coding is the rapid variations in the signal. Both types of temporal coding contain some information (cues) that can be used for speech perception. The envelope information is embedded in the variations of the average discharge rate of the auditory nerves (ANs), whereas the TFS cues are coded as the synchronization of the nerve spikes to a particular phase of the stimulus, which is known as the phase-locking phenomenon (Young 2008). It has been suggested that the E information is responsible for speech understanding, while the TFS information is linked to melody perception and sound localization. Recently, some studies pointed to the possibility that the TFS code has a role in speech perception, especially in complex background noise (Lorenzi et al. 2006). It has been observed that normal-hearing people perform much better than hearing-impaired people in fluctuating background speech intelligibility tests. This was linked to a lack of ability to use TFS cues in hearing-impaired people, which accordingly indicates that TFS has a significant role for speech intelligibility. Such possibilities may have some implications on the development of cochlear implants, which in the current systems do not provide TFS information and are concerned mainly with providing E information. However, it has been argued that these results are complicated by possible restoration of E cues by the passing of the TFS signal through the cochlear filters (e.g., Zeng et al. 2004). Indeed, the preliminary results of Heinz and Swaminathan (2008) from AN recordings in chinchillas indicate that envelope restoration is not fully eliminated even for 8 or 16 processing channels. Furthermore, envelope restoration may be more significant in humans, if human cochlear tuning is sharper than in other mammals, as has been recently suggested (Shera et al. 2002). In this study, we investigate the role of TFS coding in speech perception using a model of the auditory periphery and a cortical model of speech processing. To separate the TFS code from E information, speech signals are divided into frequency bands to extract the envelope and fine structure codes in each band. The input stimulus to our auditory model is either the TFS-only signal or auditory chimaeras. Auditory chimaeras (Smith et al. 2002) are created such that the E (TFS) is from the speech signal, while the TFS (E) is coming from a spectrally matched noise signal. The effects of human versus cat cochlear tuning on envelope restoration are explored.

40.2 Methods

40.2.1 The Auditory Periphery Model

Our model is based on Zilany and Bruce model for the cat auditory periphery (Zilany and Bruce 2006, 2007a). The model consists of several blocks representing different stages in the auditory periphery from the middle ear to the AN. The model can accurately represent the nonlinear level-dependent cochlear effects, and it provides an accurate description of the response properties of AN fibers to complex stimuli in both normal and impaired ears. However, the model is designed to simulate the auditory periphery system in cats, and there are some important differences between cat and human ears (Recio et al. 2002) that should be taken into account in the final model.

We have developed a version of the computational model for the ear that incorporates human cochlear tuning. The Q_{ERB} values for the human cochlear filter have been estimated in Shera et al. (2002) using stimulus frequency otoacoustic emissions and improved psychophysical measurements.

The cochlear frequency selectivity is represented by Q_{ERB}, defined as

$$Q_{\mathrm{ERB}}(\mathrm{CF}) = \frac{\mathrm{CF}}{\mathrm{ERB(CF)}} \tag{40.1}$$

The characteristic frequency (CF) is the center frequency of the filter, and ERB is the equivalent rectangular bandwidth, defined as the bandwidth of the rectangular filter that produces the same output power as the original filter when driven by white noise. In Fig. 40.1a, we show the human Q_{ERB} values as function of CF given in Shera et al. (2002). The Q_{ERB} values reported in Shera et al. (2002) are two or three times sharper than the previous behavioral measurements (Glasberg and Moore 1990) shown in the same figure. In our work, we mapped the Q_{ERB} values derived in Shera et al. (2002) to the corresponding Q_{10} values to set the tuning in the computational model. The mapping is illustrated in Fig. 40.1b, where Q_{10} and Q_{ERB} values are computed at each center frequency using the model's cochlear filter transfer function. A linear mapping between Q_{10} and Q_{ERB} is then estimated using least square curve fitting to obtain

$$\mathrm{Q}_{10} = 0.2085 + 0.505\mathrm{Q}_{\mathrm{ERB}}. \tag{40.2}$$

The cat Q_{ERB} versus CF curve is also shown in Fig. 40.1a, where we have calculated the cat Q_{ERB} values from the corresponding Q_{10} values in the Zilany and Bruce (2006, 2007a) model using the mapping above.

40.2.2 Speech Intelligibility Metric (STMI)

The output of the model is assessed to predict the speech intelligibility based on the neural representation of the speech. This is achieved through the spectro-temporal modulation index (STMI; Elhilali et al. 2003; Bruce and Zilany 2007; Zilany and Bruce 2007b). A simple model of the speech processing of the auditory cortex assumes an array of modulation selective filter banks, which are referred to as spectro-temporal response fields. The output of the AN model is represented by a time-frequency "neurogram." The neurogram is made up from the averaged discharge rates (over every 8 ms) from 128 AN fibers with CFs ranging from 0.18 to 7.04 kHz.

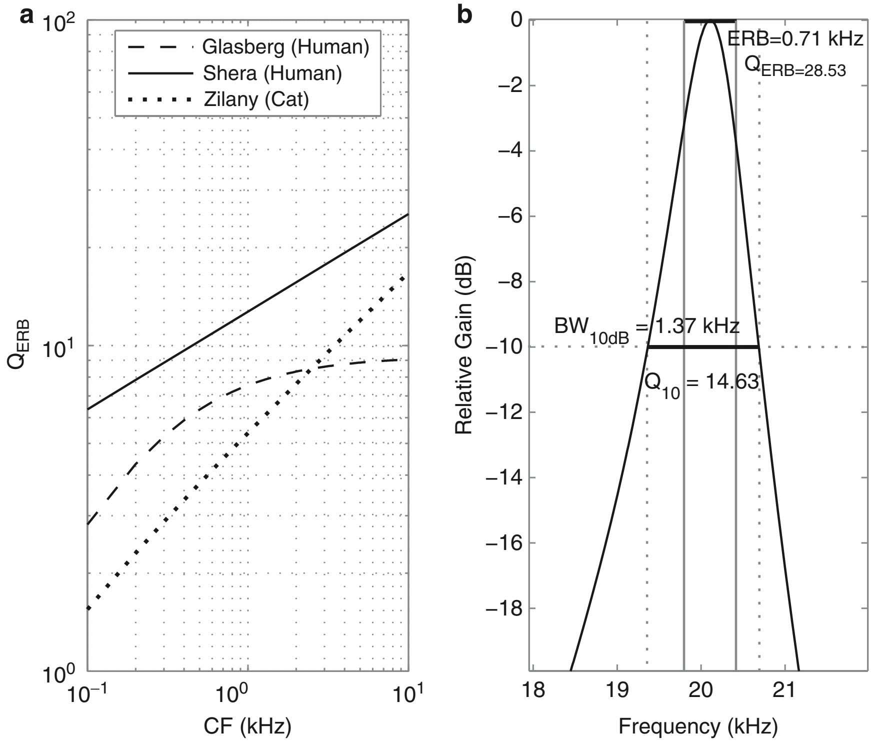

Fig. 40.1 (**a**) Comparison of the human Q_{ERB} values as a function of CF given in Shera et al. (2002), the earlier human Q_{ERB} data in Glasberg and Moore (1990), and the cat Q_{ERB} values in Zilany and Bruce (2006, 2007a). (**b**) Example illustrating Q_{10} to Q_{ERB} mapping for an AN filter at a CF of 20.107 kHz

This neurogram is processed by a bank of modulation selective filters to compute the STMI. The rates for temporal modulations of the filters range from 2 to 32 cyc/s (Hz), and the scales for spectral modulations are in the range from 0.25 to 8 cyc/oct. The STMI is computed using a template (the expected response) generated as the output (at the cortical stage) of the normal model to the stimulus at 65 dB SPL. The cortical output of the test stimulus is compared to the template, and the STMI is computed according to the formula,

$$\mathrm{STMI} = \sqrt{1-\frac{\|\mathrm{T}\text{-}\mathrm{N}\|^2}{\|\mathrm{T}\|^2}} \tag{40.3}$$

where, *T* is the cortical output of the template signal, and *N* is the cortical output of the test stimulus.

Because of the large time bins in the AN neurogram and the slow temporal modulation rates for the cortical filters, the STMI is only sensitive to spectral and temporal modulation in the neural response to speech – all phase-locking information about TFS cues is filtered out. Consequently, STMI values computed for TFS-only speech or speech TFS-noise E chimaeras are dependent only on the E cues present in the signal and any E cues restored by passing the TFS signal through the cochlear filters in the auditory periphery model.

40.2.3 Auditory Chimaeras

In Smith et al. (2002), a method to separate TFS from E cues was presented. Two acoustic waveforms are processed using a bank of band pass filters followed by Hilbert transforms to generate envelope-only and TFS-only versions of the signals. In each band, the envelope of one waveform is multiplied by the TFS of the other. The products are then summed across frequency bands to construct the auditory chimaeras. We may have speech–speech chimaeras, where both waveforms are sentences. We may also produce speech–noise chimaeras, where one waveform is the speech signal and the other is noise.

40.2.4 TFS-Only Signals

In Lorenzi et al. (2006), the role of TFS cues in speech perception is assessed by presenting TFS-only signals to a group of normal and hearing-impaired listeners and recording the intelligibility results over several sessions of training. TFS-only signals are generated in a similar method to that of Smith et al. (2002), as they both have the same technique for processing speech signals in each frequency band to separate TFS from E information. However, some distinctive differences exist between the two approaches. First, the number of frequency bands used in Lorenzi et al. (2006) is fixed (16 frequency bands), while in Smith et al. (2002) different choices are tested (from only 1 filter band up to 16 filter bands). Second, the speech TFS-only signal is used directly as the sound stimulus in Lorenzi et al. (2006), while in Smith et al. (2002) they use the speech TFS-only signal to first modulate a noise E-only signal creating auditory chimaeras, which are then used as the new stimulus.

40.3 Test Speech Material

In our work, we have used 11 sentences from the TIMIT database, randomly selected for different male and female speakers from eight major dialect regions of the United States. The sentences are used to create auditory chimaeras following the same procedure as in Smith et al. (2002). Each sentence signal is filtered using a number of band-pass filters. In our study, we have used different number of filters (1, 2, 3, 4, 6, 8, and 16) to divide the signal into frequency bands. These filters are designed as Butterworth filters of order 6, with cutoff frequencies determined such that the filters cover frequency range from 80 to 8,820 Hz with logarithmic frequency spacing (Smith et al. 2002). In each band, we compute the signal envelope using the Hilbert transform. Note that, when comparing our results to Lorenzi et al. (2006), we only use 16 frequency bands to separate the TFS and E signals.

To reproduce the stimulus signals created in Smith et al. (2002), we constructed a spectrally matched noise signal for each test sentence of the TIMIT database. The noise signal is processed in the same way as the sentence signals to produce the envelope and TFS for the noise waveform in each frequency band. The two waveforms, sentence signal and noise signal, are combined to form the speech–noise auditory chimaeras. For every sentence of the 11 TIMIT examples and for each choice of the number of frequency bands used, two sets of chimaeras are developed: speech E-noise TFS chimaeras, and speech TFS-noise E chimaeras. These chimaeras are provided to our auditory periphery model to compute the output neurogram, which is then assessed to evaluate the extent of speech intelligibility using the STMI metric. The experiment is repeated for each stimulus, and the results are averaged over all sentences in the same speech–noise chimaeras set. STMI values were computed both with the original cat cochlear tuning of Zilany and Bruce (2006, 2007a) and the human tuning of Shera et al. (2002).

40.4 Results

We have compared our STMI results to the intelligibility scores reported in Lorenzi et al. (2006). We computed the cat and human TFS-only STMI values for the case of 16 filter bands averaged over all test sentences. Our STMI result from the cat auditory model is 0.3806, while a value of 0.4849 is obtained from the human auditory model. In order to get a better understanding of these results, we computed the STMI of a white Gaussian noise (WGN) only stimulus. Our STMI results in this case are 0.2769 in cats and 0.298 for humans, indicating the lowest possible values for the STMI. It can be concluded therefore that even with 16 filter bands, there is still some restoration of E cues from the “TFS-only” speech of Lorenzi et al., and this restoration is enhanced with human cochlear tuning.

In order to reduce (or eliminate) any E cues that might be recovered by the TFS-only signal, we have generated speech TFS-WGN E auditory chimaeras. The average STMI results in this case are 0.3466 for cats and 0.4252 for humans. It can be seen that the average STMI values are reduced for these chimaeras from the TFS-only values, indicating that introducing the noise E cues does diminish somewhat the restoration of speech E cues from the speech TFS signal, but restoration is not completely eliminated.

We have computed STMI values for both cat and human tuning using the auditory chimaeras which we have generated as in Smith et al. (2002). In Fig. 40.2, we display our STMI results for cats and humans together with the intelligibility scores obtained in Smith et al. (2002). The STMI for speech E-noise TFS is monotonically increasing with the number of filter bands, while the speech TFS-noise E starts

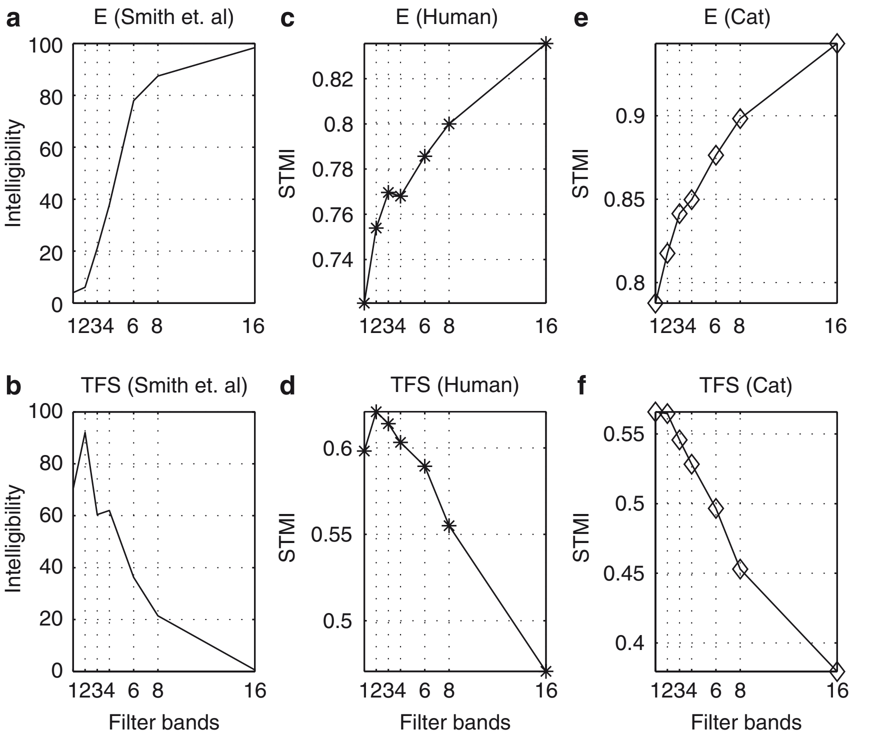

Fig. 40.2 Speech reception of sentences versus number of filter bands in (**a**) speech E-noise TFS chimaera and (**b**) speech TFS-noise E chimaera as in Smith et al. (2002). Average STMI values versus number of filter bands for (**c**) speech E-noise TFS chimaeras as the input to our human model and (**d**) speech TFS-noise E chimaeras. Average STMI values versus number of filter bands for (**e**) speech E-noise TFS chimaeras as the input to the cat model and (**f**) speech TFS-noise E chimaera

increasing with filter bands having a maximum value for two frequency bands, then it decreases with further increase in number of frequency bands. These results match well with the behavior of the intelligibility scores of Smith et al. (2002) as a function of number of filters, which are displayed in the same figure.

It is observed that the STMI values are higher for speech E-noise TFS than speech TFS-noise E over the entire range of numbers of filters. For the speech E-noise TFS signals, the STMI values for cat tuning are consistently higher than those for human tuning. This is due to the broader cat filters being less sensitive to degradation of the speech spectrum by the filter bank in the chimaeras algorithm. Comparing STMI values for cat and human tuning in the case of the speech TFS-noise E signals, scores are consistently higher with the human tuning than with the cat tuning. This observation is related to the narrower cochlear tuning incorporated in the human auditory periphery model. This narrow tuning implies better capability of the human auditory filters to restore E information from the TFS signal.

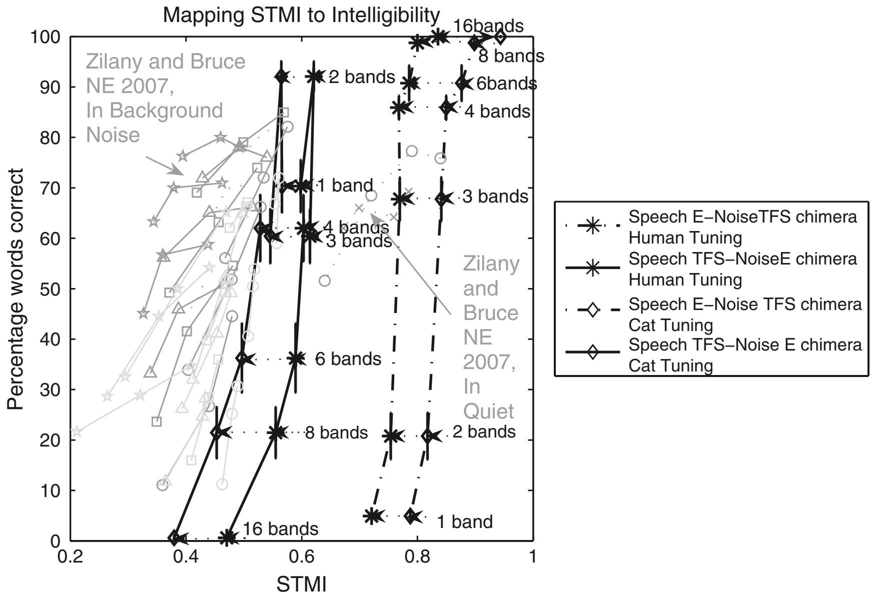

Fig. 40.3 Mapping curves between STMI and percent intelligibility explained in the legend (*thick black lines*), together with STMI-speech intelligibility mappings for cat tuning from Zilany and Bruce (2007b) for different signal-to-noise-ratio (SNR) values (*gray lines*) for comparison

Our STMI results for both cat and human tuning can be mapped to the corresponding intelligibility results obtained in Smith et al. (2002). Hence, for each (species) version of the model we have two mapping functions, one for the speech E-noise TFS chimaeras and the other for the speech TFS-noise E chimaeras. In Fig. 40.3, we display these STMI-intelligibility mapping curves (thick black lines), together with previous STMI-speech intelligibility mappings for cat tuning (Bruce and Zilany 2007; Zilany and Bruce 2007b) for different signal-to-noise-ratio (SNR) values (gray lines) for comparison. It can be observed that the mappings for the speech TFS-noise E signals with cat tuning are somewhere between the mappings obtained in Zilany and Bruce (2007b) for speech in noise and speech in quiet. The human mapping is shifted to the right, closer to the Zilany and Bruce (2007b) mappings for speech in quiet.

If the speech TFS-noise E intelligibility results of Smith et al. (2002) were due entirely to envelope restoration, then it might be expected that the mapping function for these signals would be identical to that for the speech E-noise TFS signals. This is clearly not the case for cat cochlear tuning. For human tuning, the mappings for 6–16 channels for both speech E-noise TFS and speech TFS-noise E signals do appear to be consistent with an extrapolation of the Zilany and Bruce (2007b) mappings for speech in quiet.

40.5 Conclusions

In this work, we show that STMI values for speech TFS-noise E chimaeras attain a maximum value for 1 and 2 frequency bands and then decline consistently with any further increase in bands. For 1 and 2 frequency bands, the narrowband human auditory filters can restore E cues from the TFS signal. Hence, the relatively high scores for 1 and 2 filter bands are obtained from the recovered E cues. Similar conclusions have been presented in Zeng et al. (2004), where it was argued that E recovery from TFS cues is the main reason for the relatively high intelligibility scores for small numbers of frequency bands. From our results, the STMI values obtained for the TFS-only case with 16 channels could almost completely explain the initial speech intelligibility scores for normal-hearing listeners in the Lorenzi et al. (2006) study, consistent with the observation of Heinz and Swaminathan (2008). The dependence of the envelope-restoration phenomenon on the number of filter bands in the processing algorithm and the bandwidth of the cochlear filters is illustrated by the STMI scores for the cat auditory model, where the cochlear filters are wider than the human model. In this case, the ability to recover E cues from TFS-only signals is reduced, and the STMI value is consequently less than the human tuning version. This observation is very important as it supports the theory that TFS information is used indirectly by the cochlea to recover E information, which is then used for speech understanding. This also explains the reduced ability of hearing-impaired people to benefit from TFS-only information as observed in Lorenzi et al. (2006). Since hearing-impaired people suffer from the broadening of the cochlear tuning, the recovery of E cues from TFS information is degraded and hence speech intelligibility is reduced.

However, a consistent mapping between STMI and speech intelligibility for the two types of chimaeras was not obtained for small numbers of channels. Preliminary results indicate that this may be due to the effects of the matched noise used in constructing the chimaeras on the model neural response. Future work should be concentrated on how the use of such matched noise, rather than an independent noise or a flat envelope, affects both STMI values and speech intelligibility in humans in the case of a small number of processing bands.

Acknowledgments The authors thank Dr. Zach Smith for providing speech intelligibility data. This work was funded by NSERC Discovery Grant 261736.

40.6 Comment by Michael Heinz

We recently quantified envelope recovery for chimaeric speech in recorded and modeled auditory-nerve responses and came to very similar conclusions as in your paper (Heinz and Swaminathan 2009). Your STMI model predictions provide an interesting complement ,in that they provide a prediction of recovered envelope cues at central levels, which could be present due to peripheral and/or central recovery of speech envelope cues from TFS cues. Have your STMI predictions provided any insight into the potential for envelope recovery to occur at central levels?

Heinz MG, Swaminathan J (2009) Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. J Assoc Res Otolaryngol 10 (3).

40.7 Reply Rasha Ibrahim

Thanks for your comments. Our present results indicate Envelope recovery from TFS cues at the peripheral level. It is interesting to investigate if any Envelope recovery can occur at the central level also. This might be considered in our future work to complement our study.

References

Bruce IC, Zilany MSA (2007) Modeling the effects of cochlear impairment on the neural representation of speech in the auditory nerve and primary auditory cortex. Auditory signal processing in hearing-impaired listeners, International Symposium on Audiological and Auditory Research (ISAAR), Denmark,1–10, August 2007

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun 41:331–348

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138

Heinz MG, Swaminathan J (2008) Neural cross-correlation metrics to quantify envelope and fine-structure coding in auditory-nerve responses. J Acoust Soc Am 123:3056

Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc Natl Acad Sci USA 103:18866–18869

Recio A, Rhode WS, Kiefte M, Kluender KR (2002) Responses to cochlear normalized speech stimuli in the auditory nerve of cat. J Acoust Soc Am 111:2213–2218

Shera CA, Guinan JJ Jr, Oxenham AJ (2002) Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc Natl Acad Sci USA 99:3318–3323

Smith ZM, Oxenham AJ, Delgutte B (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90

Young ED (2008) Neural representation of spectral and temporal information in speech. Philos Trans R Soc Lond B Biol Sci 363:923–945

Zeng FG, Nie KB, Liu S, Stickney G, Del Rio E, Kong YY, Chen HB (2004) On the dichotomy in auditory perception between temporal envelope and fine structure cues. J Acoust Soc Am 116:1351–1354

Zilany MSA, Bruce IC (2006) Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J Acoust Soc Am 120:1446–1466

Zilany MSA, Bruce IC (2007a) Representation of the vowel /ε/ in normal and impaired auditory-nerve fibers: model predictions of responses in cats. J Acoust Soc Am 122:402–417

Zilany MSA and Bruce IC (2007b) Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery. Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering, NJ, May 2007, pp 481–485