Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 668–683
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise

David P. Messing [a,*,1], Lorraine Delhorne [a], Ed Bruckert [b], Louis D. Braida [a], Oded Ghitza [b]

[a] *Massachusetts Institute of Technology, Cambridge, Massachusetts, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, USA*
[b] *Sensimetrics Corporation, Malden, Massachusetts, USA*

Received 16 June 2008; received in revised form 20 February 2009; accepted 20 February 2009

Abstract

Current predictors of speech intelligibility are inadequate for understanding and predicting speech confusions caused by acoustic interference. We develop a model of auditory speech processing that includes a phenomenological representation of the action of the Medial Olivocochlear efferent pathway and that is capable of predicting consonant confusions made by normal hearing listeners in speech-shaped Gaussian noise. We then use this model to predict human error patterns of initial consonants in consonant–vowel–consonant words in the context of a Dynamic Rhyme Test. In the process we demonstrate its potential for speech discrimination in noise. Our results produced performance that was robust to varying levels of stationary additive speech-shaped noise and which mimicked human performance in discrimination of synthetic speech as measured by the Chi-squared test.

Keywords: MOC efferent; Speech recognition; Noise robustness; Human confusions; MBPNL model

1. Introduction

Current models of speech intelligibility are inadequate for making predictions of speech confusions caused by acoustic interference for normal-hearing listeners. The Articulation Index (French and Steinberg, 1947; ANSI, 1969) and related measures, STI (Houtgast et al., 1980), and SII (ANSI, 1997) characterize hearing in a manner geared to the task of predicting speech intelligibility. But such measures only predict average speech intelligibility, not error patterns, and they make predictions for only a limited set of acoustic conditions (linear filtering, reverberation, and additive noise).

The performance of current speech recognition systems using front-ends such as the Mel-Filter Bank (MFB), Mel-Filtered Cepstral Coefficient (MFCC), and the Ensemble Interval Histogram (EIH) models degrades significantly in the presence of noise. At the same time however, human performance on speech recognition is more robust to noise (Lippmann, 1997; Sroka and Braida, 2005). Lippman suggests that this human–machine performance gap can be reduced by improving low-level acoustic–phonetic modeling, improving robustness with noise and channel variability, and more accurately modeling spontaneous speech.

This work is inspired by the desire to understand, predict, and mimic human speech confusions caused by acoustic interference. Our long-term goal is to formulate a matching operation, with perception-related rules of integration over time and frequency at its core, in the context of human processing of degraded speech, but in this paper we concentrate on separating the back-end development from the front-end. Our approach is to attempt to reduce

* Corresponding author. Tel.: +1 617 939 4192.
E-mail addresses: dpmessing@gmail.com (D.P. Messing), delhorne@MIT.EDU (L. Delhorne), Ebruckert@fonix.com (E. Bruckert), ldbraida@mit.edu (L.D. Braida), oded.ghitza@gmail.com (O. Ghitza).
[1] Present address: 11208 Vista Sorrento, Pkwy J306, San Diego, CA 92103, USA.

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.02.002

the influence of cognitive and memory factors while preserving the complex acoustic cues that differentiate diphones. Hence we attempt to tune the parameters of the peripheral auditory model in as much isolation as possible by reducing the effect of the back-end system.

In this paper, we focus on developing a model inspired by the non-linear efferent feedback and signal processing of the human auditory periphery, which is thought to aid speech recognition in noise environments. Specifically we develop our model to attempt to match and predict human confusions of initial consonants in speech-shaped additive Gaussian noise. We take current knowledge about medial olivocochlear (MOC) efferents and use it to create a model that can explain human performance and behavior. Other mechanisms such as neural adaptation at levels at and higher than the auditory nerve or the use of differing-rate (high, medium, and low) spontaneous-rate fibers may potentially also explain human confusions and performance. However the need for high level adaptation depends on the processing done at the level of the auditory nerve, and such theories are also speculative in nature. Instead of developing and examining all possible theories in detail (which would require much more work than is possible in the scope of a single paper such as this), we focus on the development and examination of a single MOC-based theory to explain human performance in noise.

This paper is divided into four sections. In Section 2 we discuss the background, focusing on MOC efferents; in Section 3 we describe our model. In Section 4 we fine-tune the model and match it to human performance. In Section 5 we discuss the implications of our work. Finally in Section 6 we summarize and discuss possible future research.

2. Background: MOC efferents

Mounting physiological data exists in support of the effect of MOC efferents on the mechanical properties of the cochlea and, in turn, on signal properties at the auditory nerve level, in particular when the signal is embedded in noise. MOC efferent activity is believed to reduce outer hair cell (OHC) motility and change OHC shape, resulting in increased basilar membrane stiffness, which in turn inhibits inner hair cell (IHC) response in the presence of noise. This paper develops this picture into a closed-loop model of the peripheral auditory system, a model that adaptively adjusts its cochlear mechanics based on the processed noise energy level. The next few sections summarize recent work related to MOC efferents.

2.1. Morphology and physiology

Detailed morphological and neurophysiological description of the MOC efferent feedback system are available (e.g., Gifford and Guinan, 1983; Guinan, 1996; Kawase and Liberman, 1993; Liberman, 1988; Liberman and Brown, 1986; May and Sachs, 1992; Warr, 1978; Winslow and Sachs, 1988). MOC efferents originate from neurons medial, ventral and anterior to the medial superior olivary nucleus (MSO), have myelinated axons, and terminate directly on OHCs. Medial efferents project predominantly to the contralateral cochlea (the innervation is largest near the center of the cochlea) with the crossed innervation biased toward the base compared to the uncrossed innervation (e.g., Guinan, 1996). Roughly two-thirds of MOC efferents respond to ipsilateral sound, one-third to contralateral sound, and a small fraction to sound in either ear. MOC efferents have tuning curves that are similar to, or slightly wider than, those of auditory nerve (AN) fibers (e.g., Liberman and Brown, 1986), and they project to different places along the cochlear partition in a tonotopical manner. Finally, medial efferents have longer latencies and group delays than AN fibers. In response to tone or noise bursts, most MOC efferents have latencies of 10–40 ms. Group delays measured from modulation transfer functions are much more tightly clustered, averaged at about 8 ms (Gummer et al., 1988).

Current understanding of the functional role of the MOC efferent feedback mechanism is incomplete. A few suggestions have been offered, such as shifting of sound-level functions to higher sound levels, resolution of transient sounds in a continuous masker, or preventing damage due to intense sound (e.g., Guinan, 1996). One speculated role, which is of particular interest for this work, is a dynamic regulation of the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background (e.g., Kiang et al., 1987). Several neurophysiological studies support this role. Using anesthetized cats with noisy acoustic stimuli, Winslow and Sachs (1988) showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is partly recovered. Measuring neural responses of awake cats to noisy acoustic stimuli, May and Sachs (1992) showed that the dynamic range of discharge rate at the Anterior Ventral Cochlear Nucleus (AVCN), which is tightly correlated to the rate of the AN, is only moderately affected by changes in levels of background noise. Both studies indicate that MOC efferent stimulation plays a role of regulating the AN fiber response in the presence of noise.

2.2. Psychophysics: evidence for efferent involvement in noise

A few behavioral studies indicate the potential role of the MOC efferent system in perceiving speech in the presence of background noise. Dewson (1968) presented evidence that MOC lesions impair the abilities of monkeys to discriminate the vowel sounds [i] and [u] in the presence of masking noise but have no effect on the performance of this task in quiet. More recently, Giraud et al. (1997), and Zeng et al. (2000) showed that human subjects who have undergone a vestibular neurectomy (presumably resulting in a reduced MOC feedback) exhibit degraded phoneme

perception when the speech is presented in a noisy background. These speech reception experiments may be contaminated by surgical side effects such as uncertainties about the extent of the lesion and possible damage to cochlear elements. Thus the results are somewhat controversial. Some studies (e.g. Scharf et al., 1997) showed little or no perceptual effect in situations where one might expect efferents to play a role. Ghitza (2004) explored the effects of the MOC efferent system by presenting combinations of speech and noise in various configurations (gated/continuous, monaural/binaural). His results showed a gated/continuous difference analogous to the "masking overshoot" in tone detection: the results with gated noise were worse than the results with continuous noise. These results were similar to findings by Ainsworth and Meyer (1994), Ainsworth and Cervera (2001), and Cervera and Gonzalez-Alvarez (2007) that also showed lower performance of gated noise than continuous noise. Ghitza suggested these results could be due to efferent inability to activate quickly for gated conditions, but cautioned that the results could also be due to high-order auditory and cognitive mechanisms such as those observed in the fusion of perceptual streams. Despite the concerns in all of the above studies, these results can be interpreted to support the hypothesis of a significant efferent contribution to initial phone discrimination in noise.

2.3. Recent work in modeling

Ghitza et al. (2007) presented some of our preliminary findings and work, focusing on model performance on a speech identification task. In this paper we continue the development of ideas presented there, with a focus on mimicking human performance on identification of initial consonants in consonant–vowel–consonant tokens in noise.

In tandem with our work, Ferry and Meddis (2007) have developed a similar model based on Meddis' Dual Resonance Nonlinear (DRNL) model of the auditory system. In that work, Ferry and Meddis focused on matching model parameters to basilar membrane (BM) response, AN activity, and compound action potential (CAP) responses measured in various animal studies with center frequencies (CF) between 3550 and 20 kHz. One of the challenges in doing this is that much more data exists for higher frequency fibers while not much if any exists for frequencies below 8 kHz, ranges that are essential for speech perception. Furthermore, all AN and CAP measurements are for various non-human mammals and may not match human data. Since we were more interested in the frequency range relevant to speech and human speech perception, our approach is instead to model the known biological components and focus on tuning our system parameters to match human speech confusions. We also chose to use Goldstein's Multi Band Pass Non Linear (MBPNL) model of nonlinear cochlear mechanics (Goldstein, 1990) instead of using the DRNL model. Both MBPNL and DRNL models mimic cochlear mechanics, with "tip" and "tail" paths that control the non-linear CF amplification of a signal and the filter width respectively. However, the DRNL model is a linear mixing model: the nonlinear tip compression happens before summation with the tail path of the model. The MBPNL model is a nonlinear mixing model: the nonlinear tip compression happens after summation with the tail path, resulting in a nonlinear interaction. This nonlinear mixing allows the MBPNL model to better mimic the measured nonlinear within-filter synchrony and rate suppression of two tones (in particular low frequency "tail" tone suppression of "tip" tones near CF) than linear-mixing models (for a further comparison of the DNRL and MBPNL models see Goldstein, 1990). This difference could be very important for the processing of complex signals such as speech, where multiple harmonics are found within an auditory filter. Despite these differences in approach and model development, our work, like Ferry and Meddis's, exploits the current popular theory of the role of the MOC efferent system and demonstrates performance gains in noise.

3. Model overview and qualitative evaluation

In this section we give an overview of our model and qualitatively demonstrate the ability of our closed-loop (with efferent feedback) model to produce spectrographic displays of noisy speech that are more consistent with displays of speech in quiet than are displays produced by open-loop (without efferent feedback) models. In Section 4 we will provide a quantitative analysis of our model.

We begin by describing Goldstein's Multi Band Pass Non Linear (MBPNL) model of nonlinear cochlear mechanics which is a major component of our overall model. Then we describe our open-loop model based on this MBPNL cochlear filterbank. Finally we describe our closed-loop model which adjusts the parameters of the MBPNL model depending on the efferent response to noise.

3.1. MBPNL cochlear filterbank

In this subsection, we review and discuss Goldstein's MBPNL model of nonlinear cochlear mechanics (Goldstein, 1990). This model changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior.

The MBPNL model is shown in Fig. 1. The lower path (H1/H2) is a compressive nonlinear filter that represents the sensitive, narrowband compressive nonlinearity at the tip of the basilar membrane tuning curves. The upper path (H3/H2) is a linear filter (the expanding function preceded by its inverse compressive function results in a unitary transformation) that represents the insensitive, broadband linear tail response of basilar membrane tuning curves. The gain parameter (GAIN) controls the gain of the tip of the basilar membrane tuning curves, and is used to model the inhibitory efferent-induced response in the presence of noise. For the open-loop MBPNL model GAIN is set to

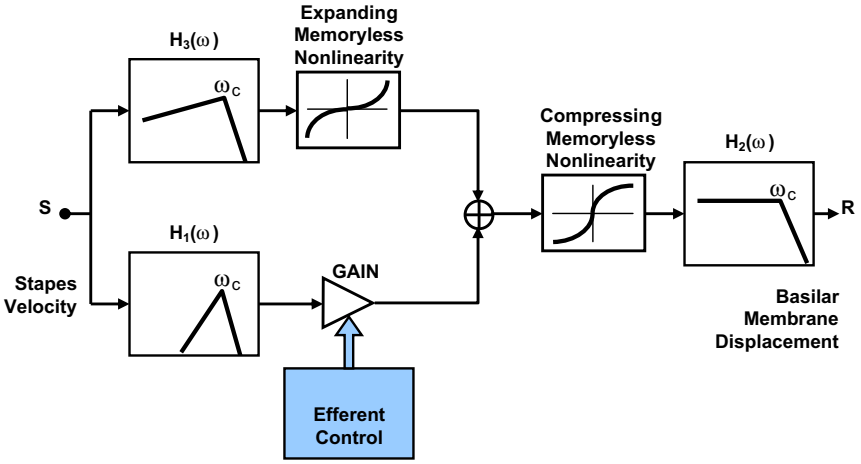

Fig. 1. MBPNL filterbank. A parameter GAIN controls the gain of the tip of the basilar membrane tuning curves. To best mimic psychophysical tuning curves of a healthy cochlea in quiet, the tip gain is set to GAIN = 40 dB (Goldstein, 1990).

40 dB, to best mimic psychophysical tuning curves of a healthy cochlea in quiet (Goldstein, 1990).

The "iso-input" frequency response of an MBPNL filter at *CF* of 3600 Hz with various tip gain settings is shown in Fig. 2. For an input signal $s(t) = A\sin(2\pi f_o t)$, with A and f_o fixed, the MBPNL behaves as a linear system with a fixed "operating point" on the expanding and compressive nonlinear curves, determined by A. For a given A, a discrete chirp signal with a slow linear ramping frequency was presented to the system in order to measure the non-linear frequency response of the system to a sinusoid at each frequency. Changes in f_o occurred only after the system reached steady-state, for a proper gain measurement. The frequency response for the open-loop MBPNL model is shown at the

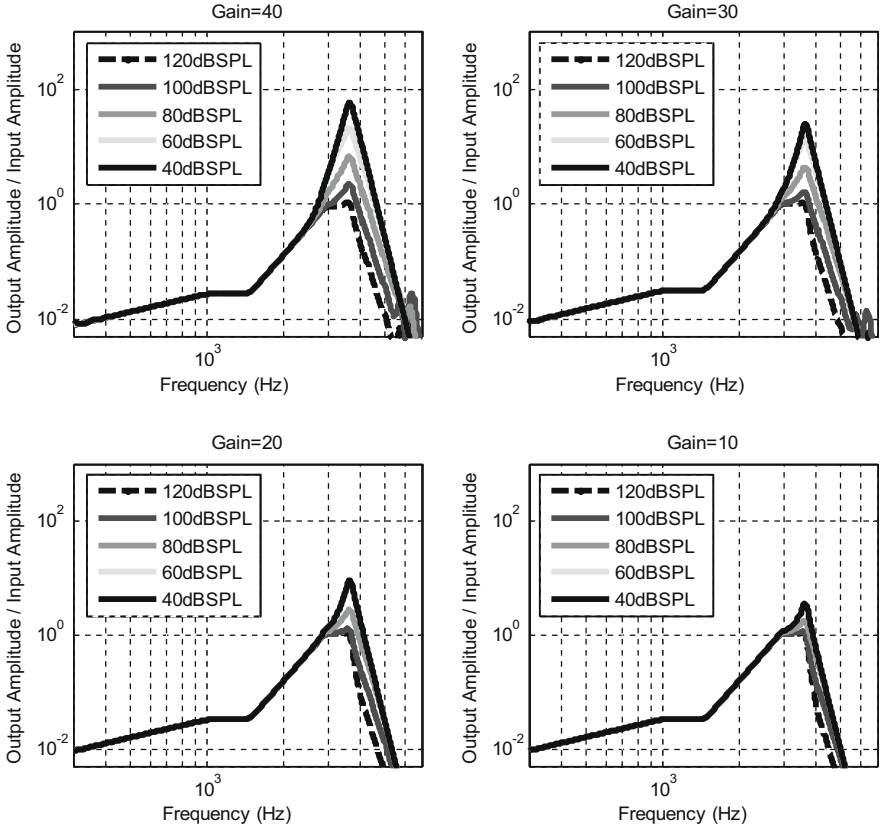

Fig. 2. MBPNL frequency responses Iso-input frequency responses of an MBPNL filter (at CF of 3641 Hz) for different values of GAIN parameter. From upper-left, clockwise: GAIN = 40, 30, 20 and 10 dB. Upper-left corner (Gain = 40 dB) is for healthy cochlea in quiet (Goldstein, 1990). Input sinusoids are varied from 40 dB SPL to 120 dB SPL.

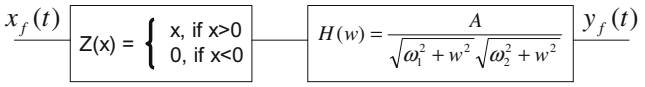

Fig. 3. Our model of the Inner Hair Cell (IHC). The model is fed the output from the cochlear filter bank. Each cochlear channel is processed by the same half-wave rectification followed by a low-pass "Johnson" filter. The "Johnson" filter is a 2^{nd} order lowpass filter with poles at 600 Hz and 3000 Hz. A is chosen to give the filter a unity gain in the pass-band. The combination of the two components produces an output that reflects nerve firing patterns while also mimicking loss of synchrony found in humans and cats as the CF of the cochlear filters is increased.

upper-left corner (i.e. for GAIN = 40 dB). Fig. 2 shows the iso-input frequency response of the system for different values of input SPL level. As the input level increases the output gain drops and the bandwidth increases, in accordance with physiological and psychophysical behavior (Glasberg and Moore, 1990). As the gain increases, the distance between the maximum and minimum peaks, which corresponds to inputs of 40 dB SPL and 120 dB SPL in Fig. 2, increases. In our closed-loop model, the tip GAIN parameter is adjusted based on the efferent response, which in turn is calculated based on the amount of noise present. For our experiments, we limited the range that GAIN can vary, based on biological observations. The adjustment of GAIN is described in more detail in Section 3.3.

3.2. Open-loop model with MBPNL filters

Our baseline open-loop model is displayed in Fig. 4. The first component is a middle ear module that mimics the high-pass frequency response of the middle ear. It consists

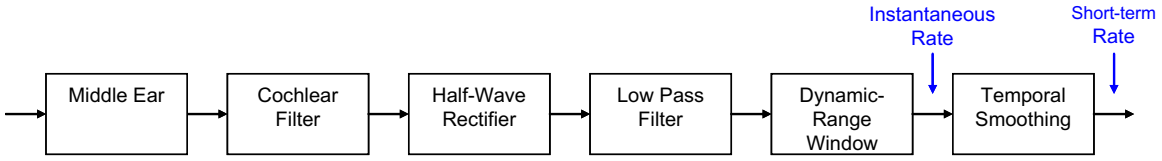

Fig. 4. Overview of one channel of the front-end model without efferent feedback. This open-loop model is composed of middle ear module followed by an MBPNL cochlear filter. The output of the filter is followed by a half-wave rectifier, low pass filter, and Dynamic Range Window (DRW), which together represent the IHC and nerve. The DRW, corresponds to the spontaneous rate and rate-saturation of the auditory nerve. After the DRW, the output is then smoothed with a trapezoidal window with 1 ms ramps that overlap to find the average short-term nerve firing rate.

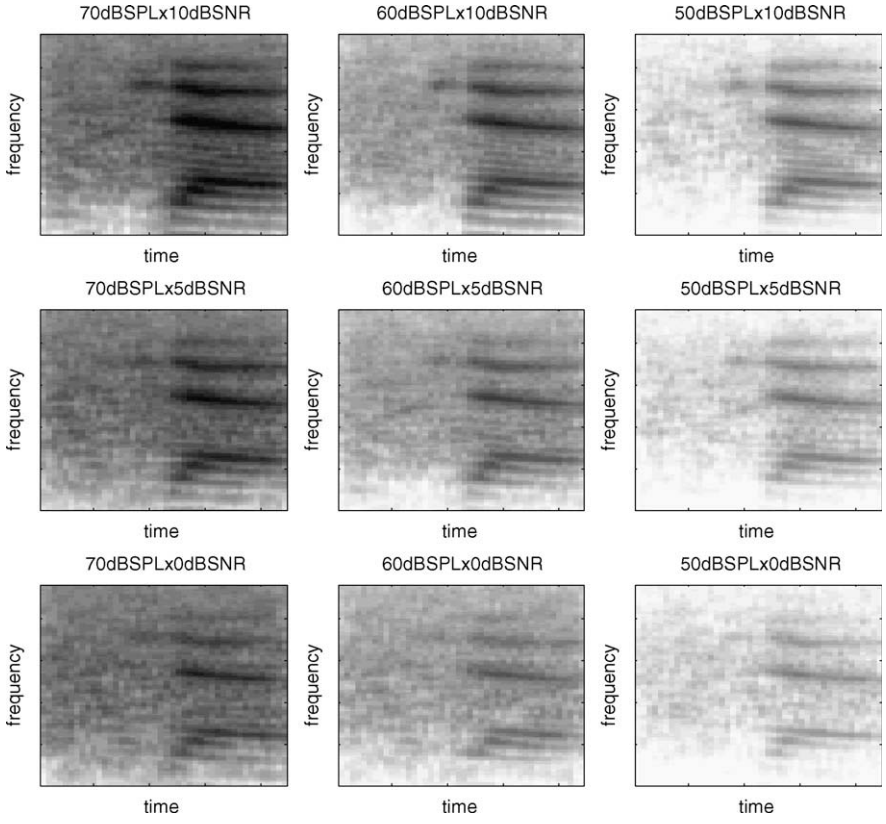

Fig. 5. Simulated IHC response to diphone/ja/ (250 ms), produced by an open-loop MBPNL model; fixed $G = 40$ dB; The upper bound of the DRW was chosen to be 130 dB to correspond roughly with the human threshold of pain. The lower bound was chosen to minimize the within-1-std metric and is 65 dB.

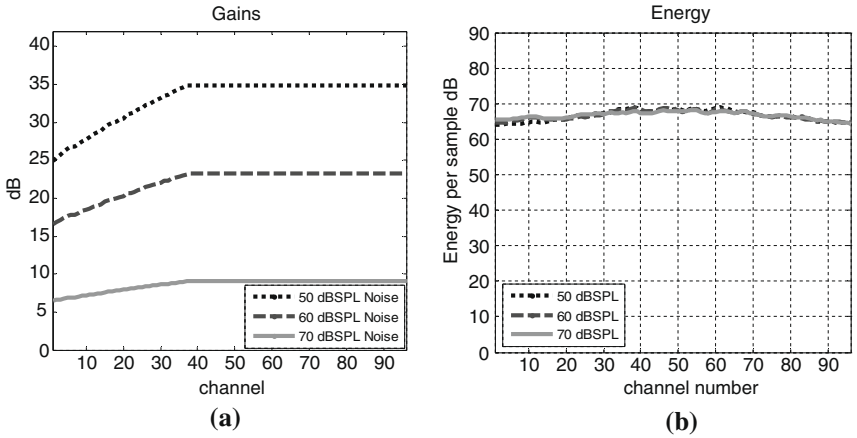

Fig. 6. Example of efferent gain regulating noise allowed above the DRW rate limiter. (a) Shows the efferent GAIN profile per cochlear channel for three different noise presentation levels (without speech) and (b) indicates the resulting noise energy per channel at the output. GAIN per noise condition are chosen to make the average total energy per channel constant across noise condition.

of a first-order high-pass filter $|H(f)|^2 = A^2 \frac{f^2}{f^2+f_0^2}$, with $A = .0014$ and $f_0 = 1000$ selected to approximate a gain of 0 dB at 1 kHz. The cochlear model is comprised of a bank of overlapping cochlear channels uniformly distributed along the ERB scale (Glasberg and Moore, 1990), four channels per ERB. Each cochlear channel comprises an MBPNL filter followed by a model of the IHC and nerve. Our model of the IHC is composed of a half-wave rectifier followed by a low-pass "Johnson" filter with poles at 600 Hz and 3000 Hz (see Fig. 3). The half-wave rectifier converts the input waveform of a cochlear channel into a nerve firing response. The filter mimics the loss of synchrony found in cats as the CF of the cochlear filters is increased (as described by Johnson, 1980); ie as CF increases, the bandwidth of the cochlear filters increase and information on the fine structure of the waveform is lost.

The dynamic range of the simulated IHC and nerve response is restricted – from below and above – to a "dynamic-range window" (DRW), representing the observed dynamic range at the AN level (i.e. the AN rate-intensity function); the lower bound and upper bound of the DRW represent the spontaneous nerve firing rate and saturation firing rate, respectively. The signal is then smoothed using overlapping N-ms (N is a variable that was adjusted and is typically set to 8, 10, or 12) trapezoidal windows with 1 ms cosine-squared ramps to find the short-term average nerve firing rate.

Fig. 5 provides a spectrographic display of the output of the open-loop system. The simulated IHC response is displayed for noise intensity levels of 70, 60, and 50 dB SPL and for SNR values of 20, 10, 5, and 0 dB (values that were used in the experiments discussed in Section 4). GAIN is 40 dB and is held constant for all SNR and noise levels. The figure contains a 3-by-3 matrix of images; the columns represent the intensity of the background noise. The rows represent SNR. Each image represents the responses to the diphone/ja/ (duration of 250 ms) spoken by a male speaker, with DRW = 65 dB. The upper bound of the DRW was chosen to be 130 dB to correspond roughly with the human threshold of pain. The lower bound was chosen to minimize our performance metrics, to best match human (discussed in Section 4), and is 65 dB. Large differences are observed across varying noise intensity and SNR levels. Note that for the DRW we chose, at 50 dB noise intensity much of the speech energy is not present in the response. Had the DRW range been shifted lower, more of the speech energy of the 50 dB noise intensity level would have been visible but also more noise. It proved impossible to find a DRW position that provides a consistent display at the output, across rows and columns.

3.3. Closed-loop model with efferent-inspired feedback

The closed-loop, efferent-inspired, MBPNL model is shown in Fig. 7. We introduce a CF dependent feedback mechanism which controls the GAIN of each MBPNL channel according to the intensity of sustained noise at that frequency band. Specifically, the GAIN parameter in Fig. 7 was adjusted to allow a prescribed amount of noise through each channel's DRW. As the lower bound of the DRW is increased, the tip-gain parameter needs to be increased to maintain the same amount of noise through each channel's DRW. Hence the choice of the lower bound affects the level of the GAIN.

Fig. 6 illustrates an example of how the efferent gain is adjusted to regulate the noise above the lower bound of the DRW rate limiter. In this figure, three separate speech-shaped Gaussian noise conditions are considered. The GAIN per channel is selected for each noise condition – 50 dB SPL (black dotted line), 60 dB SPL (grey dashed line), and 70 dB SPL (solid lighter grey line) noise – in the absence of speech. The GAIN for a given noise condition tapers off below 1 kHz, reflecting the decline in the

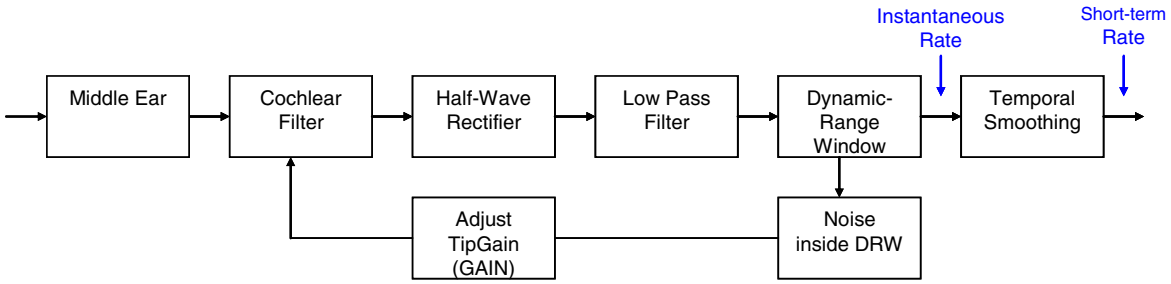

Fig. 7. Overview of one channel of the front-end model with efferent feedback. This closed-loop model is composed of the same blocks as the open-loop system with an efferent feedback response that is calculated based on the noise inside the DRW window for that particular channel. The tip-gain (GAIN) of each MBPNL filter (96 filters in total) is adjusted until the noise inside the DRW window of that channel reaches a desired level (this amount of noise was a parameter of our model and was adjusted).

number of MOC efferent nerves innervating lower frequency channels.

To determine the value of GAIN, the average noise energy per channel (with speech absent) is compared across noise conditions. If the resulting energies are not within a specified desired difference, the efferent gains are then iteratively adjusted and the resulting energies per condition are recomputed. This process is repeated until the average noise energies per channel are within a desired difference of each other. For our studies a difference of 0.1% was tolerated. As an example, when the target average noise energy per channel is 2 dB above the lower bound of the DRW, the output of the MBPNL filters for our model yields a response rate with energy per channel that fit the profile in Fig. 6b.

This adjustment of the GAIN parameter has several consequences. Besides making the energy of the noise at the output of each filter more consistent, it also affects the properties of each filter. The general effect is that loud noises reduce the non-linear amplification of small amplitude sounds while weak noises maintain the larger amplification of small amplitude sounds. Hence the overall effect

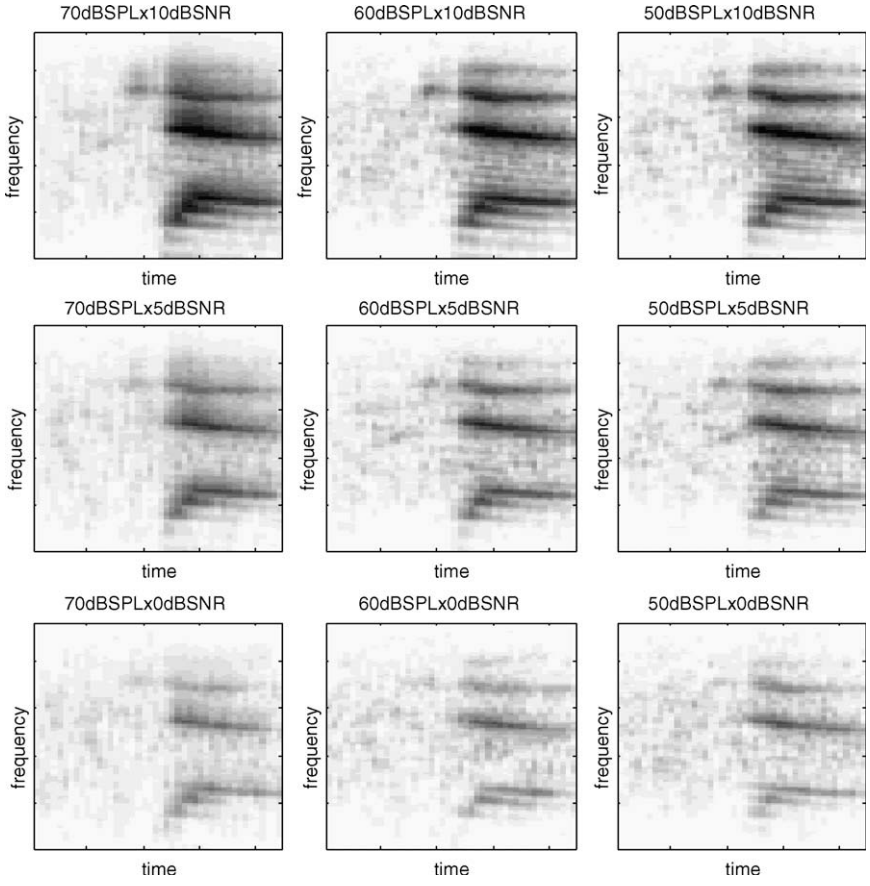

Fig. 8. Simulated IHC response to diphone/ja/, produced by the efferent-inspired closed-loop MBPNL. DRW is same as in open-loop MBPNL mode. Output of each of the 96 MBPNL filters is normalized to a fixed dynamic range.

of the efferent system in our model is to amplify small amplitude components of the speech stimulus by an amount that depends on the noise level. This point is illustrated in more detail in Fig. 2. In Fig. 2, the upper-left panel represents the nominal response (i.e. in quiet), with GAIN set to 40 dB. In this quiet condition, weaker amplitude sounds such as the 40 dB SPL sound are amplified greatly (in this case roughly 20 dB more) relative to louder sounds such as the 120 dB SPL stimulus. By increasing the efferent response in noise, we reduce the GAIN and the MBPNL response to weaker stimuli such as the 40 dB SPL tone (and background noise), as shown in the lower right pane of Fig. 2 where the GAIN parameter is set to 10 dB. Hence for high energy tone stimuli the MBPNL response is hardly affected, while the response for low energy stimuli (e.g. 40 or 60 dB SPL signals) is reduced by some 30 dB in the presence of noise.

Fig. 8 shows the spectrographic displays of the closed-loop MBPNL model. The DRW is set to 65 dB, with its position fixed at the same location as in the open-loop MBPNL model. The rate response of each MBPNL channel at the output of the DRW is stretched to the full dynamic range, i.e. the ouput of the IHC is proportionally stretched such that the minimal response rate of the signal after DRW clipping is stretched to the spontaneous level and the maximal rate of the signal after DRW clipping is stretched to the saturation level. The motivation for normalizing the IHC output stems from neurophysiological studies on anesthetized cats with noisy acoustic stimuli (Winslow and Sachs, 1988)[2]. In these studies, Winslow and Sachs show that, by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is recovered. Due to the nature of the noise-responsive feedback, the background noise is largely eliminated for all SPL × SNR conditions. Unfortunately some of the energy of the/j/ and higher vowel formants is also reduced. At a given SNR, displays of processed noisy speech, with stretching, are consistent across dB SPL noise level (rows in Fig. 8).

4. Model tuning and quantitative evaluation

As discussed in Section 1, our long-term objective is to predict consonant confusions made by normally-hearing listeners, listening to degraded speech. Our prediction engine comprises the efferent-inspired peripheral auditory model followed by a template matching operation. The extent to which this engine is an accurate model of auditory perception will be measured by its ability to predict consonant confusions in the presence of noise. This paper, however, focuses on the task of finding the parameters of the first stage with a minimal interference of the second.

[2] Concurring with this observation are measurements of neural responses of awake cats to noisy acoustic stimuli, showing that the dynamic range of discharge rate at the AN level is hardly affected by changes in levels of background noise (May and Sachs, 1992).

Ideally, to eliminate unwanted interaction between stages, errors due to template matching should be reduced to *zero*. In reality we could only try to *minimize* interaction by taking the following three steps: (1) we use the simplest possible psychophysical task in the context of speech perception, namely a binary discrimination test. In particular, we use Voiers' Diagnostic Rhyme Test (DRT) (1983) which presents the subject with a two alternative forced choice between two alternative CVC words that differ in their initial consonants. Such a task minimizes the influence of cognitive and memory factors while maintaining the complex acoustic cues that differentiate initial diphones; (2) we use the DRT paradigm with synthetic speech stimuli. An acoustic realization of the DRT word-pairs was synthesized so that the target values for the formants of the vowel in a word-pair are identical, restricting stimulus differences to the initial diphones; and (3) we use a "frozen speech" methodology (e.g. Hant and Alwan, 2003): the same acoustic speech token is used for training and for testing, so that testing tokens differs from training tokens only by the acoustic distortion.

These three steps presumably result in a reduction in the number of errors induced by the template matching. Recall that a template-match operation comprises measuring the "distance" of the unknown token to the templates, and labeling the unknown token as the template with the smaller distance. Hence, template matching is defined by the distance measure and the choice of templates. As a distance measure we use the mean- squared-error. This is an effective choice here because: (1) by using synthetic speech stimuli, the identical target values of the vowel formants for the two words results in zero error in time–frequency cells associated with the final diphone, and (2) by using frozen-speech stimuli, a distortion in a given time–frequency cell is generated *locally* (by noise component within the range of the cell) and is independent of noise at other cells.

In the rest of this section, we discuss the results of the human DRT tests, the DRT mimic, and our tuning of the model.

4.1. Human DRT

For the human DRT, six different subjects with normal hearing participated and were presented a DRT based on synthetic speech. One hundred and ninty two DRT diphones were generated with HLsyn, a modification of the Klatt synthesizer, and organized according to 96 word pairs (as per requirements for the DRT task), along four vowel quadrants (High-Front, High-Back, Low-Front, and Low-Back), and six acoustic dimensions (voicing, nasality, sustention, sibilation, graveness, and compactness). Noise was created by passing white noise through a linear filter with speech-shaped transfer function, which was obtained by averaging long-time Fourier power density spectrums for continuous speech across a large number of speakers (see Dunn and White, 1940). This noise was added to each word to obtain test tokens at various

presentation levels and SNR: 70 dB, 60 dB, and 50 dB SPL and 10 dB, 5 dB, and 0 dB SNR (levels were calculated based on rms values). Words in the database are divided into "runs" of 64 word-pairs, and the duration of one run is limited to about 3 min (to avoid fatigue). Three runs of 64 word-pairs make up a session which covers all 192 words in one repetition of one noise condition. Data for each noise condition was collected in four repetitions, with the noise condition and repetition number randomized, and with the same spoken token and a different realization of the noise used in each session. After being trained with feedback, humans performed the DRT task without feedback using Sennheiser HD580 earphones in a sound booth with double-walls made by the Industrial Acoustics Corp.

Human performance is evaluated based on percent correct responses using Voiers' DRT paradigm, and scores are broken down according to the DRT acoustic–phonetic dimensions of voicing, nasality, sustension, sibilation, graveness, and compactness. Knowledge about the acoustic correlates of the acoustic–phonetic dimensions provides diagnostic information about temporal representation of speech (for example the longer in duration continuant vs. the abrupt obstruent consonant of the sustention dimension), while the vowel quadrant identity provides information about the frequency range (i.e. location of the formants in action). Hence, human error patterns can provide a fair amount of information about the nature and patterns of the phonetic confusions.

Human performance (measured by the number of errors divided by total number of presentations) over all 9 SPL and SNR conditions with synthetic speech is shown in Fig. 9 (with a summary of results displayed in Table 1). Here, the abscissa marks the six acoustic dimensions: voicing, nasality, sustention, sibilation, graveness and compactness (denoted VC, NS, ST, SB, GV and CM, respectively). The "+" sign stands for attribute present (such as the voicing present in the/d/ in daunt) and the "-" sign for attribute absent (such as in the initial unvoiced/t/ in taunt). As one can observe in Fig. 9, overall, as SNR decreases, human performance decreases. For synthetic speech, average human errors were within 2% of each other as SPL varied and SNR was held constant (i.e. the rows of Fig. 9). Sustension was the dimension with the most errors in Fig. 9. This means that humans had the hardest time distinguishing words with sustension on the initial diphone from those without it.

The human studies produced DRT results that were sorted according to acoustic dimension, and thus provided very detailed error patterns for human listeners. The error patterns for the synthetically generated corpora were very similar to those of naturally spoken speech (Ghitza et al., 2007) and produced error rates that were stable over different noise SPL levels with sustention contributing more to errors than any other acoustic dimension. The exception to the above trend was the nasal sounds, which sounded slightly metallic and had far fewer errors (approximately zero errors) than similar tasks with humans on naturally spoken speech. We believe this metallic quality provides an unnatural cue that is exploited by the central auditory system. Hence we omitted the nasality dimension from the database. The error patterns along the other dimensions were used as targets for the machine DRT mimic described in the next section.

4.2. Machine tests: matching human performance

To find the parameter values of the closed-loop MBPNL model, the amount of noise allowed per frequency band was adjusted iteratively by tuning the DRW bounds and other parameters discussed in Section 3 to find the model settings that produced the best machine match to human scores along each acoustic–phonetic dimension. Results were computed using a single-template – one of the nine SPL (50, 60, and 70 dB) and SNR (0, 5, and 10 dB) conditions – spectrogram-like time–frequency representations of the output of the auditory filters (each filter output is a horizontal slice of the display, as shown in Figs. 5 and 8). For a given test token, a mean-squared-error (MSE) L_2-norm distance was computed between the test token and the two possible template tokens. This MSE was computed according to the following formula:

$$MSE_a(x) = \frac{\sum_{n=1}^{N_n}\sum_{i=1}^{N_i}[y_x(n,i) - y_a(n,i)]^2}{N_i N_n} \tag{1a}$$

$$MSE_b(x) = \frac{\sum_{n=1}^{N_n}\sum_{i=1}^{N_i}[y_x(n,i) - y_b(n,i)]^2}{N_i N_n} \tag{1b}$$

Here n is the index of the time frame, i is the index of the cochlear channels, N_i is the total number of frequency indices. N_n is the total number of time frames. $y_a(n,i)$ is the output of the front-end when the first template token is input to the system, $y_b(n,i)$ is the ouput of the front-end when the second template token is the input, and $y_x(n,i)$ is the ouput of the front-end when the test token is the input. For a given test token, the template producing the smaller MSE distance was selected as the simulated DRT response. For example if $MSE_a(x) > MSE_b(x)$, then the template diphone "b" was selected. Otherwise, template diphone "a" was selected. The resulting scores for each DRT presentation were then compared to find how well the system matched human performance.

For initial studies, the amount of noise power allowed over the lower bound of the DRW was incremented in 1 dB steps, with a tolerance of 0.1% difference in noise power. For later studies the amount of noise allowed over the lower bound of the DRW was set to 2 dB, 6 dB, or 10 dB, with different combinations of level per frequency band. The frequency bands examined were divided roughly according to the first formant, second formant, and third formant regions for clean speech. Specifically, the first frequency band had channels with center frequency of 266–844 Hz; the second frequency band had channels with center frequency of 875–2359 Hz; and the final frequency band

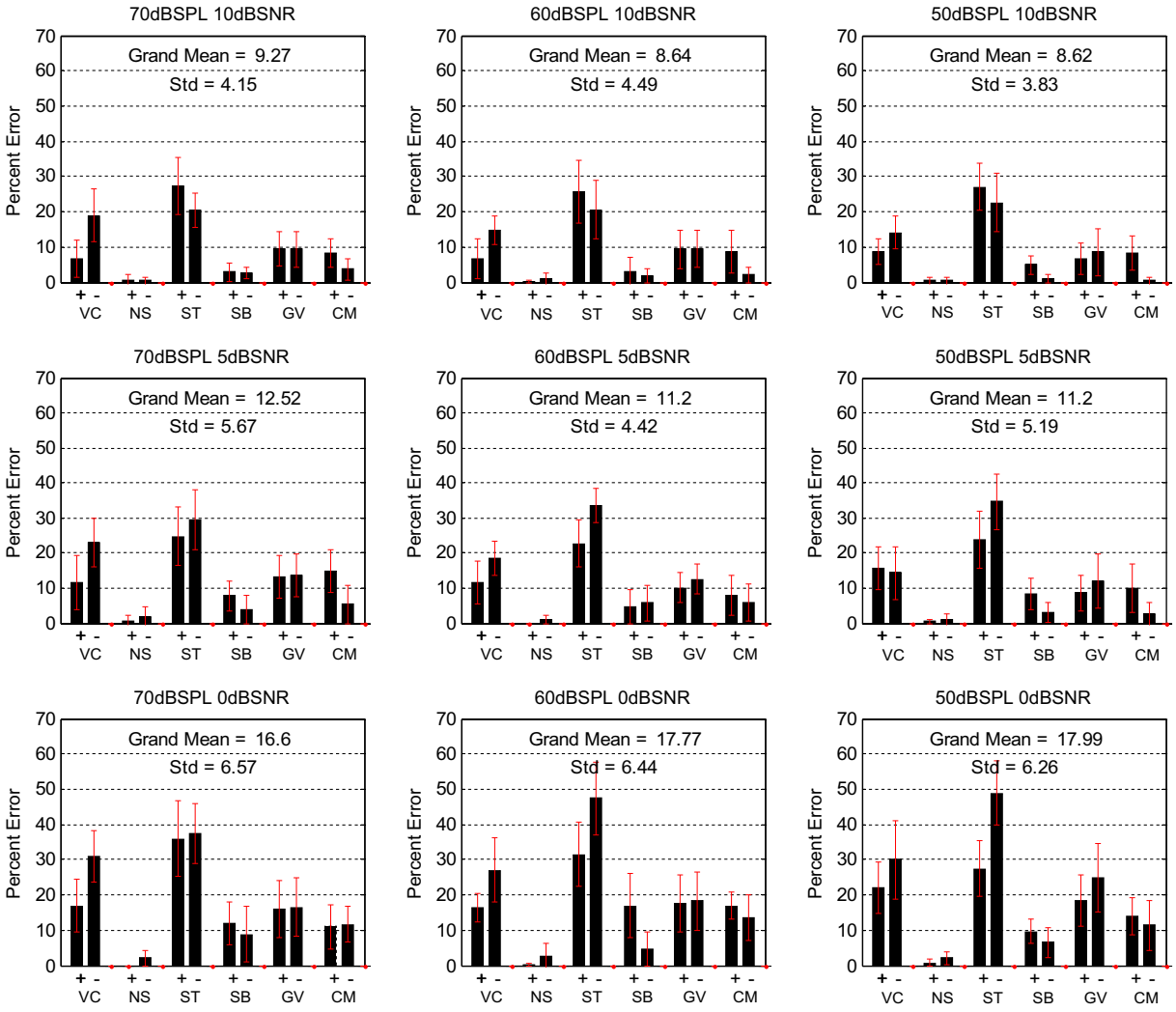

Fig. 9. Human performance on Voiers' 2AFC DRT using synthetic speech created by the HLsyn speech synthesis system. Performance is broken down into DRT dimensions having the attributes of voicing (VC), nasality (NS), sustension (ST), sibilation (SB), graveness (GV), and compactness (CM). + Indiciates diphones that have the attribute and − indicate diphones that do not have the attribute. The grand mean is computed by averaging the percent correct over all dimensions and +/− attributes. As SNR decreases, human performance decreases. Human errors moderately decrease as SPL is decreased for all conditions but the 0dB SNR cases.

Table 1
Grand mean errors per noise condition for synthetic speech.

	70 dB SPL	60 dB SPL	50 dB SPL
10 dB SNR	9.3	8.6	8.6
5 dB SNR	12.5	11.2	11.2
0 dB SNR	16.6	17.8	18.0

examined had channels with center frequency of 2422–5141 Hz.

A Chi-squared metric with a significance level of 95% based on contingency table analysis of data (Zar, 1999) was used to evaluate how closely machine performance matched that of humans, and to tune the front-end auditory model parameters. The settings that yielded the best match to human in the Chi-squared sense were a DRW lower bound of 65 dB, with noise allowed per frequency band according to Table 2, with stretching, and with a 10-ms window.

Chi-squared analysis for the DRT mimic task, with the closed-loop MBPNL model and template tokens at 60 dB SPL × 10 dB SNR, are shown in Figs. 10 and 11a. Fig. 10a shows performance averaged over SPL and SNR conditions and Fig. 11 shows a breakdown per condition (Fig. 10b – the averaged performance with an open-loop MBPNL model – was added for comparison purposes). The results suggest that the acoustic dimensions of voicing minus and sustention minus were significantly different from human for the majority of the conditions tested. When examining Fig. 10, the negative bars for the voicing minus and sustention minus categories imply that the machine is performing better than humans. The reason for this better machine performance is unknown; however it could be due to the simple L_2-norm MSE computation between time and frequency spectrogram-like token representations of our back-end. For the voicing category, timing differences between voiced and unvoiced sounds due to

Table 2
Noise allowed above the lower bound of the DRW per frequency band for the system with the best match to human.

Frequency band CF	Noise above DRW lower bound
266–844 Hz	10 dB
875–2359 Hz	6 dB
2422–5141 Hz	6 dB

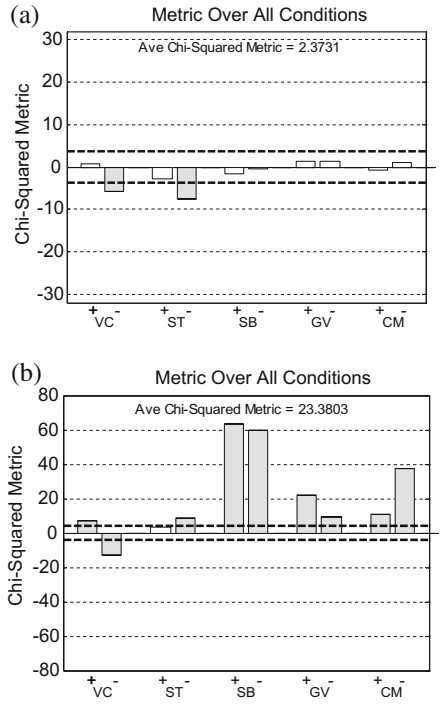

Fig. 10. Overall Chi-squared results for the system that yielded the best match to humans. Negative bars indicate human errors exceed that of machine. Positive bars indicate machine errors exceed that of humans. The absolute value of each bar is the Chi-squared value for that acoustic dimension. (a) Closed-loop MBPNL model. The performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall Chi-squared metric. (b) Open-loop MBPNL model without efferent control, added for comparison with 10(a) to show the improvement gained by adding efferent feedback. The DRW is tuned in the same manner as the closed-loop MBPNL model, chosen to match human performance as best as possible. Overall Chi-squared metric and individual values indicate a large difference with human.

voiced onset times could make discrimination easier for the machine model and hence bias results. For the sustension category, continuants (such as/f/) which belong to the ST+ category tend to occur in initial consonants that are much more gradual and spread over time while obstruents (such as/p/) which belong to the ST-category are much more abrupt and compact over time. It is possible that these timing differences are over-emphasized by the nature of our simple MSE back-end comparison on time-aligned speech, hence biasing performance in favor of the machine for these two categories.

All other DRT acoustic categories have cues that are less dependent on timing differences. Machine performance over these categories also matched humans much better with a few exceptions. The graveness plus category significantly differs for the 60 dB SPL × 5 dB SNR condition, and the graveness minus category significantly differs for the 50 dB SPL × 10 dB SNR condition.

Despite the differences for a few acoustic categories and for a few presentation conditions, the average Chi-squared metric of 2.37 suggests that on average, machine performance was close to human (and certainly within the Chi-squared significance level of 3.84).

Results with all nine noise conditions used as the template condition is displayed in Table 3. As these tables show, the 60 dB SPL × 10 dB SNR template condition produced the best Chi-squared metric results. However all nine template choice results did not vary largely, reflecting the stability of the closed-loop MBPNL representation.

5. Discussion

The main goal of this work was to describe a potential model of the signal processing of the human auditory periphery and demonstrate how several of the modeled non-linear operations of the system can be used to develop a machine that improves our capability to predict human performance in additive white noise. One of the key non-linear interactions of our system that is regulated by efferent-inspired feedback control is that of the MPBNL gain versus the lower bound of the DRW. Besides affecting filter shapes in response of noise, this interaction aids in making the output more consistent. In part, this is due to the normalizing effect that efferent control has on the output: it makes outputs fall into the DRW of interest and be consistent across input levels. However it also yields a processing advantage across SNR levels that traditional linear processing does not provide: at low noise levels, the efferent GAIN parameter is high, making the filters more responsive to small amplitude signals while the high amplitude signal response is kept roughly the same (as described in Fig. 2). This in turn amplifies small amplitude sounds such as some transients in consonants, which may be very useful for speech recognition in environments with low levels of noise. At high noise levels, the gain is low, making the filters much less responsive to small amplitude signals. Hence smaller short-time noise transients are attenuated and effectively masked below our DRW rate window of interest. At these higher noise levels, this noise masking effect allows the higher SNR regions of the speech signal to emerge from the noise background. This effectively yields an "unmasking" of sounds in noisy backgrounds, similar to the affect Ferry and Meddis (2007) describe in their work and that of (Dolan and Nuttal, 1988; Kawase et al., 1993). This overall efferent effect yields improved performance that matches humans better than a linear system, such as a normalized-input gammatone filter system (see Appendix).

By focusing on synthetic speech we were able to make time-aligned diphone comparisons, which greatly

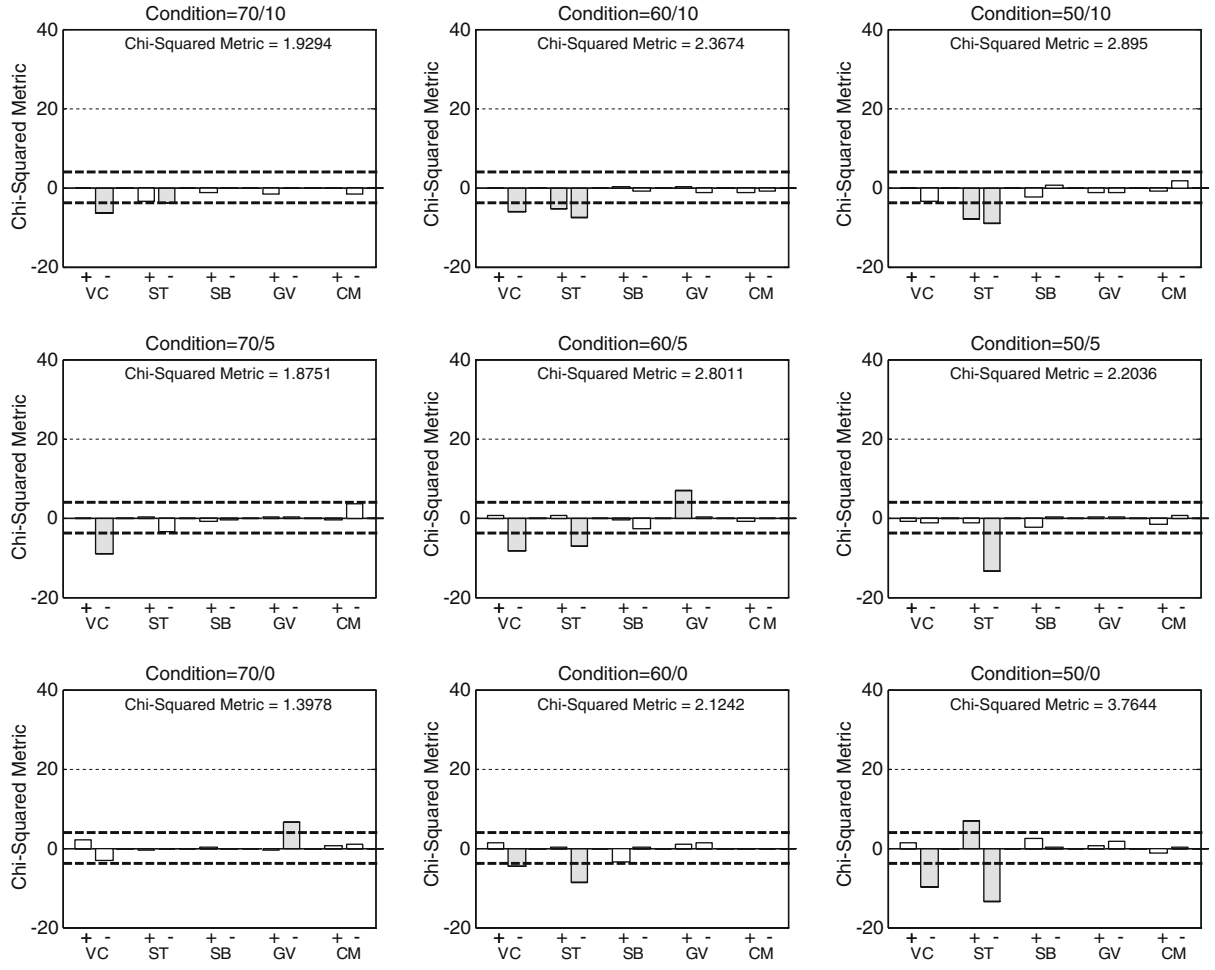

Fig. 11. Detailed Chi-squared metric results computed separately for each noise condition for the system that yielded the best match to humans. The noise condition is specificed in each panel by the SPL/SNR levels. The machine performance on a few acoustic dimensions, especially voicing-minus and sustension-minus, is significantly better than human performance. Overall the Chi-squared metrics here indicate that this system was a much better match than any other we had evaluated.

Table 3
Optimal Chi-squared metric values as a function of template condition with smoothing window length set to 10 ms. The 60 dB SPL × 10 dB SNR condition yields the best Chi-squared metric value.

	70 dB SPL	60 dB SPL	50 dB SPL
10 dB SNR	3.9069	2.3731	2.7834
5 dB SNR	2.8635	3.2876	2.9221
0 dB SNR	7.5620	7.5822	8.1114

simplified the back-end classifier. However by using synthetic speech several imperfections may have been introduced such as the metallic sound of the nasals. Furthermore, it is possible that temporal features were exaggerated, making machine discrimination easier than it should be. For example, the time-aligned nature of the speech might have made it easier for the MSE machine computation in the back-end to distinguish the duration of the initial consonant in the sustention category or the voice onset time for the voiced category (both of which are cues for those acoustic dimensions), hence resulting in better machine results. Despite this, overall machine results matched humans in a Chi-squared test and the main goals of this work were accomplished.

An interesting avenue of future work could involve exploring the use of the closed-loop MBPNL system for automatic speech recognition (ASR). Due to the nature of the noise-responsive feedback, the closed-loop system produces spectrograms that fluctuate less with changes in noise intensity and SNR compared to spectrograms produced by the open-loop system. This property is desirable for stabilizing the performance of our template-matching operation (or any other statistical pattern recognition method, e.g. HMM) under varying noise conditions. Another point is noteworthy: unlike the case of ASR, where we aim at minimum error rate, here we aimed at matching human error patterns. Towards this end, we had to inject a certain amount of noise into the DRW (see Section 4.2). Reducing the noise intensity in the DRW to zero should improve recognition results.

6. Summary

In this paper we have revised current models of auditory periphery by including the role of the descending pathway in making the cochlear response to speech sounds robust to degradation in acoustic conditions. We have qualitatively demonstrated the system's ability to produce spectrograms of noisy speech samples that are more consistent with spectrograms of speech in quiet than are spectrograms produced by open-loop models of the auditory periphery. We have also evaluated the ability of this model to preserve phonetic information quantitatively, by using it as a front-end in a machine designed to mimic human confusion patterns to initial consonants in noise. Potential applications include (1) enabling a diagnostic assessment of speech intelligibility by using the efferent-inspired model of the auditory periphery integrated with perception-based template matching, and (2) improving the performance of automatic speech recognition systems in acoustically adverse conditions.

Acknowledgement

This work was sponsored by the US Air Force Office of Scientific Research (contracts F49620-03-C-0051 and FA9550-05-C-0032) and by NIH Grant R01-DC7152.

Appendix A

The purpose of this Appendix is to highlight differences between the closed-loop MBPNL model and the standard, open-loop Gammatone model (Patterson et al., 1995). In implementing the open-loop Gammatone model we used a bank of overlapping filters (developed by Slaney, 1993) uniformly distributed along the ERB scale, four channels per ERB (same distribution as in the MBPNL model, see Section 3.2). Each filter was followed by a generic model of the Inner Hair Cell – half-wave rectification followed by low-pass filtering, representing the reduction of synchrony with CF.

The closed-loop MBPNL model and the open-loop Gammatone model are different in two major ways, in the overall architecture – open-loop vs. closed-loop, and in the type of filterbank – linear vs. nonlinear. In Section A.1 we examine the role of the filterbank by comparing the Gammatone (linear) with the MBPNL (nonlinear) filterbanks, both in an open-loop configuration. In Section A.2 we comment on the advantage in using a closed-loop configuration. Our comparisons of the various models were conducted by inspecting spectrograms (visually) and by examining numerical chi-squared results, in the same manner as in the body of this paper.

A.1. Gammatone (linear) vs. MBPNL (nonlinear) filterbanks in open-loop configuration

Here we compare the Gammatone and the MBPNL filterbanks, both without amplitude normalization at the input. Figs. 12 and 5 (a) show spectrograms of the Gammatone filterbank output and the open-loop MBPNL filterbank output, respectively, in different SPL and SNR conditions. The spectrographic display exhibits inconsistency across SPL × SNR conditions, for both models, reflecting the wide signal intensity dynamic range. Note that this inconsistency is somewhat reduced with the MBPNL due to the nonlinear nature of the MBPNL. (As shown in Fig. 2, upper left plot, for a dynamic range of 100dB at the input, the dynamic range at the output of the MBPNL is reduced by 37dB.)

This informal, visual, observation is quantified by measuring the machine performance on the DRT mimic with a Gammatone- vs. open-loop MBPNL- based machines, shown in Figs. 13 and 10 a,b, respectively. For both cases, templates were chosen to optimize the Chi-squared metric, 70dB SPL × 0 dB SNR for the Gammatone, 60dB SPL × 10 dB SNR for the open-loop MBPNL. For the Gammatone model (Fig. 13a), all chi-square bars are positive, indicating machine errors exceed that of humans, and all bars are grey, indicating that the difference between machine and human errors are greater than the significance threshold of 3.841. A similar trend is observed for the open-loop MBPNL model (Fig. 10b): all chi-square bars are positive except the non-voiced (VC-) condition, and all Chi-square bars are grey except the sustained (ST+) condition. In agreement, the average chi-squared metric is 23.3 for the open-loop MBPNL and 34.7 for the Gammatone. Although the open-loop MBPNL matches human performance better than the Gammatone system, neither model matches human performance well.

A.2. Open-loop Gammatone vs. closed-loop MBPNL

As suggested in the body of the paper, the closed-loop configuration results in a greater consistency at the output display across SPL × SNR conditions (Fig. 8). Fig. 12b shows the spectrographic display of the open-loop Gammatone model **with** amplitude normalization at the input. Normalization at the input improves consistency at the output of the system: output displays vary slightly across SNR and are very stable across SPL.

This improvement is reflected in the machine performance on the DRT mimic, shown in Figs. 10 and 13a and b. For both cases, templates were chosen to optimize the Chi-squared metric, 70 dB SPL × 5 dB SNR for the input-normalized Gammatone, 60 dB SPL × 10 dB SNR for the closed-loop MBPNL. For the input-normalized Gammatone system, sibilation-minus (ST-)shows the largest human–machine mismatch. Sustension and compactness-plus (CM+) have Chi-squared values indicating a similar human–machine performance. For the input-normalized Gammatone, the average Chi-squared metric per acoustic dimension is 6.5884 – a worse match to human performance compared to the closed-loop MBPNL performance (Fig. 10a).

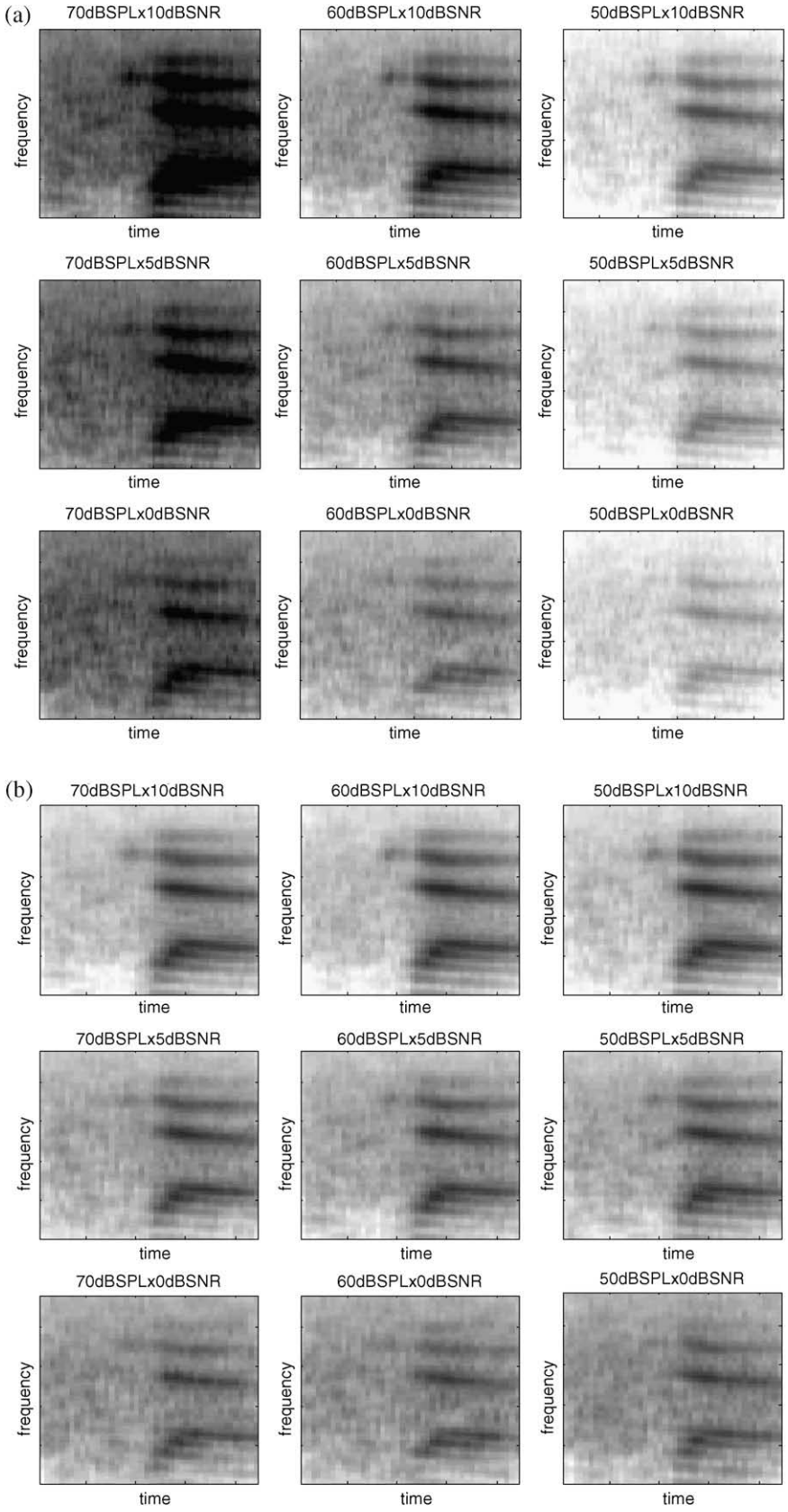

Fig. 12. Simulated IHC response to diphone/ja/, (a) open-loop Gammatone model using an 10-ms smoothing window without amplitude normalization at the input. Like the MBPNL systems, a total of 96 filters were used. A large inconsistency in the simulated IHC response spectrum is observed across varying noise intensity and SNR levels and (b) open-loop Gammatone model using an 8-ms smoothing window with amplitude normalization at the input. Compared to 12a, much better consistency is observed.

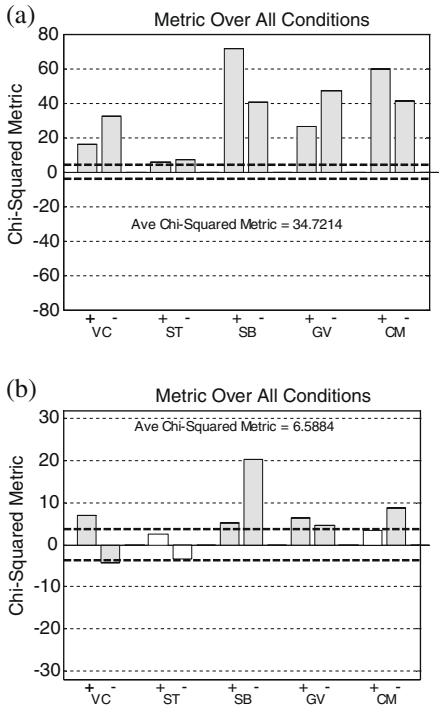

Fig. 13. Overall Chi-squared results for a linear Gammatone system that yielded the best match to humans (a) without input normalization: Overall, for the Gammatone model, machine errors significantly exceed human ones in all acoustic dimensions and (b) with normalized input: Machine errors exceed human ones in all acoustic dimensions but voicing-minus and sustension-minus. Only the sustention (ST) and compactness-plus (CM+) dimensions produce average results that do not exceed the significance threshold. The final normalized-gammatone results provide a much worse match to human performance than the MBPNL results of Fig. 10.

References

Ainsworth, W., Cervera, T., 2001. Effects of noise adaptation on the perception of voiced plosives in isolated syllables. EUROSPEECH-2001, 371–374.

Ainsworth, W., Meyer, G., 1994. Recognition of plosive syllables in noise: comparison of an auditory model with human performance. J. Acoust. Soc. Amer. 96 (2), 687–694.

ANSI, 1969. ANSI S3.5-1969, American National Standard: Methods for the Calculation of the Articulation Index. American National Standards Institute, New York.

ANSI, 1997. ANSI S3.6-1997, American National Standard: Methods for the Calculation of the Articulation Index. American National Standards Institute, New York.

Cervera, T., Gonzalez-Alvarez, J., 2007. Temporal effects of preceding band-pass and band-stop noise on the recognition of voiced stops. Acta Acust. United Acust. 93 (6), 1036–1045, 10.

Dewson, J., 1968. Efferent olivocochlear bundle: some relationships to stimulus discrimination in noise. J. Neurophysiol. 31, 122–130.

Dolan, D.F., Nuttal, A.L., 1988. Masked cochlear whole-nerve response intensity functions altered by electrical stimulation of the crossed olivocochlear bundle. J. Acoust. Soc. Amer. 83, 1081–1086.

Dunn, H.K., White, S.D., 1940. Statistical measurements on conversational speech. JASA 11 (January), 278–288.

Ferry, R., Meddis, R., 2007. A computer model of medial efferent suppression in the mammalian auditory system. J. Acoust. Soc. Amer. 122 (6), 3519–3526.

French, N.R., Steinberg, J.C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Amer. 19, 90–119.

Ghitza, O., 2004. On the possible role of MOC efferents in speech reception in noise. J. Acoust. Soc. Amer. 115 (5), Pt.2.

Ghitza, O., Messing, D., Delhorne, L., Braida, L., Bruckert, E., Sondhi, M.M., 2007. Towards predicting consonant confusions of degraded speech. In: Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S., Verhey, J. (Eds.), Hearing – from Sensory Processing to Perception. Springer-Verlag, Berlin, Heidelberg.

Gifford, M.L., Guinan, J.J., 1983. Effects of crossed olivocochlear bundle stimulation on cat auditory-nerve responses to tones. J. Acoust. Soc. Amer. 74, 115–123.

Giraud, A.L., Garnier, S., Micheyl, C., Lina, G., Chays, A., Chery-Croze, S., 1997. Auditory efferents involved in speech-in-noise intelligibility. Neuroreport 8 (7), 1779–1783.

Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Res. 47, 103–108.

Goldstein, J.L., 1990. Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering. Hearing Res. 49, 39–60.

Guinan, J.J., 1996. Physiology of olivocochlear efferents. In: Dallos, P., Popper, A.N., Fay, R.R. (Eds.), The Cochlea. Springer, New-York, pp. 435–502.

Gummer, M., Yates, G.K., Johnstone, B.M., 1988. Modulation transfer function of efferent neurones in the guinea pig cochlea. Hearing Res. 36 (1), 41–51.

Hant, J.J., Alwan, A., 2003. A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise. Speech Comm. 40, 291–313.

Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function. Acustica 46, 60–72.

Johnson, D.H., 1980. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. J. Acoust. Soc. Amer. 68 (4), 1115–1122.

Kawase, T., Liberman, M.C., 1993. Antimasking effects of the olivocochlear reflex. I. Enhancement of compound action potentials to masked tones. J. Neurophysiol. 70 (6), 2519–2532.

Kawase, T., Delgutte, B., Liberman, M.C., 1993. Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. J. Neurophysiol. 70 (6), 2519–2532.

Kiang, N.Y.S., Guinan, J.J., Liberman, M.C., Brown, M.C., Eddington, D.K., 1987. Feedback control mechanisms of the auditory periphery: implication for cochlear implants. In: Banfai, P. (Ed.), Internat. Cochlear Implant Symposium, Duren, West Germany.

Liberman, M.C., 1988. Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. J. Neurophysiol. 60, 1779–1798.

Liberman, M.C., Brown, M.C., 1986. Physiology and anatomy of single olivocochlear neurons in the cat. Hearing Res. 24, 17–36.

Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Comm. 22 (1), 1–15.

May, B.J., Sachs, M.B., 1992. Dynamic range of neural rate responses in the ventral cochlear nuecleus of awake cats. J. Neurophysiol. 68, 1589–1602.

Patterson, R.D., Allerhand, M.H., Giguere, C., 1995. Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. J. Acoust. Soc. Amer. 98, 1890–1894.

Scharf, B., Magnan, J., Chays, A., 1997. On the role of the olivocochlear bundle in hearing: 16 case studies. Hearing Res. 103, 101–122.

Slaney, M., 1993. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Apple Computer Technical Report #35.

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Comm. 45, 401–423.

Voiers, W.D., 1983. Evaluating processed speech using the diagnostic rhyme test. Speech Technol. 1 (4), 30–39.

Warr, W.B., 1978. The olivocochlear bundle: its origins and terminations in the cat. In: Naunton, R., Fernandez, C. (Eds.), Evoked Electrical Activity in the Auditory Nervous System. Academic Press, New York, pp. 43–63.

Winslow, R.L., Sachs, M.B., 1988. Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. Hearing Res. 35, 165–190.

Zar, J.H., 1999. Biostatistical Analysis, fourth ed. Prentice Hall, Upper Saddle River, NJ.

Zeng, F., Martino, K.M., Linthicum, F.H., Soli, S.D., 2000. Auditory perception in vestibular neurectomy subjects. Hearing Res. 142, 102–112.