Consonant Discrimination of Degraded Speech using an Efferent-inspired Closed-loop Cochlear Model

David P. Messing[1], *Lorraine Delhorne*[1], *Ed Bruckert*[2], *Louis D. Braida*[1], *and Oded Ghitza*[2]

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2] Sensimetrics Corporation, Somerville, Massachusetts, USA

dmessing@mit.edu, ldbraida@mit.edu, oded@sens.com

10.21437/Interspeech.2008-326

Abstract

We present a model of auditory speech processing capable of predicting consonant confusions by normal hearing listeners, based on a phenomenological model of the Medial Olivocochlear efferent pathway. We then use this model to predict human error patterns of initial consonants in consonant-vowel-consonant words. In the process we demonstrate its potential for speech identification in noise. Our results produced performance that was robust to varying levels of additive noise and which was similar to human performance in discrimination of synthetically spoken consonants.

Index Terms: speech recognition, speech processing, auditory models, olivocochlear efferent feedback.

1. Introduction

Medial olivocochlear (MOC) efferent activity is believed to regulate the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background (e.g., [1]). By reducing outer hair cell (OHC) motility and changing OHC shape, MOC stimulation increases basilar membrane stiffness, and in turn inhibits inner hair cell (IHC) response in the presence of noise. This paper develops a closed-loop model of the peripheral auditory system, a front-end model that adaptively adjusts its cochlear operating point. Specifically we develop our model to predict human confusions of initial consonants in speech-shaped additive Gaussian noise.

Our long-term goal is to formulate a template-matching operation (back-end), with perception-related rules of integration over time and frequency at its core, in the context of human perception of degraded speech, but in this paper we concentrate on separating the back-end development from the front-end. Our approach is to minimize the influence of cognitive and memory factors while preserving the complex acoustic cues that differentiate initial diphones. Hence we tune the parameters of the peripheral auditory model in as much isolation as possible by reducing the effect of the back-end system. Once the basic signal processing front-end of our model is tuned, we can then freeze the front-end and develop the back-end pattern recognition template matching system.

2. Peripheral Auditory Model (PAM)

We have developed a closed-loop model of the auditory periphery that was inspired by current evidence about the possible role of the efferent system in regulating the operating point of the cochlea. This, in turn, results in an auditory nerve (AN) representation that is less sensitive to changes in environmental conditions. In implementing the cochlear model we use a bank of overlapping cochlear channels uniformly distributed along the ERB scale [4], four channels per ERB. Each cochlear channel comprises a nonlinear filter and a model of the IHC (half-wave rect-ification followed by a low-pass filter, representing the reduction of synchrony with frequency). The dynamic range of the simulated IHC response is restricted – both below and above – to a *dynamic range window* (DRW), representing the observed dynamic range at the AN level.

The filter is derived from Goldstein's [2] model of nonlinear cochlear mechanics (MBPNL, figure 1). This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior. The lower path (H1/H2) is a compressive nonlinear filter that represents the sensitive, narrowband 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 relatively 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 (i.e. without adaptive feedback) MBPNL model GAIN is set to 40dB, to best mimic psychophysical tuning curves in quiet.

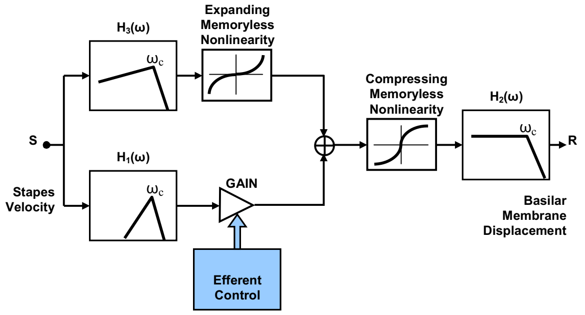

Figure 1. *MBPNL filterbank. A parameter GAIN controls the gain of the tip of the basilar membrane tuning curves. To best mimic psychophysical tuning curves of a healthy cochlea in quiet, the tip gain is set to GAIN =40dB [2]*

As for the efferent-inspired part of the model we mimic the effect of the Medial Olivocochlear efferent path (MOC). Morphologically, MOC neurons project to different places along the cochlea partition in a tonotopical manner, making synapse connections to the outer hair cells and, hence, affecting the mechanical properties of the cochlea (e.g. increasing the basilar membrane stiffness) [3]. Therefore, we introduce a frequency dependent feedback mechanism which controls the tip-gain (G) of each MBPNL channel according to the intensity level of the sustained noise in that frequency channel.

The "iso-input" frequency response of an MBPNL filter at *CF* of 3641Hz with various tip gain settings is shown in Figure 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. The frequency response for the open-loop MBPNL model is shown at the upper-left corner (i.e. for GAIN = 40dB). Figure 2 shows the iso-input frequency response of the system for different values of input 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 ratio of the maximum to minimum peaks, corresponding to inputs of 40 dBSPL and 120 dBSPL in Figure 2, increases. In our closed-loop model (i.e. with feedback), the tip GAIN parameter is adjusted based on the efferent response, which in turn is calculated based on the amount of noise present.

Figure 3 shows one cochlear channel of the efferent-inspiried closed loop model. We introduce a frequency 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 figure 3 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.

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 tuning 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 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 Figure 2. In Figure 2, the upper-left panel represents the nominal response (i.e. in quiet), with GAIN set to 40dB. In this quiet condition, weaker amplitude sounds such as the 40dBSPL sound are amplified greatly (in this case roughly 20 dB more) relative to louder sounds such as the 120 dBSPL stimulus. By increasing the efferent response in noise, we reduce the GAIN and the MBPNL response to weaker stimuli such as the 40 dBSPL tone (and background noise), as shown in the lower right pane of figure 2 where the GAIN parameter is set to 10dB. Hence for high energy tone stimuli the MBPNL response is hardly affected, while the response for low energy stimuli (e.g. 40 or 60dBSPL signals) is reduced by some 30dB in the presence of noise.

Figure 4 shows – in terms of a spectrogram – simulated IHC responses to speech in noise for two conditions (SPL levels of 50dB or 70dB, and SNR of 10dB), for an open-loop system with a linear cochlear model (left-hand side) and for the closed-loop system (right-hand side). Due to the nature of the noise-responsive feedback, the closed-loop system spectrograms fluctuate with changes in background noiseconsiderably less than are spectrosgrams produced by the open-loop model. This property is desirable for stabilizing the performance of the template-matching operation with varying noise conditions, as reflected in the quantitative evaluation reported next.

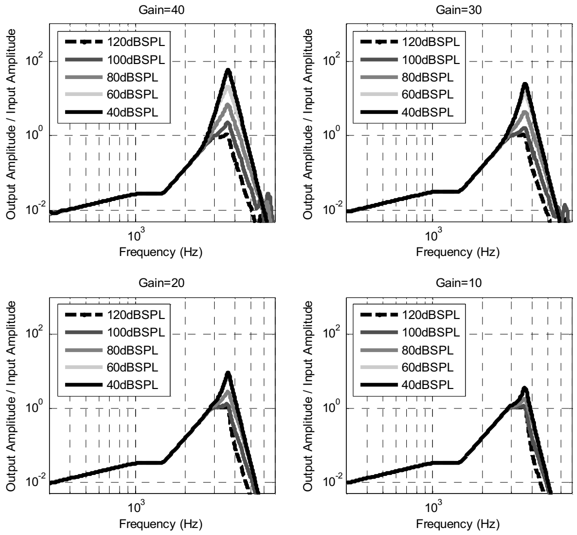

Figure 2. *Iso-input frequency responses of an MBPNL filter (at CF of 3641Hz) for different values of GAIN parameter. From Upper-left, clockwise: GAIN =40, 30, 20 and 10dB. Upper-left corner (Gain=40dB) is for healthy cochlea in quiet (Goldstein, 1990). Input sinusoids are varied from 40dBSPL to 120dBSPL*

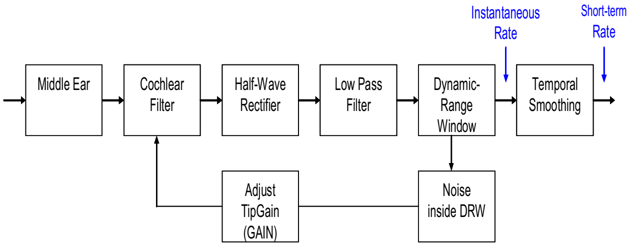

Figure 3. *Overview of one channel of the front-end model with efferent feedback. An efferent feedback response is calculated based on the noise inside the DRW window for that particular channel. The tip-gain (GAIN) of each MBPNL filter is adjusted until the noise inside the DRW window of that channel reaches a desired level (this amount of noise was a parameter of our model and was adjusted).*

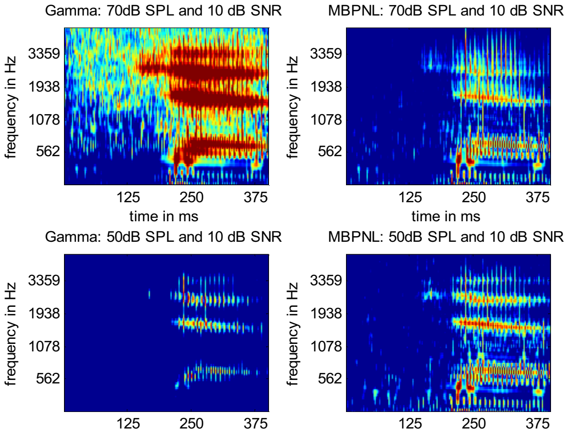

Figure 4. *Simulated IHC response for open-loop, linear PAM (left) and for closed-loop PAM (right.)*

3. Quantitative evaluation

Our long-term objective is to predict consonant confusions for degraded speech made by normally-hearing listeners. Our predictions are based on the efferent-inspired peripheral auditory model followed by a template matching operation. To make these predictions we must find the parameters of the front-end with a minimal interference of the back-end. As a back-end we used a Euclidian distance measure between the template and test tokens; this choice is justifiable because of the following relaxation steps.

To eliminate unwanted interaction between stages, errors due to template matching should be reduced to *zero*. We attempted 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' DRT [5] which presents the subject with a two alternative forced choice between two alternative CVC words that differ in their initial consonants. Such task reduces the influence of cognitive and memory factors while maintaining the complex acoustic cues that differentiate initial diphones (recall the central role of diphones in speech perception, e.g. [6]); (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. [7]): the same acoustic speech token is used for training and for testing, so that testing tokens differ from training tokens only by the acoustic distortion.

For these studies the amount of noise allowed over the lower bound of the DRW was set to 2dB, 6dB, or 10dB, 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 Hz to 844Hz; the second frequency band had channels with center frequency of 875 Hz to 2359 Hz; and the final frequency band examined had channels with center frequency of 2422Hz to 5141Hz.

A Chi-squared metric with a significance level of 95% based on contingency table analysis of data [8] 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 65dB, with noise allowed per frequency band according to table 1, with stretching, and with a 10-ms window.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	10 dB
875-2359 Hz	6 dB
2422-5141 Hz	6 dB

Table 1. *Noise allowed above the lower bound of the DRW per frequency band for the system with the best match to human.*

Cumulative Chi-squared analysis per DRT dimension using a template token at 60dBSPL and 10dBSNR are shown in figures 5 and 6; for these tests the noise level and SNR or the test token was varied between SPL=70, 60, and 50dB and SNR=0, 5, and 10dB. These 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 figure 5, 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 pixel by pixel MSE computation of our backend. For the voicing category, timing differences between voiced and unvoiced sounds due to 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 backend 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 60dBSPL x 5dBSNR condition, and the graveness minus category significantly differs for the 50dBSPL x 10dBSNR 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).

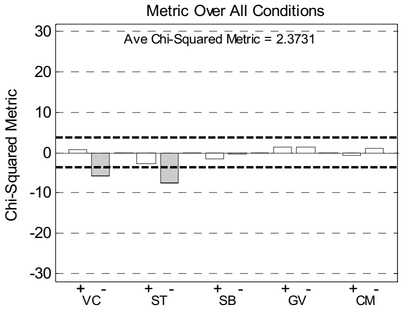

Figure 5. *Overall Chi-squared results for the system that yielded the best match to humans. Labels correspond to the following acoustic-phonetic features: VC=voicing, ST=sustension, SB=sibilation, GV=graveness, and CM=compactness. + indicates that the acoustic dimension is present; - indicates that the acoustic dimension is absent. 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. 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. Grey bars indicate differences between machine and human that were statistically significant according to the Chi-squared test.*

4. Discussion

We described a model of the signal processing of the human auditory periphery and demonstrated how several of the non-linear operations taking place in the biological system can be used to develop a system that improves our capability to predict human performance in 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 with respect to SPL level. 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 performance gain across SNR levels that traditional linear processing does not provide. At low noise levels, the gain is high, making the filters more responsive to small amplitude signals. 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 effect Ferry [9] describes in his work and Dolan and Nuttal [10], and Kawase et al. [11].

5. Acknowledgements

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

6. References

[1] Kiang, N. Y. S., Guinan, J. J., Liberman, M. C., Brown, M. C., and Eddington, D. K. "Feedback control mechanisms of the auditory periphery: implication for cochlear implants." In Banfai, P., editor, *International Cochlear Implant Symposium*. Duren,West Germany, 1987.

[2] Goldstein, J. L. "Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering," *Hearing Research,* 49, 39-60, 1990.

[3] Guinan, J. J. "Physiology of Olivocochlear Efferents," In Dallos, P., Popper, A. N. and Fay, R. R., editors, *The Cochlea*, pages 435–502, Springer, New-York, 1996.

[4] Moore, B. C. J. and Glasberg, B. R. "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. Am.,* 74, 750-753, 1983.

[5] Voiers, W. D. "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, 1(4): 30–39, 1983.

[6] Ghitza, O. "Processing of spoken CVCs in the auditory periphery. I.," *J. Acoust. Soc. Am.,* 94(5): 2507- 2516, 1993.

[7] Hant, J.J., and Alwan, A. "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Communication*, 40, 291-313, 2003.

[8] Zar, J.H. Biostatistical Analysis. 4th Edition, Prentice Hall., Upper Saddle River, NJ, 1999.

[9] Ferry, R., and Meddis, R. "A computer model of medial efferent suppression in the mammalian auditory system." *J. Acoust. Soc. Am.* **122(6)**: 3519–3526, 2007.

[10] Dolan, D. F., and Nuttal, A. L. "Masked cochlear whole-nerve response intensity functions altered by electrical stimulation of the crossed olivocochlear bundle," *J. Acoust. Soc. Am.* **83**: 1081–1086, 1988.

[11] Kawase, T., Delgutte, B., and Liberman, M. C. "Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones." *Journal of Neurophysiology,* Vol 70, Issue 6 2519-2532, 1993.

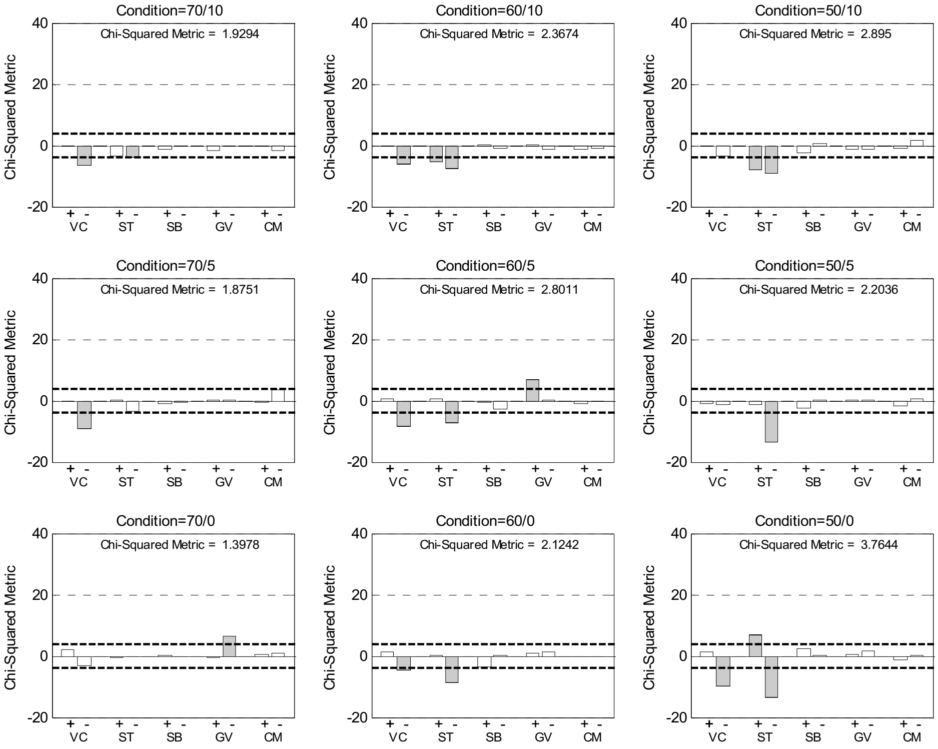

Figure 6. *Detailed Chi-squared metric results computed separately for each noise condition for the system that yielded the best match to humans. The noise condition is specificed in each panel by the SPL/SNR levels. The machine performance on a few acoustic dimensions, especially voicing-minus and sustension-minus, 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.*