USING AUDITORY FEEDBACK AND RHYTHMICITY FOR DIPHONE DISCRIMINATION OF DEGRADED SPEECH

Oded Ghitza

Sensimetrics Corporation
Somerville, MA 02144, USA
oded@sens.com

ABSTRACT

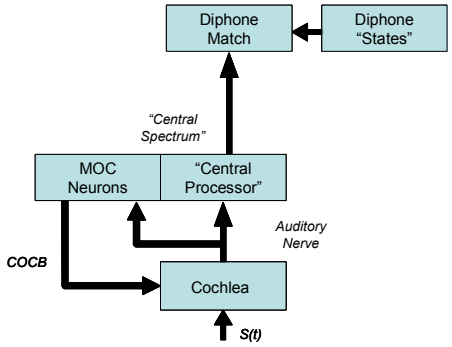
We describe a computational model of diphone perception based on salient properties of peripheral and central auditory processing. The model comprises an efferent-inspired closed-loop model of the auditory periphery connected to a template-matching neuronal circuit with a gamma rhythm at its core. We show that by exploiting auditory feedback a place/rate model of central processing is sufficient for the prediction of human performance in diphone discrimination of minimal pairs embedded in background noise – in contrast to the need for additional, temporal information when open-loop models of the periphery are used. We also demonstrate that the template-matching circuit exhibits properties, such as time-scaling insensitivity, consistent with (and desirable for) perception of spoken language.

1. INTRODUCTION

This paper examines signal processing principles used by the auditory system, in particular when the input signal is speech in the presence of background noise. A general observation is that with worsening environmental conditions, human performance in tasks related to speech intelligibility deteriorates gracefully compared to the performance of machines. This robust behavior may be attributed to either a unique form of signal processing in the auditory periphery or the use of context at higher layers (or a combination of both). What role the signal processing in the auditory periphery plays in achieving such performance? Are current models of the auditory periphery accurate enough to duplicate such performance?

Our scope is restricted to the processing that takes place prior to lexical access, on speech segments as long as 100 ms (i.e. as long as the duration of a monosyllable, e.g. diphone). At present, we have a reasonable understanding of the processing principles in the ascending pathway up through the auditory nerve (AN) − the cochlea, the inner hair cells (IHC), the outer hair cells (OHC) − and an increasing understanding of the brain-stem nuclei (such as the cochlear nucleus, the superior olivary complex, and the inferior colliculus). We have limited understanding of the neuronal circuitry by which speech is stored and retrieved in the auditory cortex, a limited understanding of the descending pathway, and little understanding of how the ascending and the descending pathways interact.

Figure 1: A block diagram of the prediction engine.



This study examines, by inference, the possible role of two mechanisms, auditory feedback and brain rhythmicity, in perceiving speech signals. We describe two computational models, an efferent-inspired closed-loop model of the auditory periphery and a template-match neuronal circuit (TMC) with an oscillatory drive at its core. From the properties of these models we infer the possible role of the underlying auditory mechanisms.

Figure 1 shows a block diagram of the model. It comprises an efferent-inspired peripheral auditory model (PAM) connected to a TMC. The extent to which this model is an accurate description of auditory perception is measured within the context of perceiving minimal word pairs (differing in their initial consonant) in the presence of additive, speech-shaped noise. In Section 2, we describe a closed-loop model of the auditory periphery that

comprises a nonlinear model of the cochlea with efferent-inspired feedback. The PAM parameters were determined in isolation from the TMC. This was achieved by analyzing confusion patterns generated in a paradigm with a minimal cognitive load (Voiers' Diagnostic Rhyme Test [DRT] [13], with *synthetic* speech stimuli to restrict phonemic variation). In Section 3, we describe initial steps towards predicting confusions of *naturally spoken* diphones (i.e. material that exhibits inherent phonemic variability). We describe a TMC inspired by principles of cortical neural processing (Hopfield, [10]). A desirable property of the circuit is insensitivity to time-scale variations of the input stimuli (associated with phonemic variability). We demonstrate the validity of this hypothesis in the context of the DRT diphone discrimination task.

2. PERIPHERAL AUDITORY MODEL

A reasonable, axiomatic assumption is that information in the auditory nerve is the only information available to the central nervous system (CNS) about *acoustic* input. While human performance in adverse conditions deteriorates only modestly, *simulated* AN representations of corrupted speech signals – generated by state-of-the-art auditory models – are markedly different from those associated with clean speech signals. For example, for speech in a typically reverberant room, there is only a slight deterioration of intelligibility (albeit with a noticeable degradation in quality) while the acoustic signature of the phonemic features in the simulated AN representations is severely compromised. Is this contrast a result of the incompleteness of current models of auditory processing?

Numerous papers have been published that examine how the response of the cochlea may be processed to provide a relevant representation of the speech signal. Each study utilizes a computational model to simulate either the direct firing activity or another related representation of the cochlear output. The manner in which this information is processed differs among the studies, reflecting differences in the structural properties of the central processor hypothesized by each study. These structural properties can be cataloged using the following three categories: (1) *place/rate* category, where the central processor possesses explicit knowledge of place (i.e. the fibers' tonotopic place of origin in the cochlear partition) but uses only short-term rate information of the neural firings, over a prescribed time window, (2) *place/temporal* category, where place information is used together with detailed temporal information of local neural responses (i.e. higher-order firing statistics, like the interspike interval statistics), and (3) *non-place/temporal* category, where place information is omitted altogether and the only sources of information are the temporal properties of the global neural response (for an excellent overview of auditory models the reader is referred to [8]). These models of auditory periphery are feed-forward models, based on our understanding of the ascending auditory pathway up through the auditory nerve. A rigorous study of the capabilities of these models to reliably represent speech signals in a variety of acoustic conditions (e.g., different sound intensities, and presence of background noise) reached the widely accepted notion that place/rate models are insufficient, and that (at least) some degree of temporal information is required.

One auditory mechanism that may play a role in the robustness of the auditory periphery in the presence of background noise is the medial olivocochlear (MOC) efferent feedback system. Numerous studies have been published providing detailed morphological and neurophysiological description of the system (e.g. Guinan [9]), as well as psychophysical accounts for its effect on the sensory representation of signals embedded in noise. MOC efferents originate from neurons in the medial superior olivary nucleus (MSO) and terminate directly on outer hair cells (OHC). They have tuning curves similar to, or slightly broader than those of AN fibers (e.g. Guinan [9]), and they project to different places along the cochlear partition in a tonotopic manner. We currently do not have a clear understanding of the functional role of this mechanism. One speculated role, which is of particular interest for the current study, is a dynamic regulation of the cochlear operating point that depends on background acoustic stimulation and which results in robust human performance in perceiving speech in a noisy background. There are a few neurophysiologcal studies consistent with this hypothesis. Using anesthetized cats with noisy acoustic stimuli, Winslow and Sachs showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is partially recovered, [14]. Measuring neural responses of *awake* cats to noisy acoustic stimuli, May and Sachs showed that the dynamic range of

discharge rate in cochlear-nucleus neurons is only moderately affected by changes in levels of background noise, [11]. Finally, a few behavioral studies indicate the potential role of the MOC efferent system in perceiving speech in the presence of background noise. Dewson presented evidence that MOC lesions impair monkeys' ability to discriminate between the vowels [i] and [u] in the presence of masking noise, but have no effect on performance in quiet, [3]. More recently, Giraud et al. ([5]) and Zeng et al. ([15]) showed, albeit inconclusively, that the performance of humans with severed MOC feedback results in relatively poor phoneme perception when the speech is presented in a noisy background.

Inspired by this evidence we have developed a closed-loop model of the auditory periphery (i.e. PAM) which uses feedback to regulate the operating point of a model of cochlear mechanics, resulting in an auditory nerve representation less sensitive to changes in environmental conditions. In implementing the PAM we use a bank of overlapping cochlear channels uniformly distributed along the ERB (equivalent rectangular bandwidth) scale, four channels per ERB. Each cochlear channel comprises a nonlinear filter and a generic model of the inner hair cell (IHC) – half-wave rectification followed by low-pass filtering,

Figure 2: A block diagram of one cochlear channel. The central processor uses place/rate strategy.

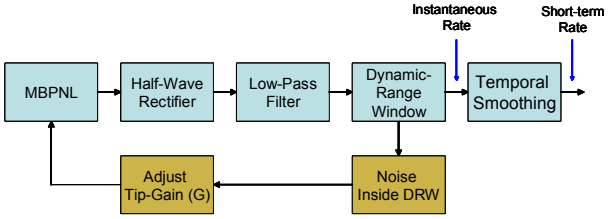
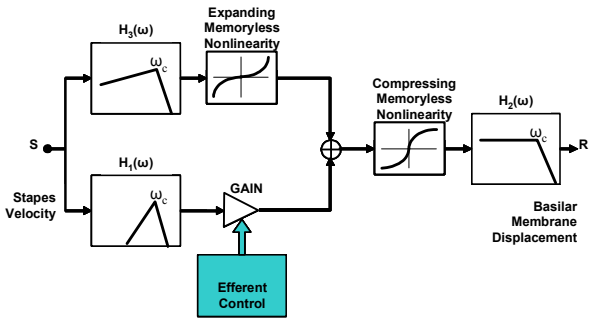


Figure 3: Goldstein's multi bandpass nonlinearity model, MBPNL, [6].

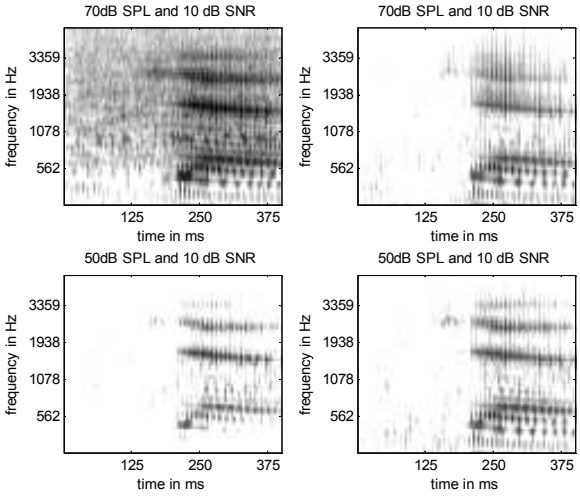


representing the reduction of neural synchrony with AN fiber characteristic frequency (CF). The dynamic range of the simulated IHC response is restricted to a dynamic-range window (DRW), representing the observed dynamic range at the AN level. The simulated IHC response (representing *instantaneous* firing rate at the AN) is smoothed temporally (temporal time integration over a 10-ms window), resulting in a short-term average-rate representation. (See Fig. 2.) The cochlear filter is Goldstein's MBPNL model of nonlinear cochlear mechanics, [6]. It 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. A parameter (GAIN) controls the gain of the tip of the simulated basilar membrane tuning curves.

As for the efferent-inspired part of the model we mimic the effect of the medial olivocochlear efferent path (MOC). Recall that morphologically, MOC neurons project to different places along the cochlear partition in a tonotopic manner, making synapse connections to the outer hair cells and hence affecting the mechanical properties of the cochlea (e.g. increasing basilar membrane stiffness). Therefore, we introduce a frequency-dependent feedback mechanism which controls the tip-gain of each MBPNL channel, permitting a prescribed intensity level of the sustained noise inside the DRW.

Figure 4 shows – in terms of a spectrogram – simulated IHC responses to diphone ***je*** (as in "jab") in two noise conditions (70 dB SPL / 10 dB SNR and 50 dB SPL / 10 dB SNR), for an open-loop MBPNL-based system (left-hand side) and for the closed-loop system (right-hand side). Due to the

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



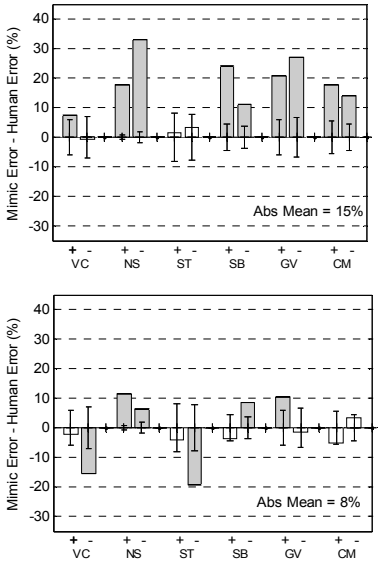
nature of the noise-responsive feedback, the closed-loop system produces spectrograms that fluctuate less with changes in noise intensity compared to spectrograms produced by the open-loop system. This property is desirable for stabilizing the performance of template-matching under varying noise conditions, as reflected in the quantitative evaluation reported in Section 2.1 .

2.1. Quantitative evaluation – isolating PAM from template matching

The evaluation system comprises a PAM followed by a TMC. *Ideally*, to eliminate PAM-TMC interaction, errors due to template matching should be reduced to zero (i.e. ideal template-matching). In *reality* we could only minimize interaction. This was acheived by using a methodology detailed in Ghitza *et al*. [4], in which the simplest possible psychophysical task in the context of speech perception was used, i.e. a binary discrimination test (Voiers' Diagnostic Rhyme Test [DRT], [13], in particular). To further reduce PAM-TMC interaction we have *synthesized* DRT word-pairs, restricting stimulus (waveform) differences to the initial diphones only. With such constraints it was reasonable to use a template-matching operation with a minimum mean squares error as the distance measure, allowing us to focus on errors attributed to the PAM alone (Ghitza *et al*. [4]).

Formal DRT sessions using human subjects have been conducted using the synthetic stimuli in quiet and in additive, speech-shaped noise at three levels (50, 60 and 70 dB SPL) and at three SNRs (0, 5 and 10 dB). Fig. 5 shows the errors produced by a DRT mimic with open-loop and closed-loop PAMs, compared to those made by human listeners. Signal conditions were the same as those used to collect the human data. Templates were created for the 60 dB SPL / 5 dB SNR condition. The abscissa marks the Jakobsonian dimensions: Voicing, Nasality, Sustention, Sibilation, Graveness and Compactness (denoted VC, NS, ST, SB, GV and CM, respectively). The "+" sign stands for an attribute being present and the "-" sign for an attribute being absent. Bars show the *difference* between the average machine and human scores. The lines indicate plus and minus one standard deviation of the human data. Gray bars indicate that the difference is greater than one standard deviation. Scores with the open-loop PAM are worse than those of the human scores. Scores with the closed-loop PAM are similar to

Figure 5: DRT mimic scores for open-loop (upper) and closed-loop (lower) PAM.



human scores except for VC- and ST-. Two points are noteworthy. First, when a sever mismatch occurs, closed-loop scores are *superior* to human scores while open-loop scores are worse. Hence, improving the open-loop system will require the exploitation of information beyond short-term rate (i.e. *temporal*). Second, although we predicted human performance in a binary task, parameters of the model were tuned to match errors between minimal pairs, *jointly* along *all* Jakobsonian dimensions. Hence we believe that the spectro-temporal patterns generated by the resulting closed-loop PAM are an adequate description of the sensory representation of degraded speech.

3. THE TEMPLATE MATCHING CIRCUIT

In developing the PAM (Section 2) we used synthetic speech stimuli, with restricted phonemic variation, hence permitting the use of a minimum mean squares error as the distance measure for template matching. In this section we consider naturally spoken speech stimuli, seeking a perceptually relevant distortion measure between speech tokens that exhibit phonemic variability.

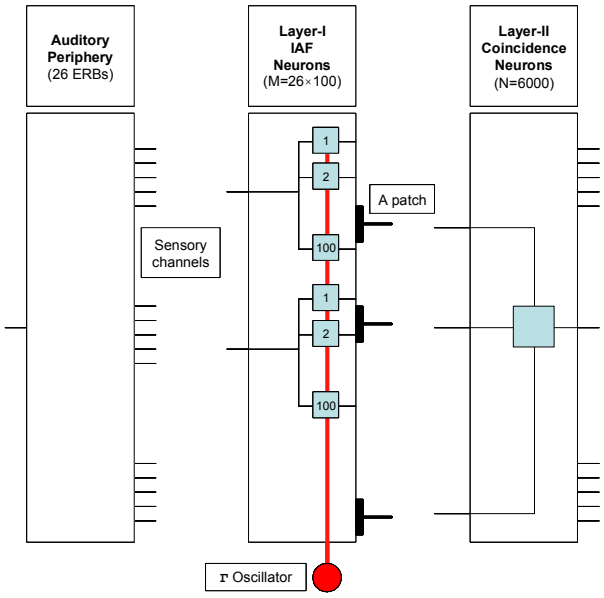
3.1. Why use models of neural computation?

In some sense speech decoding can be conceptualized as a search process, in which the search engine performs a template-matching operation comprised of two separate, but related steps. The first measures the *distance* between the

current input (e.g. a syllable) and (stored) templates. The second associates the input with the best-matching template. In this sense, template matching is defined by the *choice* of templates as well as a *distance metric*. To develop algorithms capable of emulating human performance we first need to create accurate, detailed models for both stages of the search process. An explicit, *analytical* expression is difficult to derive for such models. Instead, we seek to emulate neural computation principles that are general in nature and shared across sensory (e.g. auditory, visual, olfactory) and motor modalities. We suggest that a template-matching operation based on a *plausible* model of pertinent neural computation may implicitly incorporate characteristics essential for both the templates and the distance metric.

Ghitza *et al*. [4], have recently developed a template-matching circuit (TMC) designed to recognize and label diphone units in the speech signal. The TMC is based on principles of cortical neural processing used by Hopfield, [10]. In this circuit, a diphone is represented (or "stored") as a distinct group of neuronal states optimally tuned to

Figure 6: A block diagram of the TMC. The front-end is a filter bank with 26 critical-band channels spanning the range of the speech spectrum. Each channel drives 100 Layer-I integrate-and-fire neurons. The parameters of all Layer-I neurons are identical except for the threshold-of-firing. All Layer-I neurons are driven by a single, global, sub-threshold oscillatory current with a frequency in the gamma range. Each Layer-II coincidence neuron is driven by six randomly selected "patches" of Layer-I neurons.



the time-frequency signature of candidate diphones (whose durations range between 30-80 ms). Using the TMC we computed the precision with which diphones are recognized. Syllable-initial diphones (i.e. consonant-vowel, CV syllables) are typically identified more accurately than their syllable-final (i.e. coda, VC) counterparts. This result is consistent with both linguistic perception and with statistical analyses of conversational corpora where spectro-temporal variability of coda consonants is far greater than their consonantal counterparts in syllable onsets (Greenberg, [7]). Fig. 7 illustrates the behavior of the TMC in the DRT task. An important property of the TMC is its insensitivity to time-scale variation (consistent with Hopfield's original formulation). Such time-scale insensitivity (e.g. to variation in speaking rate) is essential for recognizing phonetic entities that are inherently variable in time and spectrum. These are the sort of properties that characterize human speech comprehension and which could prove useful for many technical applications in speech recognition, synthesis, auditory prostheses.

3.2. Why neural rhythmicity?

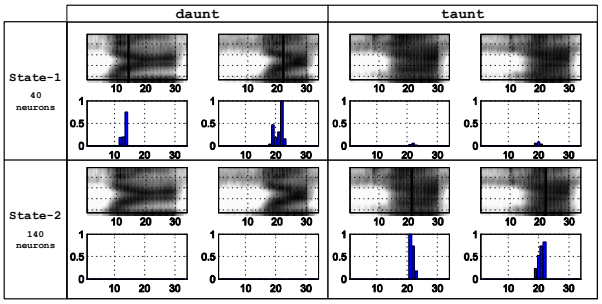
A crucial component of the TMC is a quasi-autonomous, sub-threshold oscillatory input of ca. 25 Hz. This oscillation feeds into all input neurons, serving as a synchronizing pacemaker (Hopfield, [10]), similar to the synchronization facilitator proposed by Singer and others for cortical processing (e.g. Singer, [12]; Buzsáki, [2]).

Such neural rhythms may play an important role in spoken-language comprehension. The specific timing of activation across the cortex can be visualized with electromagnetic recordings (e.g. magneto-encephalography, MEG). Typically, an increase in oscillatory activity is observed in specific rhythm bands, depending on the task. Of particular importance are the gamma (30-80 Hz) and the theta (3-10 Hz) rhythms (e.g. Bastiaansen and Hagoort, [1]). Theta oscillations are most closely associated (linguistically) with the syllable (mean duration 200 ms, [7]). Gamma oscillations are most closely associated with units important for diphone and other phonetic analyses.

4. CONCLUSIONS

In this presentation we suggest that robustness against background noise is provided principally by the signal processing performed by the

Figure 7: Performance of the TMC in the DRT task. State-1 represents 40 Layer-II neurons most sensitive to the initial diphone of the word "daunt." Analogously, State-2 represents 140 neurons for the word "taunt." The two upper-left panels show a spectrographic display of the front-end in response to the first 350 ms of two different realizations of the word "daunt". Below each spectrogram is a time-histogram of the number of State-1 neurons responding to the corresponding stimulus (shown is the pertinent fraction out of 40). The lower-right four panels show the analogous display for the response of State-2 neurons to the word "taunt". The lower-left (and the upper-right) panels show the response of the neurons to the other word. Note the strong response to stimuli of matched tokens (and weak response to opposite tokens).



peripheral circuitry. We showed that with an efferent-inspired closed-loop model of the cochlea, a place/rate model of central processor is sufficient to predict human performance in discriminating speech stimuli (with rich, relevant time-varying spectral patterns) in the presence of noise. This result is in contrast to the current notion based upon feed-forward models which suggests that a temporal (place or non-place) strategy is necessary in order to account for the robust human performance in noise. We point out that current understanding of auditory perception is based upon measurements with anesthetized animals (where the descending pathway is not functioning), and suggest that studies with awake animals (a very difficult task to perform) may alter current perspectives.

We also described a neuronal circuit with template-matching capabilities, using a gamma rhythm at the core to facilitate synchronization across the neuronal array. The circuit exhibits properties, such as time-scaling insensitivity, consistent with (and desirable for) perception of spoken language. Although the functional role of the rhythms in language decoding is unknown at present, we speculate that the range of rhythms may serve as a hierarchical synchronization mechanism by which the CNS integrates language information (phones by gamma, words by theta, and sequences of words by delta rhythms).

5. ACKNOWLEDGMENT

This work is supported by the U.S. Air Force Office of Scientific Research. The work described in Section 2 was performed in collaboration with D. Messing, L. Delhorne and L. Braida at MIT.

6. REFERENCES

[1] Bastiaansen, M. and Hagoort, P. 2006. Oscillatory neuronal dynamics during language comprehension. *Prog. Brain Res.* 159, 179-196.

[2] Buzsáki, G. 2006. *Rhythms of the Brain*. New York: Oxford University Press.

[3] Dewson, J. H. 1968. Efferent olivocochlear bundle: Some relationships to stimulus discrimination in noise. *J. Neurophysiol*. 31, 122–130.

[4] 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,* Berlin: Springer Verlag, in press.

[5] Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., Chery-Croze, S. 1997. Olivocochlear efferents involved in speech-in-noise intelligibility. *Neuroreport* 8, 1779-1783.

[6] Goldstein, J.L. 1990. Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering. *Hear. Res*. 49, 39-60.

[7] Greenberg, S. 1999. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29, 159-176.

[8] Greenberg, S. (ed.) 1988. *Representation of Speech in the Auditory Periphery. J. Phon.* 16, 1-149

[9] Guinan, J. J. 1996. Physiology of olivocochlear efferents. In: Dallos, P., Popper, A. N. Fay, R.R., (eds). *The Cochlea*, New York: Springer Verlag, 435-502.

[10] Hopfield, J.J. 2004. Encoding for computation: Recognizing brief dynamical patterns by exploiting effects of weak rhythms on action-potential timing. *Proc. Nat. Acad. Sci.* 101, 6255-6260.

[11] May, B.J., Sachs, M.B. 1992. Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. *J. Neurophysiol.* 68, 1589–1603.

[12] Singer, W. 2005. Putative role of oscillations and synchrony in cortical signal processing and attention. In: Itti, L.,,Rees, G., Tsotsos, J.K. (eds.) *Neurobiology of Attention.* Amsterdam: Elsevier, 526-533.

[13] Voiers, W. D. 1983. Evaluating processed speech using the diagnostic rhyme test. *Speech Techn*. 1 30–39.

[14] 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.

[15] Zeng, P. G., Martino, K. M., Linthcum, F. H., Soli, S. (2000). Auditory perception in vestibular neurectomy subjects. *Hearing Res*. 142, 102-112.