Optimization of sigmoidal rate-level function based on acoustic features

Víctor Poblete[1,3], *Néstor Becerra Yoma*[1], *Richard M. Stern*[2]

[1] Speech Processing and Transmission Laboratory, Universidad de Chile, Santiago, Chile
[2] Department of Electrical and Computer Engineering, and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Institute of Acoustics, Universidad Austral de Chile, Valdivia, Chile

vpoblete@ing.uchile.cl, nbecerra@ing.uchile.cl, rms@cs.cmu.edu

Abstract

This paper describes the development of an optimal sigmoidal rate-level function that is a part of many models of the peripheral auditory system. The optimization makes use of a set of criteria defined on the basis of physical attributes of the input sound that are inspired by physiological evidence. These criteria attempt to discriminate between a degraded speech signal and noise to preserve the maximum information in the linear region of the sigmoidal curve, and to minimize the effects of distortion in the saturating regions. The performance of the proposed approach is validated by text-independent speaker-verification experiments with signals corrupted by additive noise at different SNRs. Experimental results suggest that the approach presented in combination with CVN can lead to relative reductions in EER as great as 30% when compared with the use of baseline MFCC coefficients for some SNRs.

Index Terms: Auditory models, sigmoidal function, robust speaker verification.

1. Introduction

Neural processing of speech is represented by temporal patterns of neural impulses transmitted along the auditory-nerve fibers, which vary in response to the incoming sound. This dependence is summarized by curves called rate-level functions [1,2] which display a variety of forms, although they are usually sigmoidal [3,4]. Rate-level curves are characterized by: (1) discharge threshold; (2) maximum discharge rate; (3) spontaneous discharge rate; and (4) dynamic range [5]. As described by Young [6], dynamic range refers to "the range of sound levels over which the fiber changes its rate when the input changes in level." Most of the individual fibers exhibit a dynamic range of less than 35 dB [7]. In contrast, the dynamic range of loudness perception for humans is as great as 100 dB [8]. There have been a number of hypotheses concerning how humans can perceive loudness over a wide range while the range of the fibers is limited [9-11]. Currently, attention has also focused on the ability of the auditory-nerve to develop rate-level functions that adapt according to stimulus levels [12-14]. For decades the auditory system has attracted the interest of researchers in speech processing, including the use of models as part of the feature extraction process for speech recognition, speaker verification, etc. [15,16]. Most of these models begin with a bank of filters followed by a model of the rate-level nonlinearity (*eg.* [17-19]). In literature on the auditory system, mainly in mammals, there is extensive physiological evidence that demonstrates that the nonlinear cochlear transduction is strong related to the way complex sounds are perceived, especially in complex acoustic backgrounds. The evidence suggests that the cochlea is able to adapt its auditory behavior to intensity changes over a wide range of sound levels. This ability, combined with the cochlear frequency selectivity, would facilitate our skill to communicate in quiet and in varying levels of background noise. In addition, the detection of communication sounds against a background of environmental noise is a fundamental problem that not only benefits the human itself, but also many other animal species. The problem is of considerable interest and has several practical applications, which is important for research opportunities and application viewpoints. In this paper we represent the nonlinear cochlear transduction by a sigmoidal-shaped function. Our method proposes to find the optimal parameters of sigmoidal functions based exclusively on acoustic features inspired by the physiological evidence, and we did not take into consideration phonetic classification schemes. The approach is applied to a text-independent speaker verification task with signals corrupted by additive noise at different SNRs. It is worth noting that, in principle, this method is applicable to any speech processing task because all the analysis takes place in the acoustic signal domain.

2. Development of the optimization criteria

2.1. Physiological evidence in animals

Despite different species hear over different frequency ranges, their auditory nerve activities are comparable. Studies in mammals attempt to explain dynamic adaptation of the rate-level functions with respect to the intensity of the input sound, background noise intensity, and the contrast between noise and the degraded-speech signal [22-25]. In rats higher sound levels tend to move the rate-level curves to the right and increase their slopes [26,27]. Research in cats has demonstrated that the background noise causes in the rate-level curves a shift of the dynamic range to higher intensities and that their slopes can increase in presence of noise [7]. Research in ferrets has shown that another property is enhancement of spectro-temporal contrast in the acoustic environment [22,28,29], similar to the spatio-temporal contrast developed in the vertebrate retina [30-31]. These goals, in combination with reducing the nonlinear distortion of the degraded-speech and reducing differences between original clean speech and degraded-speech [20,21], lead to the definition of four optimization criteria described in Sec. 2.3.

2.2. Specification of the sigmoidal function

Let us represent the rate-level nonlinearity by the sigmoidal function $g(l)$ given by:

$$g(l) = \frac{1}{1 + e^{\omega(l-\mu)}} \quad (1)$$

where μ and ω correspond to the offset and the slope of $g(l)$, respectively. $g(l)$ allows modeling the nonlinear response in Fig. 1. μ corresponds to the location along the horizontal axis where $g(l)$ equals $1/2$. Thus, μ and ω are the parameters to be estimated. Let us now consider the output of a channel of the filter bank. The degraded-speech input signal $x_{j,k}$ at the output of filter j at the discrete-time index k is given by:

$$x_{j,k} = s_{j,k} + n_{j,k} \quad (2)$$

where $s_{j,k}$ and $n_{j,k}$ denote the clean speech and noise signal, respectively. If $x_{j,k}$ is divided into N_f frames of W samples per frame, the log-energy at frame i at filter j, $E_{j,i}$ can be written as:

$$E_{j,i} = 10 \cdot \log\left(\sum_{k \in frame\ i} w_{i-k}^2 \cdot x_{j,k}^2 \right) \quad (3)$$

where w_k represents the response of the window function.

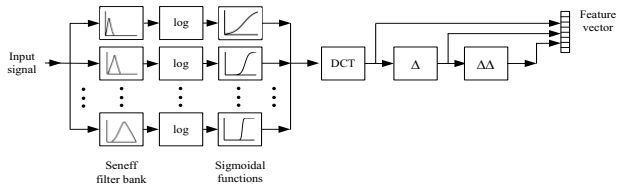

Figure 1: Block diagram of the proposed scheme.

Histograms of $E_{j,i}$ are generated to discriminate degraded-speech and noise frames by using the voice activity detector (VAD) proposed by Shin *et al.* [32]. Frames are divided into two subsets: degraded-speech and only noise frames. N_f^{sn} and N_f^n, ($N_f = N_f^{sn} + N_f^n$), indicate number of degraded-speech frames and number of noise alone frames, respectively. Finally, $E_{j,i}^x$ $(1 \le i \le N_f)$, $E_{j,m}^{sn}$ $(1 \le m \le N_f^{sn})$ and $E_{j,r}^n$ $(1 \le r \le N_f^n)$ represent the energies at filter j and at frame i, m and r for frames that are considered to belong to the original input, the subset of frames that contain degraded-speech and the subset of input frames that contain noise alone, respectively. Also, the mean and variance of the energy in the degraded-speech frames are $\mu_{j,sn}$ and $\sigma_{j,sn}^2$, respectively, while the mean and variances of the energy in the frames that contain only noise are $\mu_{j,n}$ and $\sigma_{j,n}^2$, respectively.

2.3. Specification of the objective function

We choose an objective function for the sigmoidal nonlinearity that (1) minimizes nonlinear distortion in the linear region, (2) minimizes noise power, (3) maximizes similarity between energy in the frames that represent degraded-speech and energy of the speech alone in those frames, and (4) maximizes the energy in the output signal which is presumed to be dominated by speech.

Criterion 1: Nonlinear distortion in the linear region. ω and μ should be chosen in such a way that the degraded-speech lies in the linear part of the sigmoidal curve. This nonlinear distortion, $D_j^{non-linear}$, is defined as:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\mathbf{E}\{[A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})]^2\}}{\mathbf{E}[(E_{j,m}^{sn})^2]} \quad (4)$$

where $g(\cdot)$ represents the sigmoidal function and $\mathbf{E}[\cdot]$ is the expectation operator. A_j and B_j correspond to a linear transformation that allows the comparison of $E_{j,m}^{sn}$ and $g(E_{j,m}^{sn})$ (see Appendix). By approximating the expected value by the sample mean, $D_j^{non-linear}(\omega_j, \mu_j)$ is:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})]^2}{\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn}\right)^2} \quad (5)$$

Criterion 2: Noise power. $g(l)$ can be employed to attenuate the noise due to the fact that low-energy frames can be associated with noise. The power of the noise after $g(l)$, $P_j^{noise}(\omega_j, \mu_j)$ is given by:

$$P_j^{noise}(\omega_j, \mu_j) = \mathbf{E}[g^2(E_{j,r}^n)] \quad (6)$$

$g(l)$ should minimize $P_j^{noise}(\omega_j, \mu_j)$ in order to reduce the effect of noise energy. By estimating the expected value as the sample mean, $P_j^{noise}(\omega_j, \mu_j)$ can be rewritten as:

$$P_j^{noise}(\omega_j, \mu_j) = \frac{1}{N_f^n} \sum_{r=1}^{N_f^n} g^2(E_{j,r}^n) \quad (7)$$

Criterion 3: Similarity between clean speech and the degraded-speech input. The use of a rate-level function should reduce differences between average frequency response of clean speech and average frequency response of the degraded-speech, both assessed after the sigmoidal nonlinearity [21]. Thus, the difference between the energies of the clean speech $E_{j,i}^s$ and the degraded-speech input $E_{j,i}^x$ is represented by:

$$D_j^{\text{clean-noisy}}(\omega_j, \mu_j) = \sum_{i=1}^{N_f} [g(E_{j,i}^s) - g(E_{j,i}^x)]^2 \quad (8)$$

Criterion 4: Signal variance of degraded-speech after processing by sigmoidal function. To avoid extreme compression or saturation, the variance of the resulting degraded-speech after the sigmoidal function should be maximized. This variance $V_j(\omega_j, \mu_j)$, is expressed as:

$$V_j(\omega_j, \mu_j) = \sigma^2[g(E_{j,m}^{sn})] \quad (9)$$

By expanding, $V_j(\omega_j, \mu_j)$ can be rewritten as:

$$V_j(\omega_j, \mu_j) = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} g^2(E_{j,m}^{sn}) - \left[\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} g(E_{j,m}^{sn}) \right]^2 \quad (10)$$

Specification of the complete objective function. The objective function $J(\omega_j, \mu_j)$ is defined as [33]:

$$J(\omega_j, \mu_j) = D_j^{non-linear}(\omega_j, \mu_j) + P_j^{noise}(\omega_j, \mu_j) + \\ + D_j^{\text{clean-noisy}}(\omega_j, \mu_j) - V_j(\omega_j, \mu_j) \quad (11)$$

Consequently, the optimal slope $\hat{\omega}_j$ of the sigmoidal function is estimated as:

$$\hat{\omega}_j = \arg\min_{\omega_j} \left\{ J(\omega_j, \mu_j) \right\} \quad (12)$$

In (12), μ_j is set to $\mu_j = \mathbf{E}[(E_{j,m}^{sn})]$ (i.e. centered on the mean of $E_{j,m}^{sn}$). Finally, $\hat{\mu}_j$ is estimated according to:

$$\hat{\mu}_j = \arg\min_{\mu_j} \left\{ J(\omega_j, \mu_j) \right\} \quad (13)$$

In (13), ω_j corresponds to the optimal sigmoidal slope $\hat{\omega}_j$. While we recognize that the definition of $J(\omega_j, \mu_j)$ as the simple sum of the four criteria above is a special case of the more general linear combination:

$$J(\omega_j, \mu_j) = a \cdot D_j^{non-linear}(\omega_j, \mu_j) + b \cdot P_j^{noise}(\omega_j, \mu_j) + \\ + c \cdot D_j^{\text{clean}-\text{noisy}}(\omega_j, \mu_j) - d \cdot V_j(\omega_j, \mu_j)$$

we adopted the function of (11) for simplicity in the absence of compelling evidence that other combinations of the four criteria would provide better performance.

3. The sigmoidal rate-level function

Optimal values $\hat{\omega}_j$ and $\hat{\mu}_j$ were determined using development database of speech corrupted by babble noise at an SNR equal to 10 dB, as discussed in Sec. 4. These values vary from channel to channel. This is due to the fact that the SNR varies from one filter to the other. Two subsets of frames are defined based on the VAD results: degraded-speech and only noise frames, respectively. Figure 2 describes an example of $J(\omega_j, \mu_j)$ for each the 35 analysis bands j. Figure 3 depicts an optimal sigmoidal function (solid line) and its corresponding linear mapping (dotted line). The sigmoidal curve was obtained with $\hat{\omega}_j$ and $\hat{\mu}_j$ for the filter $j = 8$. Figure 3 also depicts a histogram extracted from a testing utterance for filter $j = 8$. As can be seen in Fig. 3, the sigmoidal function compresses the noise in the nonlinearity region, for the most part the frame containing degraded-speech lie within the linear part. Figure 4 shows for the filter $j = 17$, four optimal sigmoidal functions trained with babble noise at SNRs equal to 20 dB, 15 dB, 10 dB and 5 dB, along with a fifth sigmoidal function that was trained with clean speech. As shown in Fig. 4, both $\hat{\omega}_j$ and $\hat{\mu}_j$ depend on the SNR at which the sigmoidal function was trained: as SNR is increased, the curves shift to the right and become steeper. Thus, the optimization of $\hat{\omega}_j$ and $\hat{\mu}_j$ provides an adaptation in the sigmoidal function that compensates for variations in SNR. The adaptation is consistent with the physiological literature as noted in Sec 2.1.

4. Discussion and conclusions

The utility of the optimal sigmoidal nonlinearity was evaluated using a text-independent speaker verification task, with equal error rate (EER) employed as the major figure of merit. The results were obtained using the Yoho database [34]. The database is divided into "enrollment" and "verification" segments. In this paper a subset of 70 speakers was employed. These speakers were divided: 40 background impostor speakers to train the background models; and 30 testing speakers were used in verification attempts.

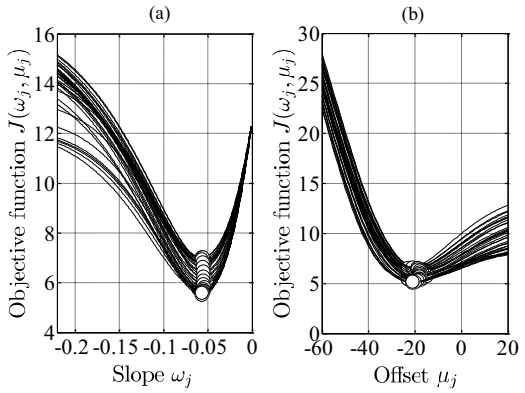

Figure 2: Objective function $J(\omega_j, \mu_j)$ plotted as a function of the sigmoidal function parameters: (a) slope and (b) offset. $\hat{\omega}_j$ and $\hat{\mu}_j$ are depicted for each of the 35 channels of the filter bank.

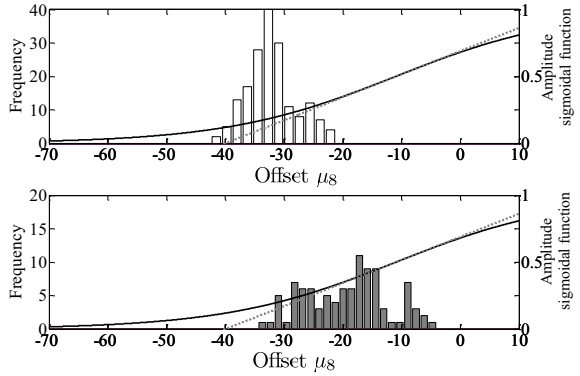

Figure 3: Example of optimal sigmoidal function (solid line) and corresponding linear mapping (dotted line). Histograms of power are also depicted for frames containing degraded-speech (filled bars) and noise only frames (open bars). Training conditions for sigmoid were babble noise at SNR=10dB. Results for filter $j = 8$ are plotted with optimal parameters: $\hat{\omega}_8 = -0.071$; and $\hat{\mu}_8 = -14$.

Additionally, a subset composed of 50 speakers and one utterance per speaker (development database) extracted from Yoho was employed to train $\hat{\omega}_j$ and $\hat{\mu}_j$. Babble noise was artificially added to the Yoho corpus to generate noisy versions of the utterances at various SNRs [35]. For all the speaker verification experiments, the system was trained with clean speech. Speech signals were filtered by a filter bank of 35 channels (extracted from the Seneff model [19]) that cover frequencies from 220-3300 Hz. After filtering, the signals were divided into 25-ms frames with 12.5-ms overlap between frames using Hamming windows. The log-energy was computed at the output of each filter. Then, an optimal sigmoidal function, estimated using the development database and the procedure explained in Sec. 2, was applied to the log-energy of each filter to both the training and testing data sets. The log-energy plus ten static cepstral coefficients, and their first and second time derivatives were estimated (see Fig. 1). Four configurations were used: (1) baseline system, (log-energies of the Seneff filter bank output); (2) baseline system with cepstral variance normalization (CVN); (3) baseline system with cepstral mean and variance normalization (CMVN); and, (4) method proposed using the optimal sigmoidal function, combined with CVN. If the entire signal were mapped into the linear region, the proposed scheme

could be considered equivalent to the CVN algorithm. Therefore, to show the effect of the nonlinearity provided by the sigmoidal function, CVN is combined with our method using the optimal sigmoidal function.

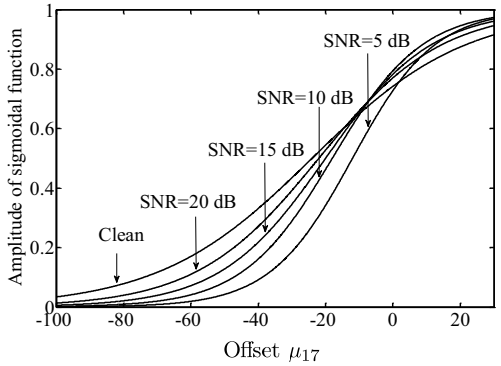

Figure 4: Sigmoidal functions plotted as a function of SNR. Training conditions: clean speech, and speech degraded by noise at SNRs equal to 20dB, 15 dB, 10 dB and 5 dB. Results for filter $j=17$ are plotted with optimal parameters.

In the verification procedure, the normalized log likelihood is estimated. The universal background model (UBM) is trained by using the background impostor speakers. A speaker-dependent Gaussian mixture model GMM is generated for each speaker by employing MAP adaptation [36]. The number of Gaussian components used in the UBM and speaker models was 256 and the covariance matrices were diagonal.

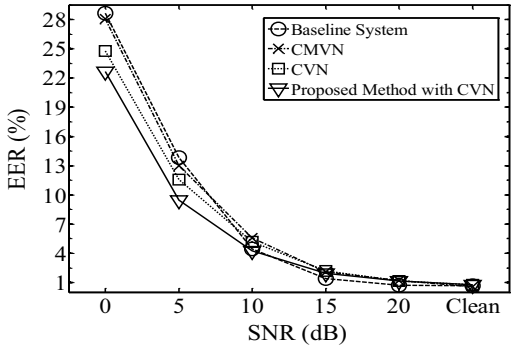

Figure 5: Comparison of EER as a function of SNR for speech in babble noise. The optimal sigmoidal functions were trained with babble noise at an SNR of 10 dB.

Figure 5 describes EER results obtained as a function of SNR for speech in the presence of babble noise. Results are compared for the baseline system, the baseline system combined with CVN, the baseline system with CMVN and the proposed method consisting of the combination of the Seneff filters, CVN, and the optimal sigmoidal nonlinearity. The use of the optimal sigmoidal function in combination with CVN provides relative reductions in EER compared with the baseline system as great as 20.9%, 31.7% and 4.5% at SNRs equal to 0 dB, 5 dB and 10 dB, respectively. The corresponding relative reductions in EER are 13.5% and 16.2% at SNR equal to 0 dB and 5 dB, respectively, when comparing the CVN system with the baseline. Similarly, the relative reductions in EER are 2.13%, 6.01% and 14.7% at SNR equal to 0 dB, 5 dB and clean speech, respectively, when comparing the CMVN system with the baseline. The use of the sigmoidal function combined with CVN provides relative reductions in EER compared to CVN alone equal to 8.48%, 18.4%, 17.8%, 9.72% and 1.33% at SNRs of 0 dB, 5 dB, 10 dB, 15 dB and clean speech, respectively. Finally, this paper describes a method that can be used to develop an optimal sigmoidal nonlinear function for auditory modeling that is based solely on the distribution of power in the degraded speech frames and the power in the frames containing noise only. The objective function that is described attempts to simultaneously minimize nonlinear distortion, minimize noise power, maximize the similarity between frames that are believed to contain noise alone and frames that are believed to contain degraded-speech, and maximize the resulting speech power.

5. Appendix

The parameters A_j and B_j are estimated according to:

$$(A_j, B_j) = \arg\min_{A_j, B_j} \left\{ D_j^{non-linear}(\omega_j, \mu_j) \right\} \quad \text{(A1)}$$

First, the partial derivative of $D_j^{non-linear}(\omega_j, \mu_j)$ with respect to A_j is estimated:

$$\frac{\partial D_j^{non-linear}}{\partial A_j} = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})] \cdot E_{j,m}^{sn} \quad \text{(A2)}$$

Then, the result obtained in (A2) is set to zero:

$$\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})] \cdot E_{j,m}^{sn} = 0$$

$$A_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn} \right)^2 + B_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} =$$

$$= \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} \cdot g(E_{j,m}^{sn}) \quad \text{(A3)}$$

$$A_j \cdot \mathbf{E}\left[(E_{j,m}^{sn})^2 \right] + B_j \cdot \mathbf{E}\left[E_{j,m}^{sn} \right] = \mathbf{E}\left[E_{j,m}^{sn} \cdot g(E_{j,m}^{sn}) \right]$$

Similarly, by estimating the derivative of $D_j^{non-linear}(\omega_j, \mu_j)$ with respect to B_j and setting the result to zero, the following equation is obtained:

$$A_j \cdot \mathbf{E}\left[(E_{j,m}^{sn})^2 \right] + B_j = -\mathbf{E}\left[g(E_{j,m}^{sn}) \right] \quad \text{(A4)}$$

By combining (A3) and (A4) and making use of the expressions:

$$\mu_j = \mathbf{E}[E_{j,m}^{sn}] \text{ and } \sigma_j^2 = \mathbf{E}[(E_{j,m}^{sn})^2] - \{\mathbf{E}[E_{j,m}^{sn}]\}^2$$

the parameters A_j and B_j are found to be equal to:

$$A_j = \frac{1}{\sigma_j^2} \left\{ \mathbf{E}\left[E_{j,m}^{sn} \cdot g(E_{j,m}^{sn}) \right] - \mu_j \cdot \mathbf{E}\left[g(E_{j,m}^{sn}) \right] \right\} \quad \text{(A5)}$$

$$B_j = \mathbf{E}\left[g(E_{j,m}^{sn}) \right] - \mu_j \cdot A_j$$

6. Acknowledgements

This research was funded by Conicyt-Chile, under grants Fondecyt 1100195 and Team Research in Science and Technology ACT 1120, and ONRG N62909-12-1

7. References

[1] Pickles, J. O., "An Introduction to the Physiology of Hearing", 3rd ed. Emerald Group, Bingley, England, ch. 4, 2008.

[2] Moore, B. C. J., "An Introduction to the Psychology of Hearing", 5th ed. Academic Press, London, 39-41, 2003.

[3] Sachs, M. B. and Abbas, P. J., "Rate versus level functions for auditory-nerve fiber in cats: tone burst stimuli," J. Acoust. Soc. Am., 56(6): 1835-47, 1974.

[4] Yates, G. K., Winter, I. M. and Robertson, D., "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range," Hear. Res., 45(3): 203-219, 1990.

[5] Nizami, L., "Dynamic range relations for auditory primary afferents," Hear. Res., 208(1-2): 26-46, 2005.

[6] Young, E. D., "Neural representation of spectral and temporal information in speech," Phil. Trans. R. Soc. Lond. B, 363(1493): 923-945, 2008.

[7] May, B. J. and Sachs M. B., "Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats," J Neurophysiol., 68(5): 1589-1602, 1992.

[8] Winslow, R. L. and Sachs, M. B., "Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise," J. Neurophysiol., 57(4): 1002-1021, 1987.

[9] Viemeister, N. F., "Intensity coding and the dynamic range problem", Hear Res., 34(3): 267-74, 1988.

[10] Heinz, M. G., Issa, J. B. and Young, E. D. "Auditory-nerve rate responses are inconsistent with common hypotheses for the neural correlates of loudness recruitment", J. Assoc. Res. Otolaryngol., 6 (2): 91-105, 2005.

[11] Moore, B. C. J., "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants", Otol. Neurotol., 24: 243-254, 2003.

[12] Barbour, D. L., "Intensity-invariant coding in the auditory system," Neurosci. Biobehav. Rev, 35(10): 2064-2072, 2011.

[13] Dean, I., Harper, N. S., and McAlpine, D., "Neural population coding of sound level adapts to stimulus statistics," Nat. Neurosci., 8(12): 1684-89, 2005.

[14] Dean, I., Robinson, B. L., Harper, N. S., and McAlpine, D., "Rapid Neural Adaptation to Sound Level Statistics," J. Neuroscience, 28(25): 6430-6438, 2008.

[15] Stern, R. M. and Morgan, N., "Features based on auditory physiology and perception," in Techniques for Noise Robustness in Automatic Speech Recognition, T. Virtanen, B. Raj and R. Singh, Eds., Wiley, 2012a.

[16] Stern, R. M. and Morgan, N., "Hearing is believing: Biologically-inspired feature extraction for robust speech recognition," IEEE Signal Processing Magazine, 29: 34-43, 2012.

[17] Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition," IEEE Transactions on Speech and Audio Processing, 2(1): 115-132, 1994.

[18] Cohen, J. R., "Application of an auditory model to speech recognition," J. Acoust. Soc. Am., 85(6): 2623-2629, 1989.

[19] Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing," J. Phonetics, 16(1): 55-76, 1988.

[20] Chiu, Y.-H. B. and Stern, R. M., "Analysis of physiologically-motivated signal processing for robust speech recognition," in Proc. of Interspeech, Brisbane, Australia, 1000-1003, 2008.

[21] Chiu, Y.-H. B., Raj, B., and Stern, R. M., "Learning-based auditory encoding for robust speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, 20(3): 900-914, 2012.

[22] Sumner, C. J. and Palmer, A. R. "Auditory nerve fibre responses in the ferret", Eur. J. Neurosci., 36(4): 2428-39, 2012.

[23] Reiss, L. A., Ramachandran, R., and B. J. May, "Effects of signal level and background noise on spectral representations in the auditory nerve of the domestic cat", J. Assoc. Res. Otolaryngol., 12: 71-88, 2011.

[24] Zilany M. S. and Carney, L. H., "Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics," J. Neurosci., 30(31): 10380-90, 2010.

[25] Taberner, A. M. and Liberman, M. C., "Response properties of single auditory nerve fibers in the mouse," J. Neurophysiol., 93(1): 557-569, 2005.

[26] Gao, F., Zhang, J., Sun, X., and Chen, L., "The effect of postnatal exposure to noise on sound level processing by auditory cortex neurons of rats in adulthood," Physiol. Behav., 97: 369-373, 2009.

[27] Bureš, Z., Grécová, J., Popelář, J., and J. Syka, "Noise exposure during early development impairs the processing of sound intensity in adult rats," Eur. J. Neurosci., 32(1): 155-164, 2010.

[28] Rabinowitz, N. C., Willmore, B., Schnupp, J., and King, A. J., "Contrast Gain Control in Auditory Cortex," Neuron, 70(6): 1178-91, 2011.

[29] Wang, K. and Shamma, S., "Self-normalization and noise-robustness in early auditory representations," IEEE Trans. on Speech and Audio Proc., 2(3): 421-435, 1994.

[30] Ohzawa, I., Sclar, G., and Freeman, R. D., "Contrast gain control in the cat's visual system," J. Neurophysiol., 54(3): 651-658, 1985.

[31] Werblin, F. S., Jacobs, A., and Teeters, J., "The computational eye," IEEE Spect., 33(5): 30-37, 1996.

[32] Shin, J. W., Kwon, H. J., Jin, S. H., and Kim, N. S., "Voice activity detection based on conditional MAP criterion," " IEEE Signal Process. Lett, 15(2): 257-260, 2008.

[33] Poblete, V., Yoma, N. B. and Stern, R. M., "Sigmoidal rate-level function optimization based on acoustic features," submitted to Speech Communication, January, 2013.

[34] Campbell, J. and Higgins, A., "YOHO speaker verification," Linguistic Data Consortium, Philadelphia, PA, 1994.

[35] Hirsch, H. G. and Pearce, D., "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," in ISCA ASR2000-Automatic Speech Recognition: Challenges for the Next Millennium, Paris: 181-188, 2000.

[36] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., "Speaker verification using adapted Gaussian Mixture Models," Digital Signal Process., 10(1-3): 19-41, 2000.