Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 19–34

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Optimization of the parameters characterizing sigmoidal rate-level functions based on acoustic features

Víctor Poblete [a,c], Néstor Becerra Yoma [a,*], Richard M. Stern [b]

[a] Speech Processing and Transmission Laboratory, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile
[b] Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[c] Institute of Acoustics, Universidad Austral de Chile, Av. General Lagos 2086, P.O. Box 5111187, Valdivia, Chile

Received 1 February 2013; received in revised form 14 June 2013; accepted 20 July 2013
Available online 29 July 2013

Abstract

This paper describes the development of an optimal sigmoidal rate-level function that is a component of many models of the peripheral auditory system. The optimization makes use of a set of criteria defined exclusively on the basis of physical attributes of the input sound that are inspired by physiological evidence. The criteria developed attempt to discriminate between a degraded speech signal and noise to preserve the maximum amount of 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 optimal sigmoidal function is validated by text-independent speaker-verification experiments with signals corrupted by additive noise at different SNRs. The experimental results suggest that the approach presented in combination with cepstral variance normalization can lead to relative reductions in equal error rate as great as 40% when compared with the use of baseline MFCC coefficients for some SNRs.

Keywords: Sigmoidal function; Auditory systems; Optimization; Acoustic features; Speech enhancement

1. Introduction

Speech sounds are pressure waves that vary as a function of time. These sounds are passed through the peripheral auditory system before being converted into electrical neural activity in the auditory nerve (Pickles, 2008). A spectral decomposition is performed in the cochlea that separates the incoming speech sounds into their constituent frequency components, and information is passed on through the auditory nerve, the brainstem, ultimately to the auditory cortex through channels that remain frequency dependent. In a natural environment the target speech sounds and background noise enter the peripheral auditory system together. Nevertheless, one of the most compelling characteristics of the auditory system is its ability to respond to and distinguish speech sounds from background noise (Darwin, 2008). In this paper we introduce a new way to improve accuracy in speaker verification tasks by incorporating a particular type of adaptation in a representation used for feature extraction that is based on processing in the auditory periphery.

1.1. Neural processing of speech signals

Neural processing of speech is represented by the temporal patterns of neural impulses (or "spikes") transmitted along the auditory-nerve fibers, which vary in time in response to the incoming sound. The dependence of the average number of spikes per second on incoming signal intensity in a particular frequency region is summarized by curves called rate-versus-level functions (e.g. Moore, 2003; Pickles, 2008). Rate-level functions display a variety of forms, although they are usually sigmoidal (e.g. Sachs

* Corresponding author. Tel.: +56 2 29784205.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.07.006

and Abbas, 1974; Yates et al., 1990). Under these circumstances, rate-level functions can be characterized by four attributes: (1) discharge threshold; (2) maximum discharge rate; (3) spontaneous discharge rate; and (4) dynamic range (Nizami, 2005). As described by Young (2008), dynamic range in this context refers to "the range of sound levels over which the fiber changes its rate when the input changes in level."

Most of the auditory-nerve fibers exhibit a dynamic range of less than 35 dB when stimulated with tones at their characteristic frequency (May and Sachs, 1992). In contrast, the dynamic range of loudness perception for humans is as great as 100 dB of sound pressure level (Winslow and Sachs, 1987). Through the years there have been a number of hypotheses concerning how humans can perceive loudness changes over such a wide dynamic range while the intrinsic dynamic range of auditory-nerve fibers is limited to 20–35 dB. These speculations have included consideration of the distributions of the thresholds of individual auditory-nerve fibers, the spread of excitation of the fibers over frequency, and possible loudness coding based on synchronous response, at least at low frequencies (<3–4 kHz) (Shamma, 1985).

In recent years, attention has also focused on the potential ability of the response at the auditory nerve to develop rate-level functions that vary according to the distribution of stimulus levels (Barbour, 2011; Dean et al., 2005, 2008). For example, experiments with cats have shown that the dynamic range in their auditory neurons is adapted for tone and noise stimuli to the distribution of sound levels. This adaptation is characterized by shifting towards the most frequently occurring level (Wen et al., 2009, 2012). Rate-level functions in the guinea pig also exhibit a restricted and mutable dynamic range. In these animals, neural responses are rapidly adjusted and tend to improve coding of the sound levels (Dean et al., 2005). Auditory-nerve fibers in the mouse also display similar behavior, although with differences in frequency ranges (Taberner and Liberman, 2005). We elaborate on these results and some of their potential consequences in Section 2.

For many years the properties of the auditory system have attracted the interest of researchers in speech processing, including the use of models of the auditory system as part of the feature extraction process for automatic speech recognition, speaker verification, etc. Some of this work has been reviewed in Stern and Morgan (2012a,b), and the earliest computational models of the peripheral auditory system that have been developed include the work of Allen (1985), Ghitza (1986, 1994), Seneff (1988), Lyon (1982), Shamma (1988), and Cohen (1989). Most of these models begin with a bank of filters tuned to different center frequencies to model the spectral decomposition of incoming sounds into the cochlea, followed by a model of transduction that includes the sigmoidal nonlinearity of the auditory transduction process that transforms the mechanical motion in the cochlea to the production of auditory-nerve spikes. As an example of the latter mechanism, the Seneff model includes a representation of the inner hair cells that consists of four stages: (1) a rate-level nonlinearity that limits the responses to signal components of a particular frequency with very small and very large amplitudes, (2) short-term adaptation that models the release of neurotransmitters during the synapse stage, (3) a low-pass filter that models the loss of synchrony in response to signal components of high frequency, and (4) an automatic gain control that maintains a presence of high-intensity sounds when the auditory nerve is saturated. Seneff (1988) proposed two parallel paths to analyze the outputs of this representation. One path measures the instantaneous overall short-time energy appearing each channel output, and in the other develops a spectral representation based on the extent to which the output signal is synchronized to best frequency of response of the fiber. Over time numerous groups have used auditory models such as the ones listed above to develop features for use in speech recognition and speaker identification, among other technologies (e.g. Kim et al., 2006; Kim and Stern, 2012).

1.2. Feature extraction for speaker verification

In speaker verification the aim has been to determine whether a given speech signal belongs to a claimed person or not based only on a voice sample (Reynolds, 1995). Usually, a speaker verification system comprises three sections: feature extraction, speaker modeling (performed from the extracted features), and decision making (Kinnunen and Li, 2010). The feature extraction section is designed to provide enough discriminative information from the speech signal to enable the speaker to be verified (Li and Huang, 2011). The development of relevant features is clearly important to discriminate one speaker from another in a fashion that preserves verification accuracy in environments that are different from the original training environment (Kinnunen et al., 2012; Li and Huang, 2011; Shao and Wang, 2008). Differences in the environment may arise from various sources, including additive interfering noise (Ming et al., 2007) and variations in the transmission channel conditions over which the speech is being recorded (Wu et al., 2007). Resolving mismatches between training and testing environments remains one of the most challenging problems to be solved for successful speaker verification in real applications (Hasan and Hansen, 2013; Saeidi et al., 2010).

The most commonly used features for speaker verification have been short-time cepstral coefficients such as Mel-frequency cepstral coefficients (MFCC) (Ajmera et al., 2011; Hanilçi et al., 2012; Wang et al., 2011). The standard MFCC method performs reasonably well when training and testing environments are matched but verification accuracy degrades seriously under noisy environments, especially when training and testing conditions are mismatched (Kinnunen et al., 2012; Li and Huang, 2011). The greatest degradation in verification performance is observed when the speech signal is degraded by additive noise at a low SNR, especially when the system is trained on clean speech (Hanilçi et al., 2012; Kinnunen et al., 2012).

Feature extraction inspired by the physiology of the human peripheral auditory system has also been proposed to improve speaker verification performance under mismatched conditions (e.g. Li and Huang, 2010, 2011; Shao and Wang, 2008; Shao et al., 2007). For instance, Shao and Wang (2008), proposed auditory-based features known as Gammatone frequency cepstral coefficients (GFCC), which effectively replace the triangular frequency weighting used in the MFCC method by the use of Gammatone filters (Shao and Wang, 2008) to achieve frequency selectivity. Gammatone filters are widely used in models of the auditory system and were developed to mimic cochlear filtering (Patterson et al., 1992). Shao et al. (2007) have demonstrated that GFCC features can provide robust speaker recognition in the presence of additive noise over a wide range of SNRs, and performance can be further improved by adding complementary auditory scene analysis (Shao et al., 2010). Similarly, Li and Huang (2011) proposed the use of cochlear filter cepstral coefficients (CFCC) for robust speaker identification in mismatched conditions (Li and Huang, 2010, 2011). The proposed CFCC features are based on a time–frequency transform called the Auditory Transform that includes several components that mimic the processing in the human peripheral auditory system (Li and Huang, 2011). CFCC features also improve speaker identification accuracy compared to conventional MFCC processing when tested under mismatched conditions (Li and Huang, 2010). Other recent auditory-based features include the Teager energy cepstrum coefficients (TECC), developed by Dimitriadis et al. (2011).

1.3. The sigmoidal rate-intensity function

In an earlier study, Chiu and Stern (2008) examined the contributions of each stage of the classic auditory model by Seneff (1988) to analyze their impact in improving recognition accuracy for speech in the presence of noise and found that the best improvement in speech-recognition accuracy is provided by the rate-level nonlinearity stage that most models of the peripheral auditory system include just after the (typically linear) bandpass filtering that models the motion of the basilar membrane in the cochlea. This nonlinearity is roughly S-shaped, and has three major regions: (1) a range of input intensities that are “below threshold” in which the function output is roughly constant at a low level, (2) a range of input intensities for which the function output is roughly linear with respect to the input intensity in decibels, and (3) a “saturated” region in which the function output is roughly constant at a higher level.

Results from recent physiological studies describe and attempt to explain various types of dynamic adaptation of the rate-level functions with respect to the intensity of the incoming sound, background noise intensity, and the contrast between noise and the degraded speech signal (Dean et al., 2005; Zilany and Carney, 2010). These adaptations enable the dynamic range of the rate-level functions, which intrinsically is rather limited, to cover a much broader range of sound levels. In general, higher input sound levels tend to move the rate-level curves to the right and increase their maximum slope (Bureš et al., 2010; Gao et al., 2009). For example, in cats the background noise causes in the rate-level functions a shift of the dynamic range to higher intensities. It has also been noted that the noise level where this shift begins can be frequency dependent (Costalupes et al., 1984), and that the slope of the rate-level functions can increase in the presence of noise (May and Sachs, 1992) in addition to increased input levels.

Similar research in ferrets has characterized enhancement of spectro-temporal contrast in the acoustic environment as another important consequence of the adaptation of the sigmoidal nonlinearity (Rabinowitz et al., 2011; Wang and Shamma, 1994). This is similar to the enhancement in spatio-temporal contrast that is developed in the vertebrate retina (Ohzawa et al., 1985; Werblin et al., 1996). As an example, Rabinowitz et al. (2011), describe auditory processing that enhances local fluctuations in the envelope of response to desired signals in the presence of noise. This cannot be accomplished by a simple gain control which simultaneously amplifies both the degraded speech and the noise components, but rather a form of adjustable nonlinear gain control corresponds that increases the dynamic range of the degraded speech while suppressing the fluctuations produced by the noise (Schneider et al., 2011).

A steepening of rate-level functions of auditory neurons has also been observed in response to increment in sound level (Kang et al., 2010; Middlebrooks, 2004; Pfingst et al., 2011) and in response to noise (Bureš et al., 2010; Gao et al., 2009). According to Garcia-Lazaro et al. (2007) who investigated rat auditory neurons, the observed rate-level curves were more or less sigmoidal in shape, with a change in the steepness of the rate-level function interpreted to be a change in the “neural response gain”. Reports on auditory neurons of marmoset monkeys have shown that the slope in the rate-level function is a measure of the sound level discriminability. A steeper slope would allow greater discrimination of sound level (Watkins and Barbour, 2011). Other studies on auditory neurons in response to continuous, dynamic sound stimuli have shown a horizontal displacement of the rate-level function relocating the dynamic region of the function toward the mean sound level, resulting in higher coding precision of the levels (Dean et al., 2005; Wen et al., 2009), (Miller et al., 2011; Schneider et al., 2011). By expanding or compressing the auditory response to incoming sound in varying degrees, contrast gain control in human audition can serve two functions: (1) to protect sensory systems from overload and (2) to enhance discriminability among selected stimuli (Schneider et al., 2011). Ideally, the rate-level function would increase its slope to correspondingly enhance contrast in its response to small amplitude of the degraded speech signal above the noise (low contrast between degraded speech and noise signals). These goals, in combination with reducing the nonlinear distortion of the degraded speech and reducing the differences between the original clean speech and the degraded speech described in Chiu et al. (2012), lead to the definition of the four optimization criteria described in Section 2.2.

With these physiological examples in mind, Chiu et al. (2012) postulated that dynamic adaptation of the rate-level nonlinearity could also improve speech recognition accuracy for speech in the presence of noise. In particular, they modeled the rate-level nonlinearity by a set of frequency-dependent logistic functions, and developed a procedure that optimized the parameters that specified the form of the sigmoidal nonlinearity for a particular additive-noise environment using an objective function based on maximizing phonemic discriminability. These authors demonstrated that the use of an adapted nonlinear rate-level function reduces differences between the shapes of spectral distributions of clean speech versus speech in noise, and they showed that the adaptation of the nonlinearity improves speech recognition accuracy in the presence of noise.

In this paper, we describe a new approach for optimizing the sigmoidal rate-level function that is based on physical attributes of the acoustical signal, rather than the phoneme discrimination that was the basis for the approach of Chiu et al. (2012). The method attempts to discriminate between the degraded speech signal and noise, preserve maximum information in the linear region of the sigmoidal curve, and minimize the effects of distortions in the saturation regions. The proposed method is applied to a text-independent speaker verification task with speech signals that are corrupted by additive noise at different SNRs.

The development of adaptation based on signal analysis (rather than phonetic analysis) is motivated by several considerations. First, and foremost, the discriminative training used by Chiu et al. (2012) is based on speech recognition at the phoneme level in order to develop the ground-truth phoneme representation of the data. We believe that parameter optimization based on speech recognition may not be best for speech tasks other than speech recognition, such as the speaker verification considered in the present paper. We also had described above the existence of adaptation of sigmoidal nonlinearities in several nonhuman species and in other sensory modalities that is similar to those modeled in the present paper. This suggests that we should search for a viable approach to adaptation of the nonlinearity that is based on something other than human phonetic discrimination.

From the computational standpoint, the discriminative training described in Chiu et al. (2012) requires a substantial amount of *a priori* information, and increases the computation needed substantially compared to the signal-processing-based approach described in this paper. Similarly, the signal-processing-based approach is much more amenable to an online or adaptive implementation than the approach of Chiu et al., in which the discriminative training must be performed offline based on training data.

An unrelated reason for revisiting the issue of adapting the sigmoidal nonlinearity is that Chiu et al. perform their computations for all the channels of frequency analysis using the same parameters. We examine in this paper the extent to which performance would be further improved by adaptation that is allowed to vary on a channel-by-channel basis, which seems reasonable as the effective SNR varies from channel to channel.

We reiterate that in principle the approach proposed in this paper is applicable to *any* speech processing task because all analysis takes place at the level of the acoustic signal. Also, the sigmoidal functions are estimated separated for each channel.

In Section 2 we describe our optimization approach and specifically the development of four criteria that optimize the sigmoidal rate-level function based on acoustic attributes of the incoming signals. In Section 3 we discuss the actual implementation of the sigmoidal rate-level nonlinearity, and in Section 4 we describe the experimental results that validate the utility of the approach.

2. Development of the optimization criteria for the sigmoidal function

In this section we describe the development of the optimization criteria for the sigmoidal rate-level function. We begin with a mathematical specification of the sigmoidal nonlinearity, and subsequently we provide a mathematical description of the four components of the objective function used to optimize the rate-level nonlinearity. We remind the reader that the goal of the adaptation is to modify the location and the slope of the sigmoidal nonlinearity so that it is best able to capture the intensity fluctuations of the speech components of the signal in each channel, and to enhance the contrast when the input speech incurs a high level of degradation from additive noise.

2.1. Mathematical specification of the sigmoidal function

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

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

where μ and ω correspond to the offset and the slope of $g(l)$, respectively. This function allows modeling the nonlinear response. The offset parameter μ corresponds to the location along the horizontal axis at which the sigmoidal curve $g(l)$ equals 1/2. The slope of $g(l)$ equals $-\omega/4$ when $l = \mu$. Consequently, the position μ and the slope ω of the sigmoidal function are the parameters to be estimated.

Let us now consider the output of a particular channel of the initial bandpass filter bank that is the first stage of every model. We represent the degraded input speech signal $x_{j,k}$ at the output of filter j at the discrete-time index k by:

$$x_{j,k} = s_{j,k} + n_{j,k} \tag{2}$$

where $s_{j,k}$ and $n_{j,k}$ denote the clean speech and noise signals, respectively. If the entire signal $x_{j,k}$ is divided into N_f frames of W samples per frame with 50% overlap, the log-energy at frame i at filter j, $E_{j,i}$, can be written as:

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

where w_k represents the response of the finite-duration window function. Histograms of the log-energies $E_{j,i}$ at filter j and at frame i are generated in order to discriminate between noise and degraded speech frames by using the voice activity detector (VAD) proposed by Shin et al. (2008), as discussed in Section 3. Hence the frames are divided into two subsets, one believed to contain degraded speech and the second subset representing frames that are assumed to contain only noise.We use the symbols N_f^{sn} and N_f^n, where $N_f = N_f^{sn} + N_f^n$, to indicate the number of frames that are assumed to contain speech degraded by noise and the number of frames that are assumed to contain noise alone, respectively. Finally, we use the symbols $E_{j,i}^x$ $(1 \leqslant i \leqslant N_f)$, $E_{j,m}^{sn}$ $\left(1 \leqslant m \leqslant N_f^{sn}\right)$ and $E_{j,r}^n$ $\left(1 \leqslant r \leqslant N_f^n\right)$ to represent the energies at filter j and at frames 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. (Recall that each frame of the input is classified as containing either degraded speech or pure noise.) In addition, the mean and variance of the energy in the degraded-speech frames are defined to be $\mu_{j,sn}$ and $\sigma_{j,sn}^2$, respectively, while the corresponding mean and variances of the energy in the frames that are assumed to contain only noise energy frames are $\mu_{j,n}$ and $\sigma_{j,n}^2$, respectively.

2.2. Specification of the objective function used to optimize the sigmoidal nonlinearity

Based on the discussion above, we choose an objective function for the sigmoidal nonlinearity that (1) minimizes nonlinear distortion in the linear region, (2) minimizes noise power, (3) maximizes the similarity between energy in the frames that are believed to represent degraded speech and the 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.

2.2.1. Criterion 1: Nonlinear distortion in the linear region

The slope ω and the position μ of the sigmoidal function should be chosen in such a way that the degraded speech lies in the linear part of the sigmoidal curve. Therefore, once the sigmoidal function is applied, the nonlinear distortion in the degraded speech would be minimized. This nonlinear distortion, $D_j^{non-linear}(\omega_j, \mu_j)$, is defined as:

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

where (as before) $E_{j,m}^{sn}$ refers to the energy of frames of degraded speech at frame index m for filter index j, $g(\cdot)$ represents the sigmoidal function and $\mathbf{E}[\cdot]$ is the expectation operator. The parameters A_j and B_j correspond to a linear transformation that allows the comparison of $E_{j,m}^{sn}$ and $g\left(E_{j,m}^{sn}\right)$ (as developed in the Appendix). By approximating the expected value by the sample mean, $D_j^{non-linear}(\omega_j, \mu_j)$ can be rewritten as:

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

where N_f^{sn} is the total number of frames containing degraded speech.

2.2.2. Criterion 2: Noise power

The sigmoidal function can be employed to attenuate the noise in the speech signal due to the fact that low-energy frames can be associated with noise. The power of the noise, after it is passed through the sigmoidal function, $P_j^{noise}(\omega_j, \mu_j)$, is given by:

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

where $E_{j,r}^n$ corresponds to energy of noise frames at frame r for filter j, $g(\cdot)$ represents the sigmoidal function, and $\mathbf{E}[\cdot]$ is the expectation operator. The sigmoidal function 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:

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

where N_f^n is the number frames that are assumed to contain noise only.

2.2.3. Criterion 3: Similarity between clean speech and the degraded speech input

According to Chiu et al. (2012), the use of a nonlinear rate-level function should reduce the differences between the average frequency response of clean speech and the average frequency response of the degraded input signal, both assessed after the sigmoidal nonlinearity. Consequently, the difference between the energy of the clean speech and the degraded speech input is represented by:

$$D_j^{clean-noise}(\omega_j,\mu_j)=\sum_{i=1}^{N_f}\left[g\left(E_{j,i}^{s}\right)-g\left(E_{j,i}^{x}\right)\right]^2 \tag{8}$$

where $E_{j,i}^{s}$ and $E_{j,i}^{x}$ correspond to the energy of clean speech and the energy of the degraded input speech, respectively, at frame i for filter j, and $g(\cdot)$ is the sigmoidal function.

2.2.4. Criterion 4: Signal variance of speech degraded by noise 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 of the degraded speech, $V_j(\omega_j,\mu_j)$, is expressed as:

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

where $E_{j,m}^{sn}$ is energy of the frames of degraded speech at frame m at filter j and $g(\cdot)$ is the sigmoidal function. By expanding the expression of the variance, $V_j(\omega_j,\mu_j)$ can be rewritten as:

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

where N_f^{sn} is the number of frames containing degraded speech frames.

2.2.5. Specification of the complete objective function

Based on the four criteria described above, we adopt for this study the objective function $J(\omega_j,\mu_j)$ that is defined as:

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

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

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

In (12), the position $\mu_j$of the sigmoidal function is set to $\mu_j=\mathbf{E}\left[\left(E_{j,m}^{sn}\right)\right]$ (*i.e.* centered on the mean of the energy of the degraded speech frames $E_{j,m}^{sn}$).

Finally, the optimal position, $\hat{\mu}_j$, of the sigmoidal function is estimated according to:

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

In (13), ω_j corresponds to the optimal sigmoidal slope $\hat{\omega}_j$.

While we recognize that the definition of the objective function $J(\omega_j,\mu_j)$ as the simple sum of the four criteria above is a special case of the more general linear combination

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

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. Implementation of the sigmoidal rate-level function

In this section we describe the adaptive procedure based on signal analysis that is used to optimize the sigmoidal rate-level function. We refer the reader to Fig. 1 for a depiction of the complete feature extraction scheme, and Fig. 2 for a depiction of the procedure for obtaining the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$. The specific values of the sigmoidal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$, as defined in Eqs. (12) and (13), respectively, were determined using a development database of speech corrupted by babble noise at an SNR equal to 10 dB, as discussed in Section 4.

The optimal values of the parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ used in our work vary from channel to channel, in contrast to the approach of Chiu et al. (2012), in which the parameters of the nonlinearity are the same for all the filters. This is helpful because the SNR varies from one filter to the other. As mentioned above, we used the voice activity detector (VAD) proposed by Shin et al. (2008) to discriminate between degraded speech and noise. Two subsets of frames are defined based on the VAD results, representing frames that contain degraded speech, and the representing frames that are assumed to contain only noise, respectively.

Fig. 3 describes the dependence of the shape of the objective function on the parameters ω_j and μ_j, which describe the slope and position, respectively, for each the 35 analysis bands j. Fig. 4 depicts a representative example of an optimal sigmoidal function (solid line) and a corresponding linear mapping (dotted line) with a slope equal to the sigmoid at its center point. The sigmoidal curve

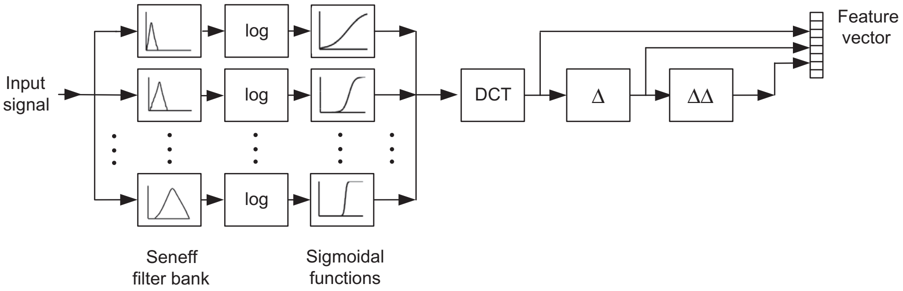

Fig. 1. Block diagram of the proposed feature extraction scheme.

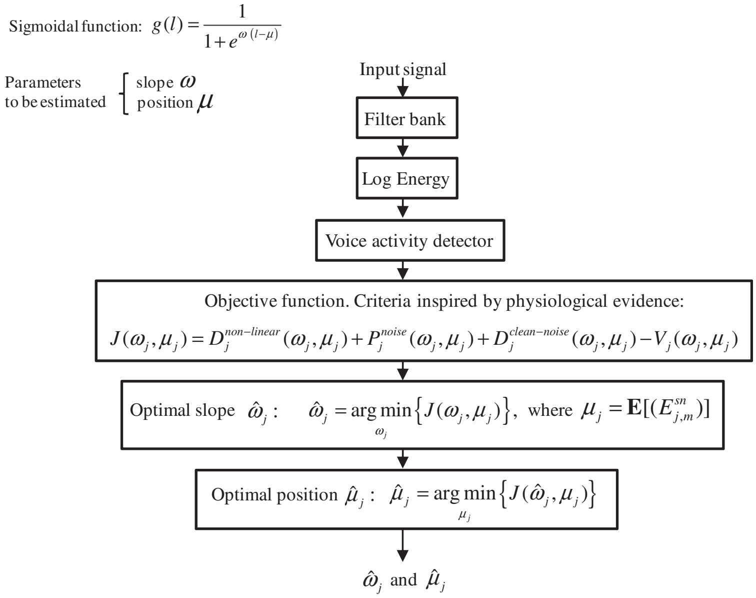
Sigmoidal function: $g(l) = \frac{1}{1+e^{\omega(l-\mu)}}$
Parameters to be estimated
slope ω
position μ
Input signal
Filter bank
Log Energy
Voice activity detector
Objective function. Criteria inspired by physiological evidence:
$J(\omega_j,\mu_j) = D_j^{non-linear}(\omega_j,\mu_j) + P_j^{noise}(\omega_j,\mu_j) + D_j^{clean-noise}(\omega_j,\mu_j) - V_j(\omega_j,\mu_j)$
Optimal slope $\hat{\omega}_j$: $\hat{\omega}_j = \arg\min_{\omega_j}\{J(\omega_j,\mu_j)\}$, where $\mu_j = \mathbf{E}[(E_{j,m}^{sn})]$
Optimal position $\hat{\mu}_j$: $\hat{\mu}_j = \arg\min_{\mu_j}\{J(\hat{\omega}_j,\mu_j)\}$
$\hat{\omega}_j$ and $\hat{\mu}_j$

Fig. 2. Block diagram for obtaining optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ of the sigmoidal function.

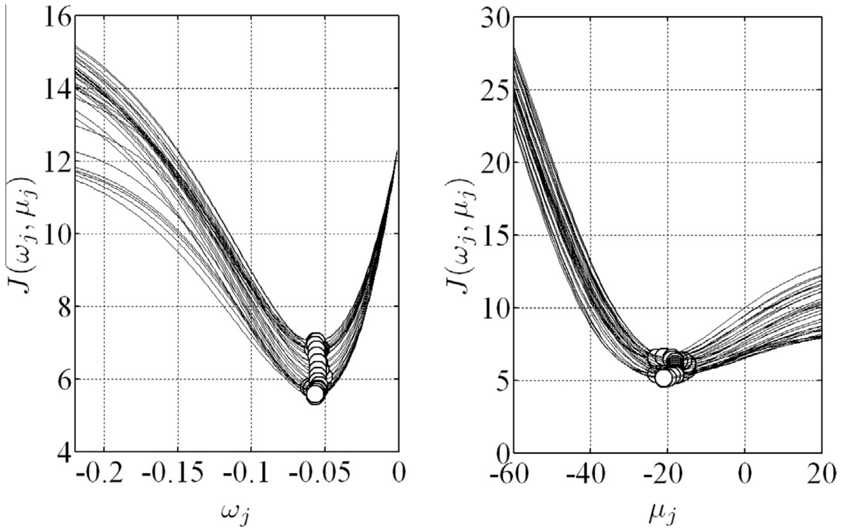

Fig. 3. The objective function $J(\omega_j, \mu_j)$ plotted as a function of the sigmoidal function parameters: (a) sigmoidal slope ω_j and (b) sigmoidal position μ_j. The optimal values of $\hat{\omega}_j$ and $\hat{\mu}_j$ are indicated by the open circles for each of the 35 channels of the filter bank.

was obtained with the optimal slope $\hat{\omega}_j$ and position $\hat{\mu}_j$ of the sigmoidal function for the filter $j = 8$. Fig. 4 also depicts the histograms extracted from a testing utterance for filter $j = 8$ with babble noise at SNR equal to 10 dB. By comparing the solid and dotted lines in Fig. 4 it can be seen that the sigmoidal function compresses the noise in the nonlinearity region, while most of the frames containing degraded speech lie within the linear part of the sigmoidal function.

Fig. 5 shows four 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. 5, both optimal slope $\hat{\omega}_j$ and position $\hat{\mu}_j$ of the sigmoidal function depend on the SNR at which the sigmoidal function was trained: as SNR is increased, the curves in Fig. 5 shift to the right and become steeper. Consequently, the optimization of sigmoidal slope $\hat{\omega}_j$ and the sigmoidal position $\hat{\mu}_j$ provides an adaptation in the sigmoidal function that compensates for variations in SNR.

Fig. 6 is a composite of all of the sigmoidal rate-level functions, plotted as a function of SNR with optimal parameters for all 35 channels, trained on speech degraded with babble noise. We observe that the sigmoidal functions adapt slightly for each channel at each SNR. Specifically, as the SNR decreases, the curves are displaced toward the right, and their slopes become steeper at the midpoints. Collectively these phenomena modify the nonlinearities to ensure that most of the speech energy falls on the relatively

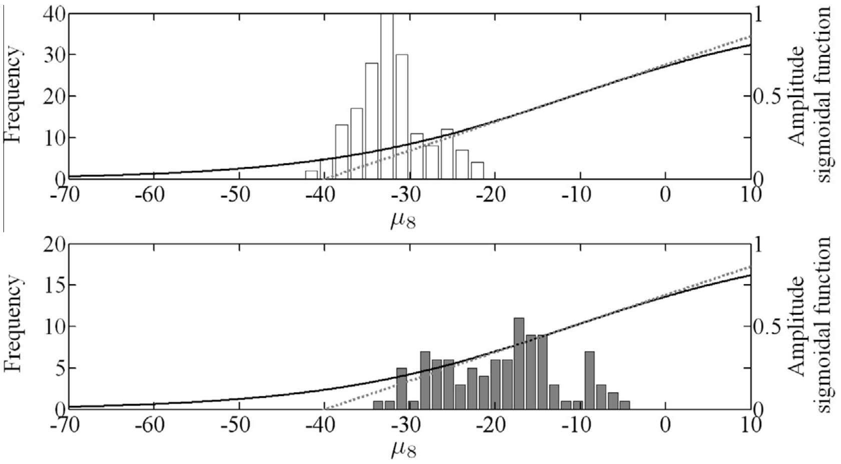

Fig. 4. Example of optimal sigmoidal function (solid line) and the corresponding linear mapping (dotted line). The training conditions for the sigmoid were babble noise at SNR = 10 dB. Results for filter $j = 8$ are plotted with optimal parameters: $\hat{\omega}_8 = -0.071$ and $\hat{\mu}_8 = -14$. In addition, histograms of power are depicted for frames containing degraded speech (filled bars) and frames assumed to contain noise only (open bars).

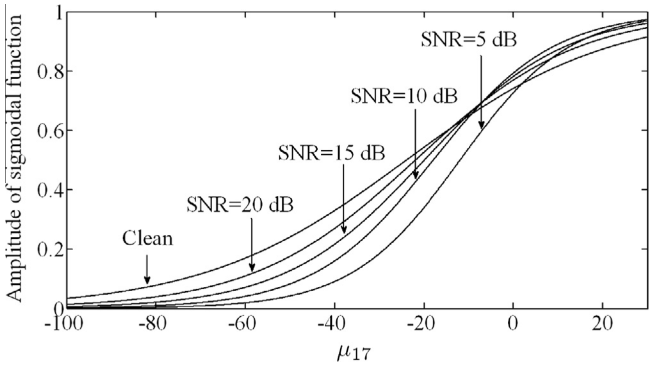

Fig. 5. Sigmoidal functions plotted as a function of SNR. Training conditions: clean speech and speech degraded by babble noise at SNRs equal to 20 dB, 15 dB, 10 dB and 5 dB. Results for filter $j = 17$ are plotted with optimal parameters.

linear part of the curve, which renders the features more robust against changes in SNR when training and testing conditions are mismatched.

Fig. 7 compares an ensemble of sigmoidal rate-level functions that were trained with different types of noise (restaurant and car noise) but at the same SNR. Results are similar to the curves of Fig. 6 in that the curves shift to the right and their slopes increase as SNR decreases, with variations in the individual responses observed from filter to filter. The interlaced patterns generated by the sigmoidal functions trained on the two types of noise indicate that the shape of the optimal non-linearity depends on the spectral distribution of the masking noise.

This dependence of the location and steepness of the sigmoidal curves in Fig. 5, Fig. 6 and Fig. 7, is consistent with the experimental results in the physiological literature described above. As noted, the steepening of rate-level functions of auditory neurons is consistent with the results of numerous physiological studies describing nonlinearity in sensory transduction (*e.g.* Bureš et al., 2010; Gao et al., 2009; Garcia-Lazaro et al., 2007; Kang et al., 2010; Middlebrooks, 2004; Pfingst et al., 2011; Watkins and Barbour, 2011).

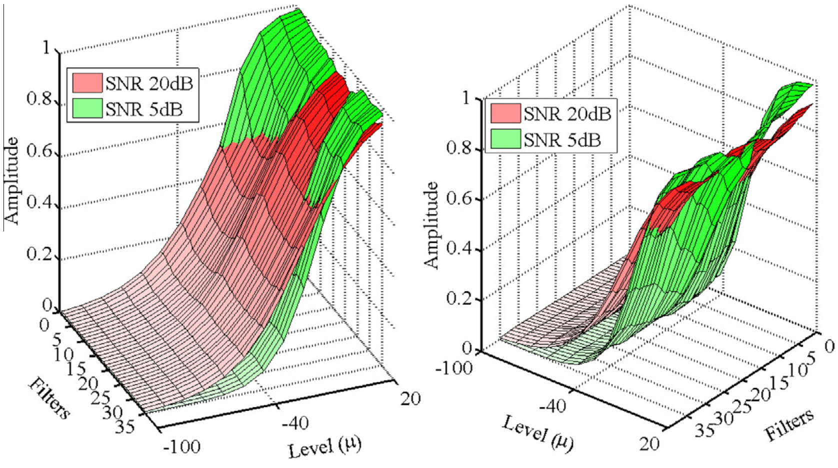

Fig. 6. Three-dimensional graphs of the sigmoidal rate-level functions trained with speech degraded by babble noise at SNR equal to 20 dB and 5 dB. The plot is rotated to show the difference in slope and horizontal displacement between both set of functions.

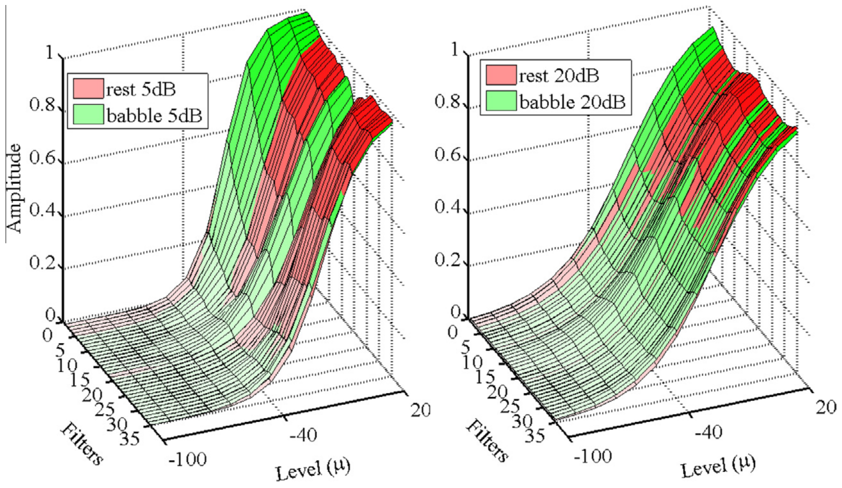

Fig. 7. Comparison of optimal sigmoidal rate-level functions trained with restaurant and car noise at SNR equal to 20 dB (right) and 5 dB (left).

We note that in this paper we describe separately a stage of logarithmic compression followed by the sigmoidal nonlinearity. We consider the logarithmic compression to be an approximation of the compressive gain of the basilar membrane (Robles and Ruggero, 2011). The subsequent sigmoidal rate-level function is intended to represent an approximation to both the existence of spontaneous AN activity at low input sound levels and the saturation of AN fibers at high sound levels, presumably because of a mixture of suppression (such as two-tone suppression (Rhode and Cooper, 1993), and a presumed limit on the discharge rate with which a fiber is capable of coding intensity.

4. Experimental results

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 presented here were obtained using the YOHO database (Campbell and Higgins, 1994), which supports the development, training, and testing of speaker verification systems. The vocabulary is composed of two-digit numbers spoken continuously in sets of three (*e.g.*, “62-31-53” is pronounced as “sixty-two thirty-one fifty-three”). The database is divided into “enrollment” and “verification” segments. Each segment contains data from 138 speakers. In this paper a subset of 70 speakers (43 males and 27 females) was employed. These speakers were divided as follows: 40 background impostor speakers (28 males and 12 females) to train the background models; and 30 testing speakers (15 males and 15 females) were used in verification attempts. For each speaker, one 96-utterance enrollment session was considered. False rejection curves were estimated with 30 speakers × 40 verification signals per client = 1200 utterances. False acceptance curves were obtained with 30 speakers × 29 impostors × 12 verification signals/per impostor = 10,440 experiments. In addition, a subset composed of 50 speakers and one utterance per speaker (development database) extracted from YOHO was employed to train the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ of the sigmoidal functions. The utterances used to train the sigmoidal function were not included in the testing data for the main speaker verification experiment. Three types of noise (babble, car, and restaurant) were selected from the AURORA database (Hirsch and Pearce, 2000). These noises were artificially added to the YOHO corpus to generate noisy versions of the utterances at various SNRs: 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. For all the speaker verification experiments, the system was trained with clean speech.

In this paper, the auditory filter bank (Stage I in the Seneff auditory model (Seneff, 1988) was obtained directly from Malcolm Slaney’s widely-used Auditory Toolbox (Slaney, 1998), which implements 35 filters with center frequencies spaced according to the Bark scale from 200 to 3300 Hz. Each filter was redesigned by reducing the sampling frequency to 8 kHz. Finally, the input signal was normalized by dividing the samples by the maximum absolute amplitude. 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, in each frame, a channel-specific optimal sigmoidal function, estimated using the development data set and the procedure explained in Section 2, was applied to the log-energy of the output of each filter, both in the training and testing data sets. Finally, the log-energy plus ten static cepstral coefficients, and their first and second cepstral time derivatives were estimated in a fashion that is similar to MFCC processing (see Fig. 1). Four configurations were considered: (1) a baseline system, which corresponds to the log-energies of the Seneff filter bank output; (2) the baseline system with cepstral variance normalization (CVN); (3) the baseline system with cepstral mean and variance normalization (CMVN), and (4) the method proposed in this paper using the optimal sigmoidal function, combined with CVN. If the entire signal were mapped into the linear region of the sigmoidal function, the proposed scheme could be considered equivalent to the CVN algorithm. Therefore, the impact of the nonlinearity provided by the sigmoidal function may be inferred by comparing results obtained with the configurations (2) and (4) as described above.

In the verification procedure, the normalized log likelihood is estimated. Given a verification attempt in which the identity of Speaker s is claimed, O denotes the observation sequence corresponding to the claimant's utterance. The output score of the system is a cohort-normalized log likelihood, log $L(O)$:

$$logL(O) = logL(O/\lambda_s) - \overline{logL(O\lambda_{\bar{s}})} \tag{14}$$

where log $L(O/\lambda_s)$ is the log likelihood of the client hypothesis and λ_s is the speaker s model, and $\overline{logL(O/\lambda_{\bar{s}})}$ is the averaged log likelihood of the cohort of impostor models. A 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 (Reynolds et al., 2000). By doing so, the correspondence of the Gaussians within each speaker-dependent GMM with those in the background GMM is preserved (Reynolds et al., 2000).

4.1. General dependence on SNR and the presence of the sigmoidal nonlinearity

Fig. 8 describes results provided using the optimal sigmoidal functions for the speaker verification task in the presence of three types of background noise: speech babble, car noise, and restaurant noise, all as a joint function of the SNR at which the sigmoidal functions were trained and the SNR of the incoming speech. The optimal sigmoidal functions were trained with babble noise and SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. As can be seen in Fig. 8, the lowest EERs are achieved when the sigmoidal function is trained with 10 dB for all testing conditions. We note that these results are consistent with similar findings observed by Chiu et al. (2012) where the optimal sigmoidal function was trained at an SNR of 10 dB by using a criterion based on phoneme discrimination. In contrast, the objective function $J(\omega_j, \mu_j)$ is based entirely on the physical characteristics (and especially the power distribution) of the incoming speech, and does not take phonetic content into account at all. Consequently, the fact that the best speaker verification results are achieved with the sigmoidal function trained with signals at SNR 10 dB means only that for these SNRs the benefits provided by noise suppression are more significant than the distortion introduced by saturation at higher levels. Most of the results we described below are carried out using sigmoidal functions trained at an SNR equal to 10 dB.

We also note that the sigmoidal function trained at 0-dB SNR provides poor performance in speaker verification of speech at all testing SNRs. This is most likely a consequence of the fact that at 0-dB SNR the speech and noise distributions are nominally overlapping. Hence, the speech-plus-noise and noise-alone distributions are not separable by SNR, and any nonlinear compression applied to the noise will be applied to the speech as well. The estimation of the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ for the sigmoidal distribution is less reliable as well.

Fig. 9 describes EER results obtained as a function of SNR for speech in the presence of three types of background interference: speech babble, car noise, and restaurant noise. Results are compared for the baseline system, the baseline system combined with CVN, the baseline

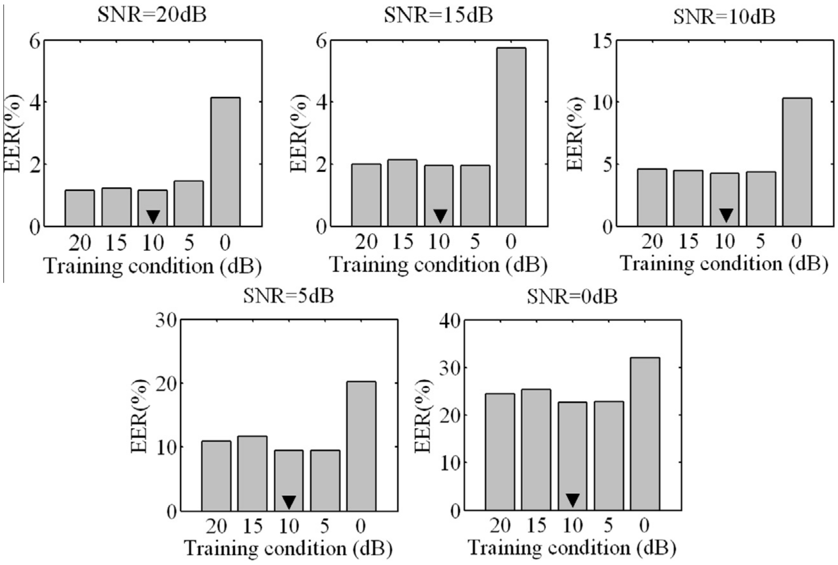

Fig. 8. EER for speaker verification as a function of the SNR of the testing data and the SNR used to develop the parameters for the sigmoidal function. The data obtained from each SNR used for testing are described in a single graphic, with the testing SNR indicated at the top. The optimal sigmoidal functions were trained with babble noise and SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB, as indicated by the scale at the bottom of each panel.

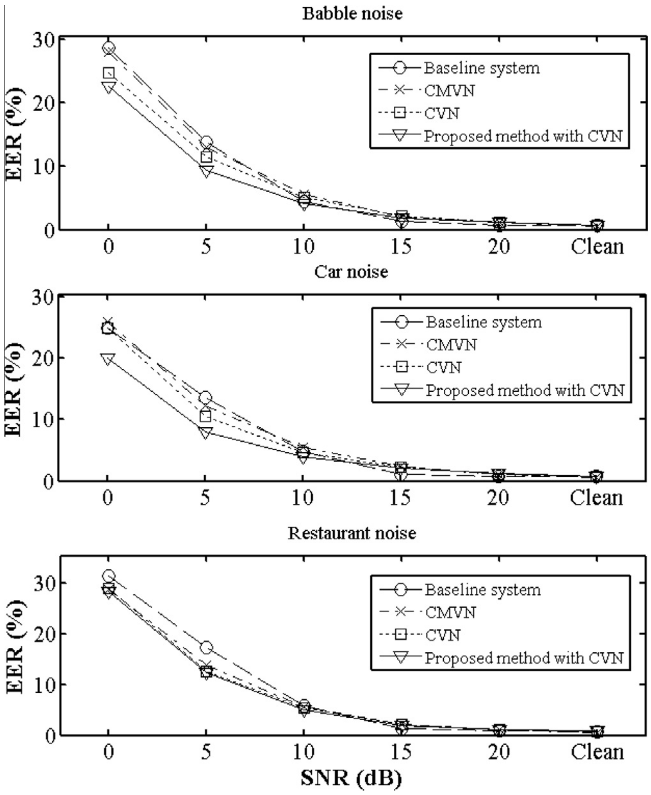

Fig. 9. Comparison of EER as a function of SNR for speech in babble, car and restaurant noise, respectively. Depicted separately are results for the baseline system, the baseline system with CMVN, the baseline system with CVN; and the system using the optimal sigmoidal function combined with CVN. The optimal sigmoidal functions were trained and tested in matched noisy condition at an SNR of 10 dB.

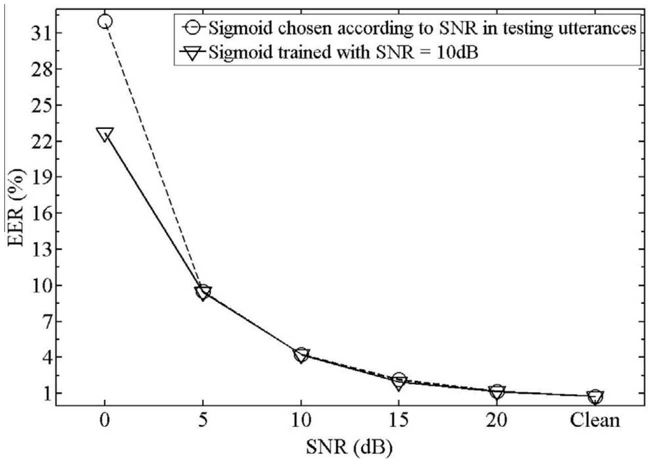

Fig. 10. Comparison of EER as a function of SNR for speech in babble noise using the sigmoidal function combined with CVN. In one curve the sigmoid trained with SNR equal to 10 dB was applied to both training and testing utterances. In the second curve, training and testing utterances were processed with the sigmoid trained at the same SNR used in the utterance.

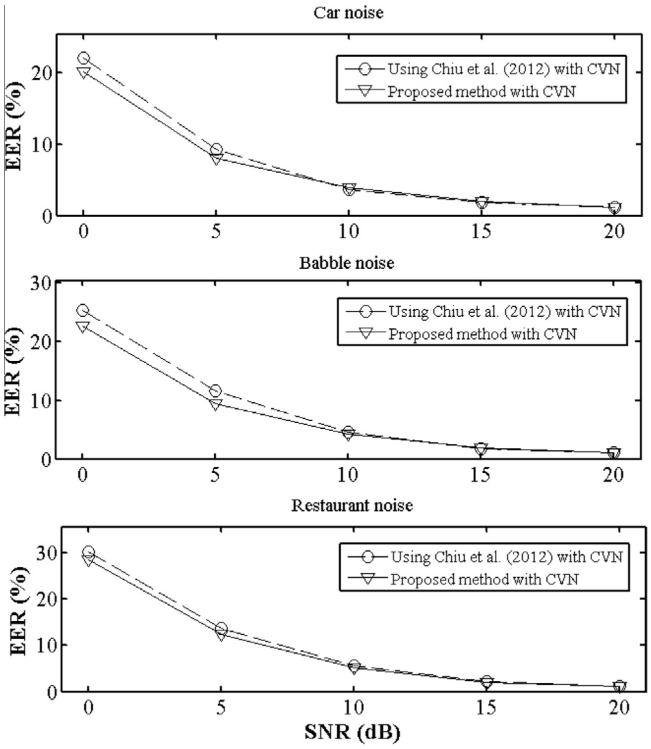

Fig. 11. Comparison of EER (%) obtained using the method with parameters of the nonlinearity given by Chiu et al. (2012), combined with CVN, and the proposed method in this paper combined with CVN, for speech in car, babble and restaurant noise, respectively.

system with cepstral mean and variance normalization (CMVN) and the proposed method combining the proposed optimal sigmoidal nonlinearity and CVN, as described above. The use of the optimal sigmoidal function in combination with CVN improves the effective SNR of the system in all three types of noise, typically on the order of 1–2 dB. Maximum relative percentage improvements in SNR compared to baseline at selected SNRs are approximately 31.7%, 40.6%, and 28.4% for the three types of background noise. Best performance is always obtained using the proposed optimal sigmoidal nonlinearity, but at least for car noise, the performance of a system with CVN only comes close.

4.2. Comparison of training the nonlinearity parameters at fixed versus matched SNRs

As noted above, the results depicted in Fig. 9 were obtained by estimating the parameters characterizing the sigmoidal functions using speech in the presence of speech babble at an SNR of +10 dB, which appears to the best single training SNR according to the results depicted in Fig. 8. Nevertheless, we would always expect better performance to be obtained when the parameters characterizing the sigmoidal functions are estimated in environmental conditions that match the testing environment. In an effort to quantify the magnitude of the improvement to be expected, we repeated some of the conditions with the SNRs for parameter estimation matched to the SNRs used in the speaker verification experiment itself. Fig. 10 compares EERs for speaker verification when the sigmoids are trained at the testing SNR with the corresponding EERs obtained when the parameters are always estimated from signals at 10-dB SNR. As can be seen from the figure, very little difference in results is observed, except for the lowest SNR, 0 dB, which actually provides very poor performance when the parameters are estimated at a matched SNR (presumably because the data are too noisy to provide reliable parameter estimation). For this reason we continue to use sigmoids trained at an SNR of 10 dB for all utterances, because they provide similar performance to sigmoids trained to match the testing data at most SNRs, and substantially better performance at 0-dB SNR.

4.3. Comparison to the results of Chiu et al.

As noted above, Chiu et al. (2012) described a method of adapting the sigmoidal rate-level function using a criterion based on discrimination analysis at phonetic level. Fig. 11 compares results obtained using the method described in this paper with results obtained using the method described by Chiu et al., with CVN included in obtaining both sets of results. The parameters obtained for the sigmoidal functions of Chiu et al. (2012) were $\alpha = 0.05$; $\omega_0 = 0.613$; and $\omega_1 = 0.521$. Results are presented for three types of noise: babble, car, and restaurant at SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB, respectively. The experimental results shown in Fig. 11 indicate that both the method proposed in this paper and the method proposed by Chiu et al. are effective in maintaining good performance at most SNRs, but the method proposed in the present paper performs somewhat better for all three noise types at the lower SNRs.

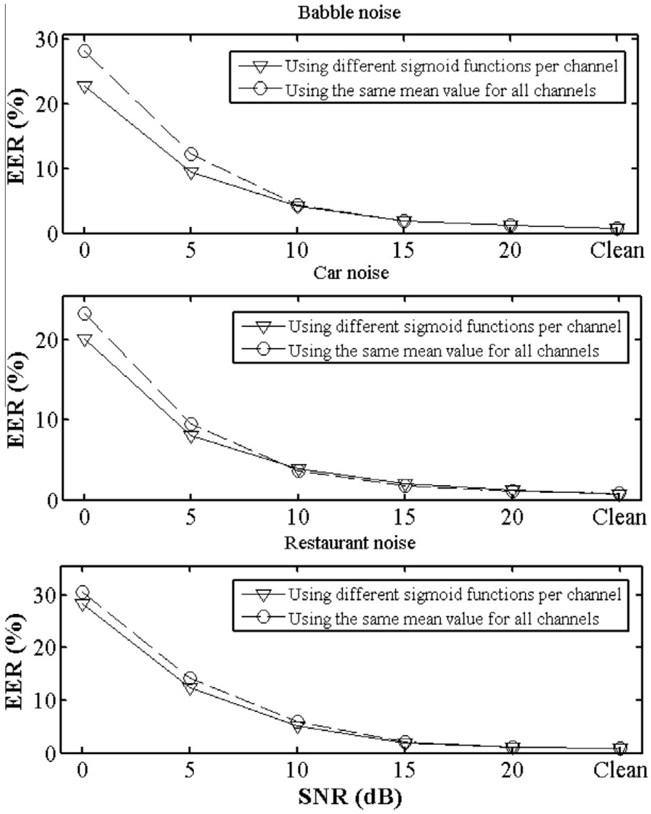

Fig. 12. Comparison of EER (%) obtained using the proposed method in this paper combined with CVN, using different sigmoidal functions per channel and the same mean value for all channels, for speech in car, babble and restaurant noise, respectively.

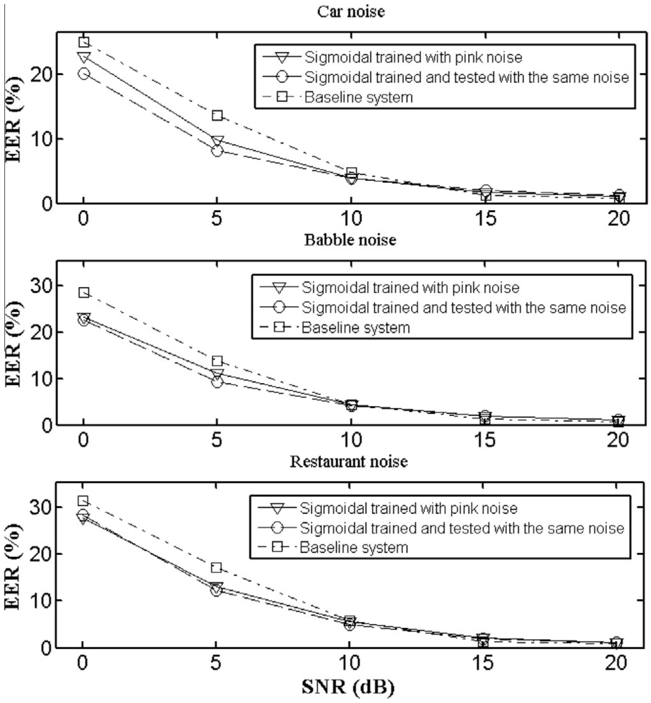

Fig. 13. Comparison of EER (%) obtained using the proposed method with parameters of the nonlinearity trained (at SNR equal to 10 dB) and tested with the same type of noise, and the proposed optimization method with the parameters of the sigmoidal function trained with pink noise (at SNR equal to 10 dB). In both cases the speech was degraded with car, babble and restaurant noise, and the sigmoidal function was combined with CVN.

4.4. Impact of channel-specific estimation of sigmoidal nonlinearities

Fig. 12 compares results obtained using the sigmoidal nonlinearities estimated on a channel-specific basis as described in this paper with results obtained using a single nonlinearity for all 35 frequency channels. It can be seen that the allowing the sigmoidal nonlinearities to vary from channel to channel is advantageous at SNRs of 0 and +5 dB, most likely because the local SNRs exhibit greater variation from channel to channel at the lower SNRs. Thus, at lower SNRs, the performance of the optimization improves speaker verification accuracy, due to the fact that the adaptation enables to increase the dynamic range of the degraded speech above the noise and minimizes nonlinear distortions in the linear region while suppressing fluctuations produced by noise.

4.5. Comparisons to optimal sigmoidal functions trained and tested with different type of noise

Fig. 13 compares results obtained by using the sigmoidal nonlinearities estimated with pink noise at SNR equal to 10 dB with results shown in Fig. 9 where the same kind of noise was employed in training and testing. When compared to baseline processing, the sigmoidal functions trained with pink noise in combination with CVN leads to average relative reductions in EER equal to 23.5% and 13% at SNR equal to 5 dB and 0 dB, respectively, with car, babble and restaurant noise. This result strongly validates the proposed optimization method. However, the highest reductions in EER are obtained when the sigmoidal nonlinearities are trained and tested with the same noise, except with restaurant noise at SNR equal to 0 dB where both sigmoidal functions provide almost the same result.

4.6. General comments

The improvements provided by the use of the sigmoidal function are consistent with results from other studies in speech recognition based on auditory processing. Auditory-based approaches typically provide significant improvements at lower SNRs, but at higher SNRs they may achieve performance that is no better than (or worse than) the performance that is observed using conventional signal processing based on MFCC or PLP features (e.g. (Ghitza, 1986), (Chiu and Stern, 2008; Chiu et al., 2012), (Jankowski and Lippmann, 1992; Kim et al., 1999). We reiterate that the results presented here are consistent with those described in Chiu et al. (2012) where the sigmoidally-shaped rate-intensity function has been identified as an important component of auditory-based feature extraction systems for speech recognition.

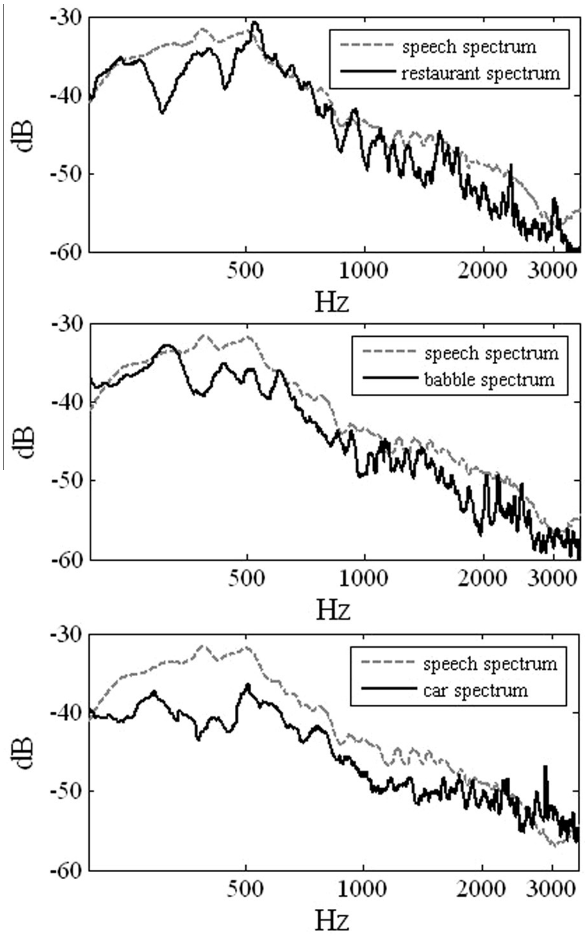

Fig. 14. Comparison of the average power spectrum for 50 utterances of clean speech with the power spectra of noise. The mean-squared errors (MSE) between the speech and restaurant, babble and car noise spectra are 58.7, 70.1, and 95.4, respectively.

Nevertheless, the use of the sigmoidal nonlinearity trained at 0-dB SNR failed to provide a significant improvement. As we have noted above, at 0-dB SNR the distributions of power of the frames containing degraded speech overlap with the power distributions of the noise-only frames. In addition, it appears that the power spectra of the speech signals and background noise are more similar in the case of restaurant noise than in the cases of the other two noise types considered. This is illustrated in Fig. 14, which shows the average power spectrum of 50 utterances extracted from clean speech and the power spectrum of restaurant, babble and car noises in each of the three panels. Spectra were estimated by using a FFT with 2^{15} points. An averaging filter was applied to smooth the speech and noise spectra. Finally, for comparison purposes, each spectrum was normalized according to its energy. We observed differences in mean-squared error (MSE) between the speech and noise spectra equal to 58.7, 70.1 and 95.4 for restaurant, babble and car noises, respectively. The corresponding differences between EERs obtained using the sigmoid nonlinearity combined with CVN compared with the use of CVN alone are 0.57%, 2.1%, and 4.9%, respectively, at SNR equal to 0 dB. Hence, we believe that the sigmoidal nonlinearity fails to improve EER for the speaker-verification task in restaurant noise at 0-dB SNR because the spectra of speech and noise are very similar, causing the power curves for degraded speech and noise to overlap at all frequencies.

5. Conclusions

This paper describes a method that can be used to develop an optimal sigmoidal nonlinear rectifier 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 noise power, minimize nonlinear distortion, maximize the similarity between clean speech and the degraded speech input, and maximize the signal variance of the speech degraded by noise after processing by sigmoidal function. The optimal sigmoidal functions obtained are frequency dependent because the output SNR from the channels of the initial bandpass filter bank varies from one channel to the other. Finally, we note the proposed approach differs from cepstral mean and variance normalization (CMVN), which in effect produces a linear function that relates input and output, similar to the linear approximations of Fig. 4. The observed improvements in speaker identification accuracy obtained using the optimal sigmoidal nonlinearity (compared to results obtain with CMVN) demonstrate the potential of the nonlinearities that are part of human auditory processing.

The resulting sigmoidal nonlinearities are demonstrated to exhibit a location and slope that change as a function of SNR in a fashion that is consistent with the corresponding dependencies that are described in the physiological literature. The utility of the optimal sigmoidal nonlinearities derived in this fashion is considered in a series of experiments measuring speaker verification accuracy using the YOHO database. Our results indicate that the use of a sigmoidal nonlinearity defined strictly from the physical characteristics of the input (as apposed to phoneme discrimination) can lead to average relative reductions in EER compared to baseline processing as great as 12%, 33.6% and 16.6% at SNR equal to 10 dB, 5 dB and 0 dB, respectively, with speech degraded by babble, car and restaurant noise. The sigmoidal nonlinearity provides smaller benefit at higher SNRs, consistent with previous experiments with auditory models in speech recognition. The consistency of results between the two optimization schemes (using discrimination based on phoneme classes and discrimination based on waveform characteristics), reinforces the notion that optimal sigmoidal functions can reduce the mismatch between the training conditions and testing. In principle, our results obtained appear to be generic suggesting that this optimization approach could be applicable to any image or sound recognition system in which the feature extraction employs a nonlinear function based on rate-level responses.

Acknowledgements

This research was funded by Conicyt-Chile under grants Fondecyt 1100195 and Team Research in Science and Technology ACT 1120.

Appendix A

In this Appendix we develop the parameters A_j and B_j of the nonlinear distortion factor $D_j^{non-linear}(\omega_j, \mu_j)$, which are defined in Section 2.3.

The parameters A_j and B_j are estimated according to:

$$(A_j, B_j) = \underset{A_j, B_j}{\arg\min}\left\{D_j^{non-linear}(\omega_j, \mu_j)\right\} \tag{A1}$$

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

$$\begin{aligned}\frac{\partial D_j^{non-linear}}{\partial A_j} = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2\cdot\left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right)\right] \\ \cdot E_{j,m}^{sn}\end{aligned} \tag{A2}$$

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

$$\begin{aligned}&\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2\cdot\left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right)\right] \cdot E_{j,m}^{sn} = 0 \\ &A_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn}\right)^2 + B_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} \\ &\quad = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} \cdot g\left(E_{j,m}^{sn}\right) \\ &A_j \cdot \mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right] + B_j \cdot \mathbf{E}\left[E_{j,m}^{sn}\right] = \mathbf{E}\left[E_{j,m}^{sn} \cdot g\left(E_{j,m}^{sn}\right)\right]\end{aligned} \tag{A3}$$

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:

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

By combining (A3) and (A4) and making use of the expressions $\mu_j = \mathbf{E}\left[E_{j,m}^{sn}\right]$ and $\sigma_j^2 = \mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right] - \left\{\mathbf{E}\left[E_{j,m}^{sn}\right]\right\}^2$, the parameters A_j and B_j are found to be equal to:

$$\begin{aligned}&A_j = \tfrac{1}{\sigma_j^2}\left\{\mathbf{E}\left[E_{j,m}^{sn} \cdot g(E_{j,m}^{sn})\right] - \mu_j \cdot \mathbf{E}\left[g\left(E_{j,m}^{sn}\right)\right]\right\} \\ &B_j = \mathbf{E}\left[g\left(E_{j,m}^{sn}\right)\right] - \mu_j \cdot A_j\end{aligned} \tag{A5}$$

References

Ajmera, P.K., Jadhav, D.V., Holambe, R.S., 2011. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition 44 (10–11), 2749–2759.

Allen, J.B., 1985. Cochlear modeling. IEEE ASSP Magazine 2 (1), 3–29.

Barbour, D.L., 2011. Intensity-invariant coding in the auditory system. Neuroscience and Biobehavioral Reviews 35 (10), 2064–2072.

Bureš, Z., Grécová, J., Popelář, J., Syka, J., 2010. Noise exposure during early development impairs the processing of sound intensity in adult rats. European Journal of Neuroscience 32 (1), 155–164.

Campbell, J., Higgins, A., 1994. YOHO speaker verification. Linguistic Data Consortium, Philadelphia, PA.

Chiu, Y.-H.B., Stern, R.M., 2008. Analysis of physiologically-motivated signal processing for robust speech recognition. In: Proceedings of Interspeech, Brisbane, Australia, pp. 1000–1003.

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

Cohen, J.R., 1989. Application of an auditory model to speech recognition. Journal of the Acoustical Society of America 85 (6), 2623–2629.

Costalupes, J.A., Young, E.D., Gibson, D.J., 1984. Effects of continuous noise backgrounds on rate response of auditory nerve fibers in cat. Journal of Neurophysiology 51 (6), 1326–1344.

Darwin, C.J., 2008. Listening to speech in the presence of other sounds. Philosophical Transactions of Royal Society B: Biological Science 363 (1493), 1011–1021.

Dean, I., Harper, N.S., McAlpine, D., 2005. Neural population coding of sound level adapts to stimulus statistics. Nature Neuroscience 8 (12), 1684–1689.

Dean, I., Robinson, B.L., Harper, N.S., McAlpine, D., 2008. Rapid neural adaptation to sound level statistics. Journal of Neuroscience 28 (25), 6430–6438.

Dimitriadis, D., Maragos, P., Potamianos, A., 2011. On the effects of filterbank design and energy computation on robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing 19 (6), 1504–1516.

Gao, F., Zhang, J., Sun, X., Chen, L., 2009. The effect of postnatal exposure to noise on sound level processing by auditory cortex neurons of rats in adulthood. Physiology & Behavior 97, 369–373.

Garcia-Lazaro, J.A., Ho, S.S., Fair, A., Schnupp, J.W., 2007. Shifting and scaling adaptation to dynamic stimuli in somatosensory cortex. European Journal of Neuroscience 26 (8), 2359–2368.

Ghitza, O., 1986. Auditory nerve representation as a front-end for speech recognition in a noisy environment. Computer Speech & Language 1 (2), 109–131.

Ghitza, O., 1994. 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.

Hanilçi, C., Kinnunen, T., Ertaş, F., Saeidi, R., Pohjalainen, J., Alku, P., 2012. Regularized all-pole models for speaker verification under noisy environments. IEEE Signal Processing Letters 19 (3), 163–166.

Hasan, T., Hansen, J.H.L., 2013. Acoustic factor analysis for robust speaker verification. IEEE Transactions on Audio, Speech and Language Processing 21 (4), 842–853.

Hirsch, H.G., Pearce, D., 2000. 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, pp. 181–188.

Jankowski, C.R., Lippmann, R.P., 1992. Comparison of auditory model for robust speech recognition. In: Proceedings of the Workshop on Speech and Natural Language, Stroudsburg, PA, pp. 453–454.

Kang, S.Y., Colesa, D.J., Swiderski, D.L., Su, G.L., Raphael, Y., Pfingst, B.E., 2010. Effects of hearing preservation on psychophysical responses to cochlear implant stimulation. Journal of the Association for Research in Otolaryngology 11 (2), 245–265.

Kim, D.S., Lee, S.Y., Kil, R.M., 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing 7 (1), 55–69.

Kim, C., Chiu, Y.-H.B., Stern, R.M., 2006. Physiologically-motivated synchrony-based processing for robust speech recognition. In: Proceedings of Interspeech, Pittsburgh, Pennsylvania, pp. 1975–1978.

Kim, C., Stern, R.M., 2012. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proceedings Acoustics, Speech and, Signal Processing, pp. 4101–4104.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52 (1), 12–40.

Kinnunen, T., Saeidi, R., Sedlák, F., Lee, K.A., Sandberg, J., Hansson-Sandsten, M., Li, H., 2012. Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Transactions on Audio, Speech and Language Processing 20 (7), 1990–2001.

Li, Q., Huang, Y., 2010. Robust speaker identification using and auditory-based feature. In: Proceedings of Acoustics Speech and, Signal Processing, pp. 4514–4517.

Li, Q., Huang, Y., 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Transactions on Audio, Speech and Language Processing 19 (6), 1791–1801.

Lyon, R.F., 1982. A computational model of filtering, detection, and compression in the cochlea. In: Proceedings of the IEEE-International Conference on Acoustics, Speech, and Signal Processing, Paris, pp. 1282–1285.

May, B.J., Sachs, M.B., 1992. Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. Journal of Neurophysiology 68 (5), 1589–1602.

Middlebrooks, J.C., 2004. Effects of cochlear-implant pulse rate and inter-channel timing on channel interactions and thresholds. Journal of the Acoustical Society of America 116 (1), 452–468.

Miller, C.A., Woo, J., Abbas, P.J., Hu, H., Robinson, B.K., 2011. Neural masking by sub-threshold electric stimuli: animal and computer model results. Journal of the Association for Research in Otolaryngology 12 (2), 219–232.

Ming, J., Hazen, T.J., Glass, J.R., Reynolds, D.A., 2007. Robust speaker recognition in noisy conditions. IEEE Transactions on Audio, Speech and Language Processing 15 (5), 1711–1723.

Moore, B.C.J., 2003. An Introduction to the Psychology of Hearing, 5th ed. Academic Press, London, pp. 39–41.

Nizami, L., 2005. Dynamic range relations for auditory primary afferents. Hearing Research 208 (1–2), 26–46.

Ohzawa, I., Sclar, G., Freeman, R.D., 1985. Contrast gain control in the cat's visual system. Journal of Neurophysiology 54 (3), 651–658.

Patterson, R.D., Holdsworth, J., Allerhand, M., 1992. Auditory models as preprocessors for speech recognition. In: Schouten, M.E.H. (Ed.), The Auditory Processing of Speech: From Sounds to Words. Mouton de Gruyter, Berlin, Germany, pp. 67–83 (Chapter 1).

Pfingst, B.E., Bowling, S.A., Colesa, D.J., Garadat, S.N., Raphael, Y., Shibata, S.B., Strahl, S.B., Su, G.L., Zhou, N., 2011. Cochlear infrastructure for electrical hearing. Hearing Research 281 (1–2), 65–73.

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

Rabinowitz, N.C., Willmore, B., Schnupp, J., King, A.J., 2011. Contrast gain control in auditory cortex. Neuron 70 (6), 1178–1191.

Reynolds, D., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17 (1–2), 91–108.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing 10 (1–3), 19–41.

Rhode, W.S., Cooper, N.P., 1993. Two-tone suppression and distortion production on the basilar membrane in the hook region of cat and guinea pig cochleae. Hearing Research 66, 31–45.

Robles, L., Ruggero, M., 2011. Mechanics of the mammalian cochlea. Physiological Reviews 81 (3), 1305–1352.

Sachs, M.B., Abbas, P.J., 1974. Rate versus level functions for auditory-nerve fiber in cats: tone burst stimuli. Journal of the Acoustical Society of America 56 (6), 1835–1847.

Saeidi, R., Pohjalainen, J., Kinnunen, T., Alku, P., 2010. Temporally weighted linear prediction features for tackling additive noise in speaker verification. IEEE Signal Processing Letters 17 (6), 599–602.

Schneider, B.A., Parker, S., Murphy, D., 2011. A model of top down gain control in the auditory system. Attention, Perception and Psychophysics 73 (5), 1562–1578.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics 16 (1), 55–76.

Shamma, S.A., 1985. Speech processing in the auditory system I: the representation of speech sounds in the responses of the auditory nerve. Journal of the Acoustical Society of America 78 (5), 1612–1621.

Shamma, S.A., 1988. The acoustics features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. Journal of Phonetics 16, 77–91.

Shao, Y., Srinivasan, S., Wang, D.L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proceedings of Acoustics Speech and Signal Processing, vol. IV, pp. 277–280.

Shao, Y., Wang, D.L., 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In: Proceedings of Acoustics Speech and, Signal Processing, pp. 1589–1592.

Shao, Y., Srinivasan, S., Jin, Z., Wang, D.L., 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language 24 (1), 77–93.

Shin, J.W., Kwon, H.J., Jin, S.H., Kim, N.S., 2008. Voice activity detection based on conditional MAP criterion. IEEE Signal Processing Letters 15, 257–260.

Slaney, M., Auditory Toolbox, Version 2, Technical Report No. 1998–010, Interval Research Corporation, 1998.

Stern, R.M., Morgan, N., 2012a. Features based on auditory physiology and perception. In: Virtanen, T., Raj, B., Singh, R. (Eds.), . In: Techniques for Noise Robustness in Automatic Speech Recognition. Wiley.

Stern, R.M., Morgan, N., 2012b. Hearing is believing: biologically-inspired feature extraction for robust speech recognition. IEEE Signal Processing Magazine 20 (6), 34–43.

Taberner, A.M., Liberman, M.C., 2005. Response properties of single auditory nerve fibers in the mouse. Journal of Neurophysiology 93 (1), 557–569.

Wang, K., Shamma, S., 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Transactions on Speech and Audio Processing 2 (3), 421–435.

Wang, N., Ching, P.C., Zheng, N., Lee, T., 2011. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE Transactions on Audio, Speech and Language Processing 19 (1), 196–205.

Watkins, P.V., Barbour, D.L., 2011. Level-tuned neurons in primary auditory cortex adapt differently to loud versus soft sounds. Cerebral Cortex 21 (1), 178–190.

Wen, B., Wang, G.I., Dean, I., Delgutte, B., 2009. Dynamic range adaptation to sound level statistics in the auditory nerve. Journal of Neuroscience 29 (44), 13797–13808.

Wen, B., Wang, G.I., Dean, I., Delgutte, B., 2012. Time course of dynamic range adaptation in the auditory nerve. Journal of Neurophysiology 108 (1), 69–82.

Werblin, F.S., Jacobs, A., Teeters, J., 1996. The computational eye. IEEE Spectrum 33 (5), 30–37.

Winslow, R.L., Sachs, M.B., 1987. Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. Journal of Neurophysiology 57 (4), 1002–1021.

Wu, W., Zheng, T.F., Xu, M.-X., Soong, F.K., 2007. A cohort-based speaker model synthesis for mismatched channels in speaker verification. IEEE Transactions on Audio, Speech and Language Processing 15 (6), 1893–1903.

Yates, G.K., Winter, I.M., Robertson, D., 1990. Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hearing Research 45 (3), 203–219.

Young, E.D., 2008. Neural representation of spectral and temporal information in speech. Philosophical Transactions of Royal Society B: Biological Science 363 (1493), 923–945.

Zilany, M.S., Carney, L.H., 2010. Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics. The Journal of Neuroscience 30 (31), 10380–10390.