On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition

Dimitrios Dimitriadis, *Senior Member, IEEE*, Petros Maragos, *Fellow, IEEE*, and Alexandros Potamianos, *Senior Member, IEEE*

***Abstract*—In this paper, we examine how energy computation and filterbank design contribute to the overall front-end robustness, especially when the investigated features are applied to noisy speech signals, in mismatched training-testing conditions. In prior work ("Auditory Teager energy cepstrum coefficients for robust speech recognition," D. Dimitriadis, P. Maragos, and A. Potamianos, in *Proc. Eurospeech'05*, Sep. 2005), a novel feature set called "Teager energy cepstrum coefficients" (TECCs) has been proposed, employing a dense, smooth filterbank and alternative energy computation schemes. TECCs were shown to be more robust to noise and exhibit improved performance compared to the widely used Mel frequency cepstral coefficients (MFCCs). In this paper, we attempt to interpret these results using a combined theoretical and experimental analysis framework. Specifically, we investigate in detail the connection between the filterbank design, i.e., the filter shape and bandwidth, the energy estimation scheme and the automatic speech recognition (ASR) performance under a variety of additive and/or convolutional noise conditions. For this purpose: 1) the performance of filterbanks using triangular, Gabor, and Gammatone filters with various bandwidths and filter positions are examined under different noisy speech recognition tasks, and 2) the squared amplitude and Teager–Kaiser energy operators are compared as two alternative approaches of computing the signal energy. Our end-goal is to understand how to select the most efficient filterbank and energy computation scheme that are maximally robust under both clean and noisy recording conditions. Theoretical and experimental results show that: 1) the filter bandwidth is one of the most important factors affecting speech recognition performance in noise, while the shape of the filter is of secondary importance, and 2) the Teager–Kaiser operator outperforms (on the average and for most noise types) the squared amplitude energy computation scheme for speech recognition in noisy conditions, especially, for large filter bandwidths. Experimental results show that selecting the appropriate filterbank and energy computation scheme can lead to significant error rate reduction over both MFCC and perceptual linear predicion (PLP) features for a variety of speech recognition tasks. A relative error rate reduction of up to $\sim$30% for MFCCs and $\sim$39% for PLPs is shown for the Aurora-3 Spanish Task.**

Manuscript received February 17, 2010; revised June 05, 2010, September 27, 2010; accepted October 05, 2010. Date of publication November 15, 2010; date of current version May 25, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

D. Dimitriadis was with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece, Athens GR-15773, Greece. He is now with AT&T Labs, Florham Park, NJ 07932 USA (e-mail: ddim@research.att.com).

P. Maragos is with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece, Athens GR-15773, Greece (e-mail: maragos@cs.ntua.gr).

A. Potamianos is with the Department of Electronics and Computer Engineering, Technical University of Crete, Chania GR-73100, Greece (e-mail: potam@telecom.tuc.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2092766

***Index Terms*—Bandpass filters, cepstrum analysis, error analysis, parameter estimation, robustness, spectral analysis, speech processing, speech recognition, time–frequency analysis.**

I. Introduction

ROBUST feature extraction is a complex problem much studied over the years. Despite recent progress in the domain of robust automatic speech recognition (ASR), many questions, such as how the energy estimation process and the filterbank design affect ASR performance under noise, especially for various levels of additive/convolutional noise and acoustic model mismatch, remain open. The effect of noise on the features employed in a speech recognition front-end is nontrivial and can greatly influence the overall system performance. In this context, much work has been done minimizing this mismatch [2], [3] by using transformations of the noisy features to a "cleaner" feature domain, and thus improving their invariability to certain noise types. Other related work includes speech enhancement [4], normalization of the noisy features statistical properties [5]–[7], and dynamic feature combinations [8]. In [9]–[11], the effect of environmental noise on the statistical speech models was investigated and two algorithms (CDCN and MFCDCN) were proposed for compensating it. However, the feature robustness problem remains unsolved in a globally optimal way. Our goal, in this paper, is to analyze both theoretically and experimentally, how the filterbank design parameters and energy computation scheme affect the robustness of speech recognition systems in noisy recording conditions.

The use of filterbanks in ASR front-ends was motivated by the human hearing process [12]–[14], where the energy across frequencies of the audio spectrum is resolved by using auditory filters. Although the human hearing process is for the most part heavily researched, machines have been unable to match the robustness that human beings exhibit in speech recognition in noise [15]. Efforts to model the human audio processing to further improve the robustness of speech recognition front-end have had limited success, e.g., perceptual linear prediction features (PLP) [16], relative spectral transform features (RASTA) [17], dynamic spectral subband centroids [18], or the auditory-based features [19]. However, for the past two decades, the Mel frequency cepstrum coefficients (MFCCs) [20] have remained the most widely used features for ASR applications mainly because they combine good discrimination capabilities with low computational complexity. These features incorporate some aspects of the human hearing process, such as the nonlinear filter placing (mel-scale) and subband energy estimation, and perform well in relatively clean and well-matched conditions. On

the other hand, MFCCs lack robustness in adverse recording or noise mismatch conditions.

Recently, the *Teager energy cepstrum coefficients (TECCs)* have been proposed and shown to outperform the MFCCs, especially in noisy recognition tasks and under mismatched training/testing conditions [1]. The TECCs employed an alternative energy estimation scheme, i.e., use of the Teager–Kaiser instead of the square amplitude energy operator [21], and human hearing-inspired filterbanks, i.e., Gammatone filters placed on the equivalent rectangular bandwidth (ERB) curve. The ERB is a measure used in psychoacoustics, approximating the bandwidths of the filters in human hearing by rectangular bandpass filters. It was first introduced for speech processing applications in [22] and [23].

The main goals of this paper are to: 1) adequately present the TECC "family" of features, i.e., the TECCs and other front-ends employing similar design parameters, 2) investigate under what noise conditions this new family of features outperforms the MFCCs, 3) provide theoretical and experimental results on the optimality of the energy computation scheme (squared amplitude versus Teager–Kaiser energy operator), and 4) investigate the optimal design of the filterbank (number of filters, filter bandwidth, and shape) for noisy speech recognition tasks. Specifically, we compare the *mean Teager–Kaiser (MTE)* or *mean square amplitude (MSE)* energy schemes for cepstrum-based feature extraction, when applied to speech signals corrupted by additive and/or convolutional noise. Further, we analyze the performance of the energy computation schemes as a function of the filterbank design parameters, e.g., bandwidth in conjunction with the noise spectral characteristics. Overall, different key parameters of the feature extraction process are investigated and ASR experiments are undertaken to examine their impact on the corresponding recognition results. This work builds upon theoretical results in [21].

The paper is organized in sections as follows. In Section II, the clean speech and the harmonic noise models are introduced. Herein, the input signals are bandpass filtered and the respective filter bandwidths are examined, as well. A unified energy estimation scheme is presented, where the *Teager–Kaiser energy operator* (TEO) and the *square amplitude energy operator* (SEO) are only two cases of the general scheme (Section II-C). It is shown that the energy estimation performance is much dependent on the filter bandwidth. The proposed feature extraction process is presented in Section III. In Section IV, it is investigated how additive and convolutional noise types affect the proposed features. The performance of these features in speech processing applications is presented in Section V; both energy estimation and speech recognition in noise are investigated. Finally, the conclusions and discussion of future work are provided in Section VI.

II. Background

In most speech processing applications, speech signals $x(t)$ are filtered by filterbanks yielding $r_j(t) = g_j(t) * x(t)$, where $g_j(t)$ is the impulse response of the jth analysis filter and "$*$" stands for convolution. The AM–FM speech model suggests the decomposition of the speech signal into J (resonance inspired) signals $r_j(t)$, where J the number of deployed filters in the analysis filterbank [24], [25]

$$x(t) \approx \sum_{j=1}^{J} r_j(t) = \sum_{j=1}^{J} a_j(t) \cos\left(\int_0^t \omega_j(\tau) d\tau + \theta_j\right) \quad (1)$$

where $a_j(t)$, $\omega_j(t)$ are the instantaneous amplitude and frequency modulating signals and θ_j is a phase offset. Herein, the underlying assumption is that the information-carrying signals $a_j(t)$,$\omega_j(t)$ are slowly varying compared to the carrier frequencies. Next, we summarize the main theoretical results from [21].

A. Harmonic Noise Modeling

An approximation of a bandpass noise signal $v_j(t)$ was first proposed in [26] and [27] and used in [21]. The noise signal is modeled as a sum of stationary sinusoids $v_{jk}(t)$ $(k = 1, \ldots, K_j)$, with fixed amplitudes b_{jk}, phase offsets θ_{jk} that are independent random variables uniformly distributed over $[-\pi, \pi]$ and frequencies ω_{jk} placed equidistantly with spacing ω_R,

$$v_j(t) \approx \sum_{k=1}^{K_j} v_{jk}(t) = \sum_{k=1}^{K_j} b_{jk} \cos\left(\omega_{jk}\, t + \theta_{jk}\right) \quad (2)$$

The number of sinusoid components K_j is given by $K_j \triangleq \lceil B_j/\omega_R \rceil$, where B_j is the jth-filter passband. Thus, we approximate noise with more components K_j when the filter passband is broader.

B. Noisy Teager–Kaiser Energy Estimation

If we apply the *Teager–Kaiser energy (TEO)* operator [24] to the bandpassed noisy signal $s_j(t) \triangleq r_j(t) + v_j(t)$, its long-term *mean Teager–Kaiser energy* (MTE) [21] is a sum of two components

$$\langle \Psi[s_j(t)] \rangle \approx \langle a_j^2(t) \omega_j^2(t) \rangle + \sum_k b_{jk}^2 \omega_{jk}^2 \quad (3)$$

where $\langle \cdot \rangle$ denotes the time-averaging process.

The *normalized deviation* $D_{\mathcal{T}}$ provides a measure of the robustness of energy estimation in additive noise and is defined as the ratio of the difference between the mean noisy and clean energy estimates over the mean clean estimates

$$D_{\mathcal{T}}[s_j, r_j] \approx \frac{\sum_{k=1}^{K_j} b_{jk}^2 \omega_{jk}^2}{\langle a_j^2(t) \omega_j^2(t) \rangle}. \quad (4)$$

The normalized deviation $D_{\mathcal{T}}$ is proportional to the squared product of ω_{jk} with the amplitude coefficients b_{jk}, and inversely proportional to the mean instantaneous frequency $\omega_j^2(t)$ weighted by $a_j^2(t)$. Therefore, the $D_{\mathcal{T}}$ estimates depend on the *relative* spectral energy distribution (within the frequency band of interest) of the noise and speech signals, as detailed in [21].

C. Noisy Squared Amplitude Energy Estimation

The *mean squared amplitude energy* (MSE) for $s_j(t)$ is given by

$$\left\langle s_j^2(t)\right\rangle \approx \frac{1}{2}\left(\left\langle a_j^2(t)\right\rangle + \sum_k b_{jk}^2\right). \tag{5}$$

Similarly, the *normalized deviation* $D_{\mathcal{S}}$ for the MSE case is

$$D_{\mathcal{S}}[s_j, r_j] \approx \frac{\sum_{k=1}^{K_j} b_{jk}^2}{\left\langle a_j^2(t)\right\rangle}. \tag{6}$$

The $D_{\mathcal{S}}$ estimates are approximately equal to the inverse signal-to-noise ratio (SNR) values in the filter passband. Henceforth, the signal arguments, i.e., the signals $s_j(t), r_j(t)$, will be ignored in $D_{\mathcal{T}}$ and $D_{\mathcal{S}}$ for notational simplicity.

The MTE normalized deviation (4) can be formulated as the ratio of the second-order spectral centroid of the noise over the clean signal [25], while, the MSE deviation (6) is the ratio of the zeroth-order spectral centroids [18]. We can express both of these deviations with a compact notation

$$D^{(p)} = \frac{\sum_k b_{jk}^2 \omega_j^p}{\int_{B_j} \omega^p |X(\omega)|^2 d\omega}. \tag{7}$$

For $p = 0:\ D^{(0)} \equiv D_{\mathcal{S}}$, whereas for $p = 2:\ D^{(2)} \equiv D_{\mathcal{T}}$.

Based on the equations above, the spectral energy distribution (pth-order spectral moments) within the frequency band of interest determines the relative performance of the MSE and MTE[1] estimates. In general, the MTE values present smaller estimation errors (deviations) when compared to the MSE ones when the high-energy noise components are concentrated over low frequencies (within the passband), and vice-versa, all due to the weighting term ω^p that affects the overall spectral energy distribution of the input signal [21]. The MTE and MSE estimates are obviously related, due to this.[2]

D. Medium and Short-Time Properties of Energy Operators

The analysis above assumes that the duration of the averaging window is long enough to ignore all transient terms. However, the estimation errors of the MTE and MSE schemes depend on the window length, as well. In the case when medium- and short-time windows (less than 15 ms) are considered, transient terms contribute to the estimation error and should be taken into further account in the analysis. In this context, the MTE deviation values are expected smaller than those of the MSE ones. Finally, all the transient terms are inversely proportional to the frequency content, e.g., filter center frequency ω_c. Therefore, these deviation terms are further emphasized for smaller frequency values. A more detailed description can be found in [21].

[1]The *relative* performance of MSE versus MTE scheme does not solely depend on the signal-to-noise ratio in the frequency band.

[2]Higher-order derivatives of the input signal correspond to larger values of p, [21].

E. Narrowband Signal Analysis

For narrowband signals the signal $s_j(t)$ is approximated by a two-cosine sum, i.e., the noise has a single frequency component

$$s_j(t) = \underbrace{a_j(t)\cos(\omega_{cj}t + \theta_r)}_{r_j(t):\text{Clean Signal}} + \underbrace{b\cos(\omega_{cj}t + \theta_v)}_{v_j(t):\text{Noise Signal}}$$

where ω_{cj} is the jth filter center frequency. Then,

$$\Psi[s_j(t)] \approx \omega_{cj}^2 a_j^2(t) + \omega_{cj}^2 b^2 + 2b\omega_{cj}^2 a_j(t)\cos(\theta_r - \theta_v).$$

Assuming that $\dot{a}_j(t) \approx 0$ and $\ddot{a}_j(t) \approx 0$, the noisy signal MTE estimate (3) is given by

$$\langle \Psi[s_j]\rangle = \omega_{cj}^2\left(\left\langle a_j^2(t)\right\rangle + b^2\right)$$

and the normalized deviation $D_{\mathcal{T}}$, (4), is given by

$$D_{\mathcal{T}}[s_j, r_j] = \frac{b^2}{\left\langle a_j^2(t)\right\rangle}. \tag{8}$$

Correspondingly, for the MSE case,

$$\begin{aligned} s_j^2 = {} & \frac{1}{2}a_j^2(t)[1 + \cos(2\omega_{cj}t + 2\theta_r)] \\ & + \frac{1}{2}b^2[1 + \cos(2\omega_{cj}t + 2\theta_v)] \\ & + ba_j(t)[\cos(\theta_r - \theta_v) + \cos(2\omega_{cj}t + \theta_r + \theta_v)] \end{aligned}$$

and, the MSE deviation, (6), is

$$D_{\mathcal{S}}[s_j, r_j] = \frac{b^2}{\left\langle a_j^2(t)\right\rangle}. \tag{9}$$

From (8) and (9), it is concluded that both long-term $D_{\mathcal{T}}$ and $D_{\mathcal{S}}$ are equal when narrow bandpass filters are used. Consequently, no significant difference is expected when employing different energy operators on narrowband signals (this is the case of approximately monochromatic signals). However, the MSE estimates include time-decaying transient phenomena, as opposed to the MTE scheme where these phenomena are not present (in the case of shorter averaging windows). In general, the MTE estimates are expected to present smaller deviations than the MSE ones, as outlined in Section II-D. The experimental verification of this analysis is presented in Section V.

III. Generalized Cepstrum Coefficient Front-Ends

Next, we investigate cepstral features that are computed using different filterbanks and energy computation schemes, i.e., the mel Teager-energy cepstral coefficients and their generalizations.

A. ERB and Maximally Smooth Filterbanks

The *Equivalent Rectangular Bandwidth* (ERB) has been introduced to measure the bandwidth of asymmetrical IIR filters, such as the Gammatone filters. Given that $|G(\omega_c)|$ is the maximum gain of a bandpass filter with frequency response $G(\omega)$, reached at frequency ω_c, then the filter ERB is defined as

$$\text{ERB} = \frac{\int |G(\omega)|^2 d\omega}{|G(\omega_c)|^2}. \tag{10}$$

In other words, the ERB is the bandwidth of a rectangular shaped filter when its energy (the integral of its frequency response magnitude squared) is normalized by the maximum gain squared, $|G(\omega_c)|^2$. By normalizing the filter ERBs, their design parameters have to be modified accordingly.

A Gabor filter impulse response is given by

$$g_{\mathrm{Gab}}(t) = e^{-b^2t^2}\cos(\omega_c t) \tag{11}$$

where b is a parameter controlling the filter bandwidth and ω_c is its center frequency. According to [28], the corresponding ERB value is $B_{\mathrm{Gab}} = b/\sqrt{2\pi}$.

Further, the impulse response of a Gammatone filter is given by

$$g_{\mathrm{Gamm}}(t) = t^3 e^{-2\pi 1.1019bt}\cos(\omega_c t) \tag{12}$$

where b is a bandwidth controlling parameter and ω_c is its center frequency. Its ERB value is given by $B_{\mathrm{Gamm}} = b$ [23].

When the filters have equal bandwidth parameters-b

$$B_{\mathrm{Gab}} = 1/\sqrt{2\pi}\cdot B_{\mathrm{Gamm}} \quad \text{or} \quad B_{\mathrm{Gab}} \simeq 0.4\cdot B_{\mathrm{Gamm}} \tag{13}$$

meaning that for the same design parameter b Gabor filters are *narrower* than the corresponding Gammatone ones. By considering (13), the Gabor filter bandwidths should be *normalized* by a factor of approximately (times) 2.5 to achieve the same equivalent filtering passband as with the Gammatone filter passbands. Henceforth, equal ERB values are assumed when comparing ASR results corresponding to Gabor and Gammatone filterbanks.

B. Generalized Cepstrum Coefficients

MFCCs are typically computed using a filterbank of 22–25 triangular filters with 50% bandwidth overlap[3]; the (log) mean mel energy coefficients are estimated and then transformed to the Cepstrum domain via the discrete cosine transform (DCT). The feature sets analyzed in this paper, as proposed in [1], employ smoother and broader filters. The use of such filters, i.e., Gammatone or Gabor filters, for estimating the cepstral coefficients, is supported by the broader filter approach, as presented in [29]. In addition to that, different energy estimation schemes have been investigated, providing additional robustness to the proposed features (depending though, on the spectral fingerprint of the clean and noise signals).

The feature extraction algorithm consists of the following steps. Fig. 1:

1) Filter the speech signal using a mel-spaced filterbank. The filterbank consists of 25–100 smooth filters and uses Gabor, Gammatone or Gammachirp filters.[4]
2) Estimate the MTE or MSE mel-energy coefficients of the framed bandpassed signals.
3) Transform these energy coefficients into the Cepstrum domain. Only the first low-order cepstral coefficients are kept for recognition (the *de facto* standard is to keep the first 13 coefficients, including C0).
4) Estimate their first and second order time derivatives and perform cepstral mean subtraction (CMS).[5]

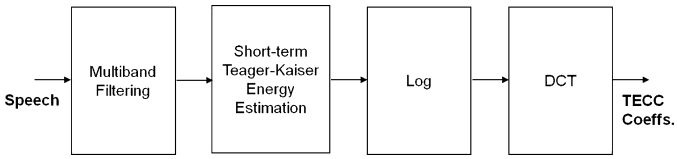

Fig. 1. Block diagram of the TECC feature extraction process.

In [14] and [19], it is conjectured that the Gammatone filters equidistantly placed in the Mel-frequency scale, resemble the human ear. The first two of the steps substantially differentiate the proposed algorithm from the typical MFCC algorithm. The following two steps, i.e., the cepstral coefficient estimation and the truncation process, remain the same as in [20]. The ASR results presented in [1] and in Section V below, show significant improvement, especially for recognition tasks in noise. The additional robustness to noise can be attributed to the use of wider filters and the use of alternative energy estimation schemes, i.e., the MTE scheme.

IV. Error Analysis for Cepstrum Features in Noise

Until now, the bandpass filters were considered ideal where their amplitude response was rectangular with fixed amplitude equal to unity. Herein, the aforementioned analysis is generalized for a wider "family" of bandpass filters.

Under the conditions detailed in [30] and [31] for speech and [32] for image signals, a filtered bandpass AM-FM signal $r_i(t)$ can be approximated by

$$r_j(t) \approx a_j(t)|G_j[\omega_j(t)]|\cos\left\{\int_0^t \omega(\tau)d\tau + \angle G_j[\omega_j(t)] + \theta_j\right\} \tag{14}$$

where $G_i[\cdot]$ is the frequency response of the jth filter. The approximation is exact when $r_j(t)$ is monochromatic, i.e., $\omega_j(t) = \mathrm{constant}$. Further, in the case of real, symmetric filters, e.g., Gabor filters, $\angle G_j[\omega_j(t)] = 0$ and the filtering process affects only the instantaneous amplitude signal $a(t)$ [30]. Similarly to (14), the noise signal can be rewritten as

$$v_j(t) = \sum_{k=0}^{K_j} b_{jk}|G_j[\omega_{jk}]|\cos\{\omega_{jk}\,t + \theta_{jk} + \angle G_j[\omega_{jk}]\}. \tag{15}$$

In the case of filtering the speech signals, the instantaneous amplitude signals are given by

$$a_j(t)|G_j[\omega_j(t)]| \quad \text{and} \quad b_{jk}|G_j[\omega_{jk}]|.$$

The phase offsets, i.e., $\angle G_j[\omega_j(t)]$ and $\angle G_j[\omega_{jk}]$, are averaged out.[6] Only in the cases of short- and medium-term energy averaging, these phase offsets should be considered.

[3]The triangular filters present finite passband support therefore, the overlap is, usually, estimated over them.

[4]Herein, results only for the first two types of filters are reported.

[5]The experimental results using features without CMS are similar. However, these results appear more noisy making conclusions less clear.

[6]Assuming that $\omega_j(t)$ is smooth enough, then $\angle G_j[\omega_j(t)] \approx$ constant.

A. Cepstral Coefficient Error Analysis

As shown above, the noisy speech energy coefficients, (3), (5), are the sum of the speech and the noise energy coefficients (given sufficient length for the averaging window), i.e., $P_s[j,m] = P_r[j,m] + P_v[j,m]$, where $j = 1,\ldots,J$, $m = 1,\ldots,M$, J the number of filters and M the total number of frames. For the case of MTE, $P_r[j,m] = \langle\Psi[r_j(t)]\rangle$ and $P_v[j,m] = \langle\Psi[v_j(t)]\rangle$,[7] for $t \in m$th-time frame. Henceforth, to simplify the notation we shall drop the frame index m from all equations. Note that this analysis holds true for each one of the frames.

With a unified notation for both energy schemes, similarly to (7), the *cepstral Mel energy coefficients* [7], [26] are given by

$$\mathbf{C}_s^{(p)} = \mathbf{W}\cdot\log(\mathbf{P}_s^{(p)}) \tag{16}$$

where $\mathbf{C}_s^{(p)} = (C_s^{(p)}[1],\ C_s^{(p)}[2],\ \ldots,\ C_s^{(p)}[I])^T$ is the vector of the estimated noisy cepstral coefficients with length I, $\mathbf{W}$ is an $I \times J$ discrete cosine transform matrix and $\mathbf{P}_s^{(p)} = (P_s^{(p)}[1],\ P_s^{(p)}[2],\ \ldots, P_s^{(p)}[J])^T$ and $\mathbf{P}_r^{(p)} = (P_r^{(p)}[1],\ P_r^{(p)}[2],\ \ldots, P_r^{(p)}[J])^T$ are the noisy and clean speech Mel-energy coefficient vectors estimated over the J filter passbands. Depending on the energy estimation scheme, the parameter $p = 0$ or 2 (when $p = 0$, we refer to the MSE values, and for $p = 2$ to the MTE coefficients). To further simplify the notation, we shall, henceforth, drop the superscript p, as well. The analysis below holds true for either $p = 0$ or $p = 2$.

Equation (16) is rewritten element-wise, as

$$C_s[i] = \sum_{j=1}^{J} W_{ij}\log(P_s[j]) \tag{17}$$

where $1 \le i \le I$ and $W_{ij} = \sqrt{2/J}\cdot\cos[\pi i(j-1/2)/J]$.

Inspired by the analysis in [10], [11], we introduce the *cepstral coefficient deviation* $\Delta C[i]$ as the difference of the noisy and the clean speech cepstral coefficients, i.e., $C_s[i]$ and $C_r[i]$,

$$\Delta C[i] = C_s[i] - C_r[i] = \sum_{j=1}^{J} W_{ij}\log\left(\frac{P_s[j]}{P_r[j]}\right). \tag{18}$$

From the analysis in Section II, (18) leads to

$$\Delta C[i] = \sum_{j=1}^{J} W_{ij}\log\left(1+\frac{P_v[j]}{P_r[j]}\right)$$

where the quantity $P_v[j]/P_r[j]$ is the *normalized MTE or MSE Mel energy deviations* (7), within the jth filter passband. Therefore,

$$\Delta C[i] = \sum_{j=1}^{J} W_{ij}\log(1+D_j) \tag{19}$$

where $D_j \triangleq P_v[j]/P_r[j]$ is the estimated energy deviation for the jth filter index, assuming $P_r[j] \neq 0,\ \forall j$. The ΔC deviation values provide an indication of how noise (of different spectral characteristics) corrupts the MTE- and MSE-based cepstral coefficients. These deviations consist of a linear combination of the log energy deviations weighted by W_{ij}, across all filters. Therefore, the energy deviation values corresponding to different frequency bins linearly affect all the cepstral coefficients. Consequently, smaller energy estimation errors will yield smaller cepstral feature deviations from the clean ones.[8]

B. Convolutional Noise Analysis

In the presence of both additive and convolutional noise the corrupted speech signal equals to $r_j(t) * h_j(t) + v_j(t)$. As defined in the previous sections, the normalized Mel-energy coefficient deviation is given by

$$D_j = \frac{M_j[v_j] + M_j[r_j * h_j] - M_j[r_j]}{M_j[r_j]} \tag{20}$$

where r_j, v_j and h_j are the framed ($t \in m$th time frame) bandpassed clean speech, additive and convolutional noise signals, respectively. Further,

$$M_j[v_j] = \int_{B_j} \omega^p |G_j(\omega)N(\omega,m)|^2 d\omega$$
$$M_j[r_j * h_j] = \int_{B_j} \omega^p |G_j(\omega)X(\omega,m)|^2 |H(\omega,m)|^2 d\omega$$

and

$$M_j[r_j] = \int_{B_j} \omega^p |G_j(\omega)X(\omega,m)|^2 d\omega$$

where $G_j(\omega)$ is the jth filter frequency response and B_j its passband, whereas $X(\omega,m), N(\omega,m)$ and $H(\omega,m)$ are, respectively, the periodograms of the clean, additive and convolutional noise signal frames, and p defined as above.

By substitution, we obtain

$$D_j = \frac{\int_{B_j}\omega^p|G_j(\omega)N(\omega,m)|^2 d\omega}{\int_{B_j}\omega^p|G_j(\omega)X(\omega,m)|^2 d\omega} + \frac{\int_{B_j}\omega^p(|H(\omega,m)|^2-1)|G_j(\omega)X(\omega,m)|^2 d\omega}{\int_{B_j}\omega^p|G_j(\omega)X(\omega,m)|^2 d\omega} \tag{21}$$

The normalized deviations[9] D_j consist of two terms accounting for the two different noise types, i.e., the additive and the convolutional noise parts:

$$D_j = D_j^{\text{conv}} + D_j^{\text{add}} \tag{22}$$

where

$$D_j^{\text{conv}} \triangleq \frac{\int_{B_j}\omega^p(|H(\omega,m)|^2-1)|G_j(\omega)X(\omega,m)|^2 d\omega}{\int_{B_j}\omega^p|G_j(\omega)X(\omega,m)|^2 d\omega}$$

[7]Herein, only the MTE case is presented. However, the same equation holds true for the case of MSE, as well.

[8]The energy-related errors can be attributed to both the estimation process and the existence of noise.

[9]We assume that D_j is non-negative and in the rare occasions when it takes negative values we suggest thresholding it.

and

$$D_j^{\text{add}} \triangleq \frac{\int_{B_j} \omega^p |G_j(\omega)N(\omega,m)|^2 d\omega}{\int_{B_j} \omega^p |G_j(\omega)X(\omega,m)|^2 d\omega}.$$

Assuming that $|H(\omega,m)|$ remains almost constant for a certain time frame and across all frequency bands, then $|H(\omega,m)|^2 - 1 \approx H_j$ and

$$D_j = H_j + D_j^{\text{add}}$$

Finally, after substituting the noise model, we obtain

$$D_j = H_j + \frac{\sum_k \omega_k^p b_k^2 |G_j(\omega_k)|^2}{\int_{B_j} \omega^p |G_j(\omega)X(\omega,m)|^2 d\omega}. \quad (23)$$

The assumption of convolutional noise with constant spectral characteristics for each time frame, adds a constant deviation term H_j to the total normalized deviation. This constant error term can be easily removed via an energy normalization post-processing scheme, e.g., mean value subtraction [6].

In the general case of noisy signals contaminated by both additive and convolutional noise, the cepstral deviation $\Delta C[i]$ (19) will, now, contain an additional term (22)

$$\Delta C[i] = \sum_{j=1}^{J} W_{ij} \log\left(1 + D_j^{\text{conv}} + D_j^{\text{add}}\right). \quad (24)$$

Similar results are presented in [9] and [10] for the case of the MFCCs. In this context, it should be highlighted the importance of the weighting term ω^p that emphasizes certain parts of the signal power spectrum (according to the values of p) and thus, can provide smaller cepstral coefficient deviations $\Delta C[i]$ when set accordingly. One of the paper contributions is based on the introduction of this weight ω^p to the feature extraction process.

V. Energy Estimation and Speech Recognition Experiments

In this section, various parameters of the feature extraction process are investigated experimentally in terms of noisy cepstral coefficient deviations from the clean case and their respective speech recognition performance. Specifically, the following parameters are evaluated: 1) the filter shape: *Gabor* or *Gammatone* filterbanks, 2) the number of filters: ranging from 25 to 100 filters,[10] 3) the filter bandwidth (while keeping the number of filters fixed), and 4) the energy scheme: *MTE* or *MSE* approaches. In all cases, the filters are equidistantly placed following the mel frequency scale. The bandwidth overlap is estimated by considering the filters' ERB values. The same design parameters are used for both Gabor and Gammatone filterbanks, i.e., same number of filters, filter placing and normalized ERB bandwidths.

A. Experimental Setup

For the experimental part of this paper three speech databases are used, i.e., the Aurora-3 (Spanish task), Aurora-4 and the TIMIT + Noise speech databases. The fundamental difference between these databases is that the first database contains real-life data, while the second and third databases contain data corrupted by artificially added noises. The Aurora-3 database is recorded inside the cabin of a moving car using both a close-talking and a near-field microphone. Thus, the data contain both convolutional and additive noise. Finally, the Aurora-4 task is a large-vocabulary speech recognition task (LV-ASR), contrary to the rest of the tasks that have a limited vocabulary and use all-pair grammars.

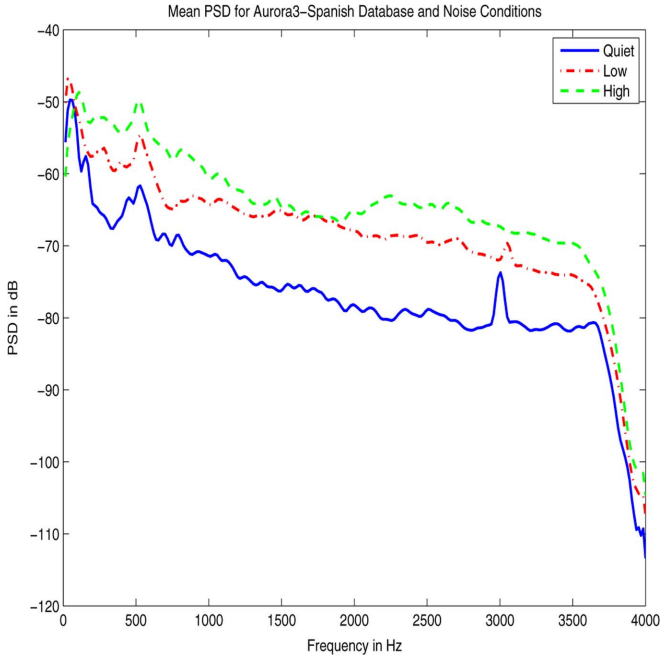

Fig. 2. Mean normalized PSD for the three different Aurora-3 noise conditions: quiet, low, and high noise levels. The mean PSDs are averaged over all noise frames of the same noise condition signals.

In more detail, the Aurora-3 database contains recordings of two different microphones and three noise levels with average SNR levels at 12, 9, and 5 dB, respectively. Three different training-testing scenarios are examined, i.e., the well-matched (WM), the medium-mismatch (MM) and the high-mismatch (HM) conditions. In the WM scenario, all microphone combinations and SNR levels are included in both the training and the testing sets. In the MM scenario, training and testing is performed using only the hands-free microphone recordings. In the HM condition, the close-talking microphone recordings are used for training, while the hands-free recordings are used for testing. Typically, car noise is assumed low-pass. However, the analysis of the mean normalized power spectrum density (PSD), shown in Fig. 2, does not fully support this assumption. Specifically, a high-pass noise component between 1500–2500 Hz, is present in the high-noise scenario, and an additional spike-like noise component around 3 kHz, can be noted for the quiet and low-noise scenarios. The first high-pass component can be attributed to the wind noise from the open windows while driving in high speed and/or the car-radio playing music, and the second one to the engine noise. This analysis is especially relevant for interpreting the results of the speech recognition task; as explained in Section II, the spectral shape of the noise determines

[10]In this set of experiments, the bandwidth overlap percentage between adjacent filters remains fixed. Consequently, changing the number of filters also affects the filter bandwidth.

the relative performance of the *MTE*- versus *MSE*-based cepstral features.

The Aurora-4 database has been created to investigate LV-ASR tasks in the presence of noise. The database is based on the WSJ database and the 5 k-words task for training and testing, respectively [33]. For training, the 16-kHz sampled noisy set is used. It contains a variety of noises added to the clean speech, and mixes data from several microphones. The test set was created by adding seven different noise types, i.e., clean, street traffic, train station, car, babble, restaurant, and airport, to two-microphone recordings yielding 14 different testing conditions [33]. The language model used is the baseline model provided by the ETSI configuration.

Finally, the third database (TIMIT + Noise) is created by artificially adding different types of noise to the TIMIT database. For this purpose, the NOISEX-92 noise database is used, containing ten typical noise samples, each with different spectral characteristics [34]. These noise signals are down-sampled to 16 kHz and added to the speech sentences[11] from the TIMIT database, while keeping the global average SNR fixed at SNR $=$ 5 dB.[12] The training is performed on the clean TIMIT data while the test sets consist of the noise-corrupted versions of the original TIMIT test set. Further, the clean speech signals are used as reference for comparing the normalized deviation and log distortion difference values of the estimated features.

The HMM-based HTK Tools platform is used for all ASR experiments. The statistical model for the Aurora-3 task consists of 11 context-independent, left-right, word HMMs that are trained using the ETSI WI007 training scripts. For the TIMIT + Noise tasks, the model consists of 46 phoneme-based, 3-state, left–right HMMs with 16 Gaussians per state. The grammar used for both cases is the all-pair, unweighted grammar. The MFCC, PLP, MSE- and MTE-based feature vectors consist of 39 coefficients, i.e., 13 cepstral coefficients (including C0) and their first and second time-derivatives.

The principal motivation behind including experiments on both real and artificial data is twofold: 1) using artificial data allows for the exact computation of the deviations (from the clean ones) for the ASR features, and 2) using real-life data presents different unaccounted sources of noises that degradate the ASR performance, i.e., Aurora-3 data. On the contrary, the underlying phenomena in TIMIT + Noise task are clearly presented and anticipated by the theoretical analysis.

B. Speech Signal Energy Deviations

Typically, the estimation of the signal time–frequency energy distributions is the first step in the feature extraction process. We compare the MTE and MSE computation schemes across all filters in the presence of additive noise. The normalized MTE and MSE energy deviations defined in (4) and (6) are actually the inverse subband SNRs, where the Mel-energy coefficient deviation from the clean estimates is the "noise" and the clean-case estimate is the "desired signal." Consequently, the SNR of the MSE scheme is defined[13] as $\mathrm{SNR}_{\mathcal{S}} \triangleq -10\log_{10}(\widehat{D}_{\mathcal{S}})$, and similarly for the MTE case, i.e., $\mathrm{SNR}_{\mathcal{T}} \triangleq -10\log_{10}(\widehat{D}_{\mathcal{T}})$. Energy estimation results are presented in terms of *mean SNR differences* (in dBs), or $\mathrm{SNR}_{\mathcal{S}} - \mathrm{SNR}_{\mathcal{T}}$. The differences assume negative values only when the averaged MTE-based deviations are smaller than the corresponding MSE ones. In that sense, the Teager–Kaiser operator provides more robust energy estimates than those based on MSE.

1000 instances of the phonemes */aa/* and */sh/* are extracted from the TIMIT + Noise database for each of the *babble*, *car*, and *white* noise types. Two different mel-spaced Gammatone filterbanks, using 25 or 100 filters (with constant 3-dB-bandwidth overlap of 50%) are used [1]. MTE and MSE coefficients are computed for each bandpassed signal using an analysis window of 30 ms, updated every 10 ms. The log root-mean-square (RMS) differences between the true and estimated MTEs and MSEs are computed and averaged over all frames and 1000 phonemic instances.

In Fig. 3(a)–(c), the mean log RMS error is shown for a Gammatone filterbank with 25 filters, while in Fig. 3(d)–(f) the error is shown for 100 filters. Given that for both cases the filter overlap is fixed at 50%, the bandwidths in the first case are four times larger than the later ones. As explained in Section II-E, the differences between the MSE and MTE estimates are expected to be more prominent for the filterbank with larger bandwidth filters. Indeed for narrowband filters, as those employed in a 100-filter filterbank, the deviation differences become nontrivial only for the first and last few filter indices [see Fig. 3(d)–(f)]. For filters positioned in low frequencies, the difference is due to transient phenomena that are not fully averaged out. For wider filter passbands, the differences between the MSE and MTE deviations become significant, depending on the spectral shape of the signal and on the noise type.

Overall, the MTE estimates are significantly more robust, i.e., yield smaller deviation values than the MSE ones, when the major spectral energy content of noise is concentrated in lower frequencies compared to that of the speech signal, e.g., in the case of Volvo noise [see Fig. 3(b)]. Mixed results are obtained for other noise types (babble and white noise) as shown for the case of phoneme /aa/. In addition, transient phenomena play a key role, especially for the lower frequencies (or smaller filter indices) [21]. The MTE estimation scheme outperforms the MSE one for smaller filter indices, due to these transient phenomena. The difference in performance is more pronounced for wider filters and fricative sounds. In the cases detailed above, the MTE-based estimated deviations (from the clean energy coefficients) are presented significantly smaller than the respective MSE ones.

C. Cepstral Coefficient Deviations

Next, we compare the performance of MSE- and MTE-derived cepstral coefficients. These coefficients are estimated as

[11]The noise signals have a duration of approximately 235 s, so a portion of the noise signal is randomly selected and added to each speech signal. Their sampling frequency is 19.98 kHz.

[12]The SNR value is estimated as the mean ratio of the speech over the noise signal energies per frame. Then, the noise signals are scaled so that the global mean SNR is 5 dB. Therefore, this value refers to the wide-band speech signal and suggests that the SNR level is, on the average, 5 dB.

[13]The "$\wedge$" stands for mean estimates averaged over 1000 phoneme instances.

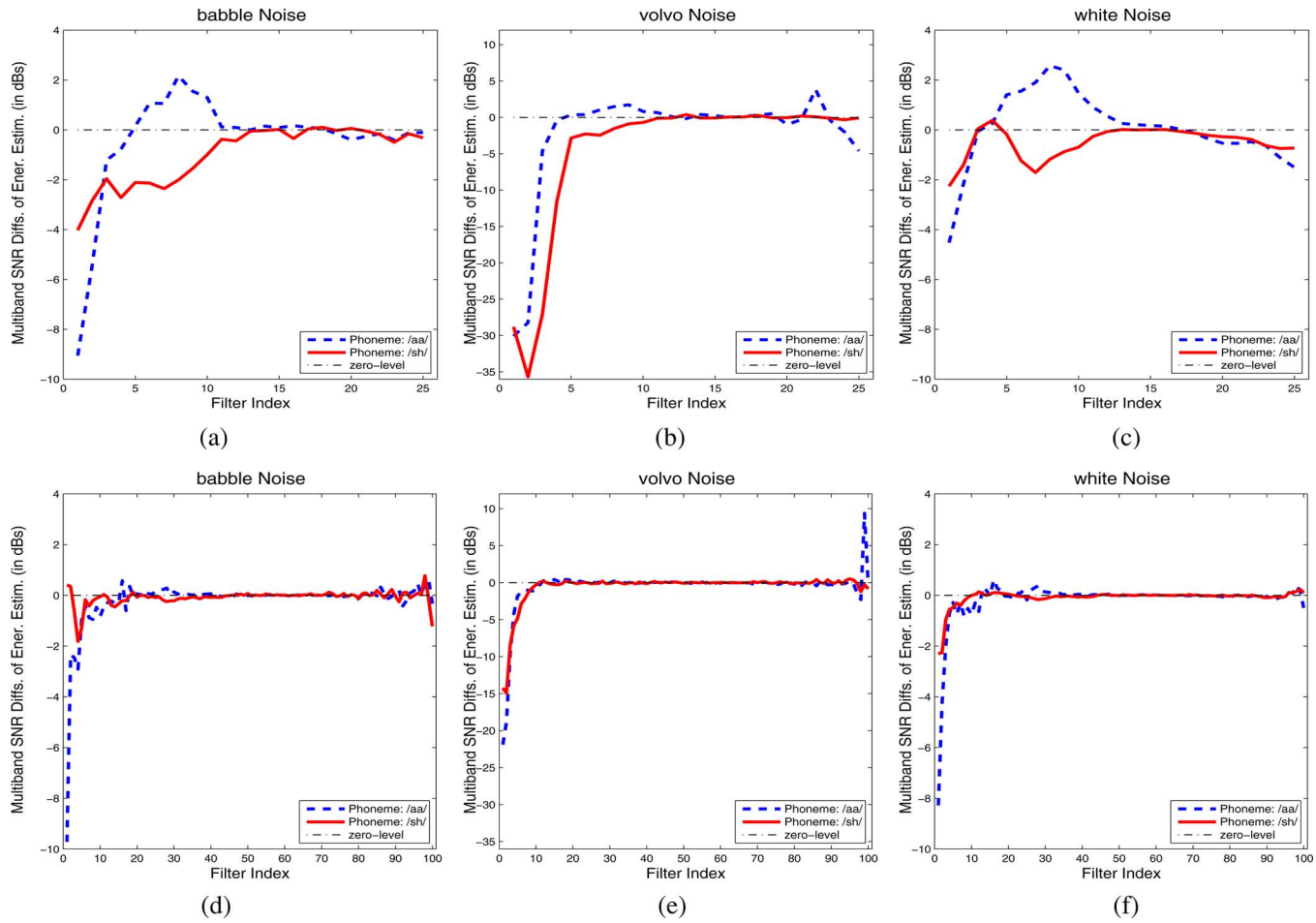

Fig. 3. Multiband SNR energy estimation differences for MSE and MTE schemes averaged over 1000 instances for the phonemes: /aa/ and /sh/, extracted from the TIMIT database, corrupted by babble (a), (d), car (b), (e) and white (c), (f) noises at $\mathrm{SNR} = 5$ dB (on average). The filterbanks consist of (a)–(c) 25 and (d)–(f) 100 Mel-spaced, Gammatone filters with fixed overlap of 50%. Positive values mean that the MSE scheme is more robust than the MTE one. Negative values indicate better performance of the MTE scheme.

discussed in Section III. One possible way to explore the features' robustness is to estimate the normalized mean cepstral feature deviation from the clean case (in dBs) as follows[14]:

$$\mathrm{Dev}C[i] \triangleq 20\log_{10}\left(\frac{\widehat{\Delta C}[i]}{\widehat{C}_r[i]}\right) \tag{25}$$

where $\widehat{\Delta C}[i]$, provided by (19), are the RMS differences between the noisy and the clean cepstral coefficients $\forall i \in \{1,\ldots,I\}$, normalized by the RMS values $\widehat{C}_r[i]$ of the clean ones. These deviations are indicative of how noise of different spectral characteristics affects the cepstral coefficients. Similar error analysis is, also, applied to the MFCCs (using a triangular filterbank) and is used as a baseline. The experimental setup remains the same as in the previous experiment, i.e., MSE- and MTE-based cepstral coefficients are computed for 1000 TIMIT instances of the phoneme /aa/ corrupted by additive noise, when filtered by mel-spaced Gammatone filterbanks with either 25 or 100 filters and fixed bandwidth overlap of 50%.

[14]The normalization scheme ensures that the coefficient magnitude range cannot affect the overall experimental results (lack of filter magnitude normalization can cause this mismatch across different filterbanks). Equation (25) is inspired by [10] and [11].

In Fig. 4, the normalized RMS cepstral deviations (in dBs) are presented as a function of the cepstral coefficient index for babble [Fig. 4(a) and (d)], car [Fig. 4(b) and (e)], and white noises [Fig. 4(c) and (f)]. The deviations of the MTE- and MSE-based features are, on average, smaller, outperforming the MFCC baseline. Further, MSE-based and MFCC features present very similar performance for some of the noise types. The differences are more pronounced when wider filters are employed (25-filters), as shown in Table I. As expected, the MTE-based features present smaller deviations than the MSE-based features for volvo noise, as shown in Fig. 4(e) and, especially, in Fig. 4(b). For babble and white noise, all three front-ends perform similarly. This is consistent with the Mel-energy coefficient deviations presented in the previous section. Similar results are also reached in the case of the MTE/MSE cepstral coefficient scheme for other phonemes. Concluding, we observe that the MTE-based features outperform, on average, all other studied features, i.e., MFCCs and MSE-based cepstral coefficients, for most phonemes and types of noise, see Table I. These differences are especially pronounced for low-pass noises, e.g., car (Volvo) noise. Finally, the proposed features present significantly smaller deviations w.r.t. the clean feature version, compared to the MFCC-based

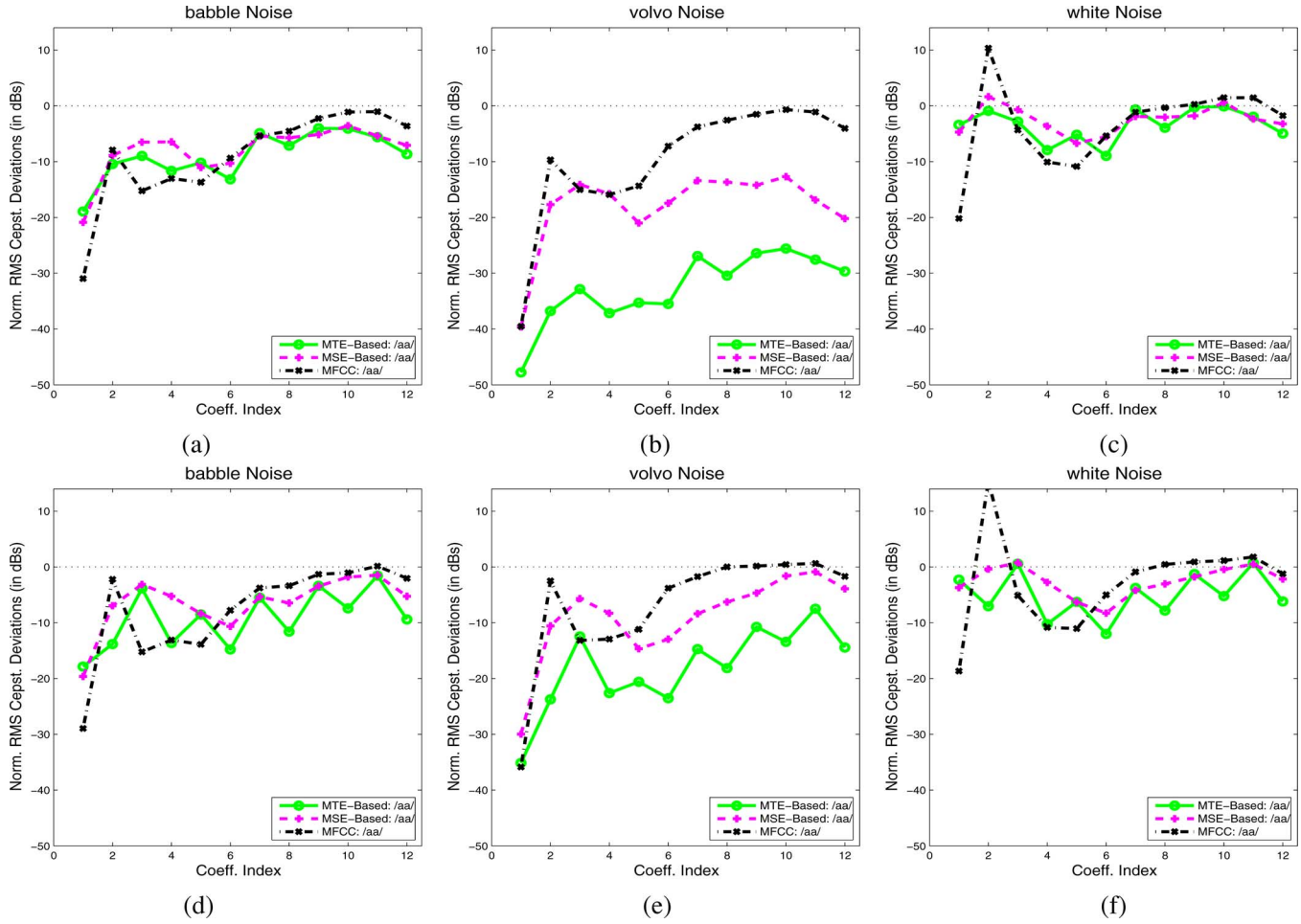

Fig. 4. Normalized RMS cepstral deviations (in dBs) computed over 1000 instances of the phoneme /aa/ extracted from the TIMIT database. Results shown as a function of coefficient index for babble [in (a) and (d)], car [in (b) and (e)], and white [in (c) and (f)] noise at an average $\mathrm{SNR} = 5$ dB. The filterbank consists of (a)–(c) 25 and (d)–(f) 100 Mel-spaced, Gammatone filters with fixed overlap of 50%. Smaller values indicate enhanced robustness in noise.

TABLE I
MEAN NORMALIZED DEVIATIONS (IN dB) FOR THREE FEATURE SETS: MFCC, MTE- AND MSE-BASED CEPSTRAL COEFFICIENTS FOR THREE NOISE SCENARIOS: BABBLE, CAR, AND WHITE NOISE. CEPSTRAL DEVIATIONS ARE ESTIMATED USING 25- AND 100-FILTER FILTERBANKS FOR 1000 INSTANCES OF THE PHONEMES /aa/ AND /sh/. SMALLER VALUES INDICATE ENHANCED ROBUSTNESS TO NOISE

Mean Normalized Cepstral Deviations (in dBs)						
	Cepstral Features	Noise Types			Aver.	Num. of Filts.
		Babble	Car	White		
Phone /aa/	MFCC	-0.01	1.16	4.42	1.86	25 Filt.
	MSE-	-5.80	-13.52	-1.62	-6.98	
	MTE-	**-7.08**	**-28.62**	**-2.58**	**-12.76**	
	MFCC	0.84	2.66	6.31	3.27	100 Filt.
	MSE-	-2.47	-2.30	-0.38	-1.72	
	MTE-	**-4.74**	**-10.93**	**-2.28**	**-5.98**	
Phone /sh/	MFCC	9.01	7.56	4.72	7.10	25 Filt.
	MSE-	**4.55**	8.70	**1.51**	4.92	
	MTE-	5.69	**3.88**	2.36	**3.98**	
	MFCC	8.01	6.76	3.61	6.13	100 Filt.
	MSE-	3.11	9.01	-0.22	3.97	
	MTE-	**0.75**	**4.75**	**-2.71**	**0.93**	

deviation values, according to Table I, providing additional robustness to the feature extraction process.

D. Speech Recognition Experiments

Next, speech recognition performance is evaluated when the following parameters vary: filter shape, energy scheme, number and bandwidth of the filters. Word and phone error rates are estimated for various types and levels of noise, i.e., the Aurora-3 (Spanish Task), Aurora-4, and the TIMIT + Noise databases, respectively. The results are presented as a function of the first filter ERB value and the total number of filters (the filter bandwidth overlap percent is a dependent parameter taking values between 30%–85%). For example, for the left-most Fig. 6(a), the first filter ERB takes values between 22–44 Hz that correspond to ERB overlap (with the adjacent filters) percent of 30%–85%. The ERB overlap percent is fixed across all filters of the filterbank. In the case of the 100-filter filterbanks in Figs. 5, 6(c), the filter ERB values are set proportional to those of the 25- and 50-filter filterbanks (when examining their first filters and the ERB overlap percent ranges in 30%–85%). Results (word accuracy) for the Aurora-3 database are shown in Fig. 5 for Gammatone filterbanks and for MSE/MTE estimation.[15] Further, results (phone accuracy) for

[15]The results for the word-level LV-ASR task (Aurora-4) appear to be similar to those of the Aurora-3 task and are omitted due to lack of space.

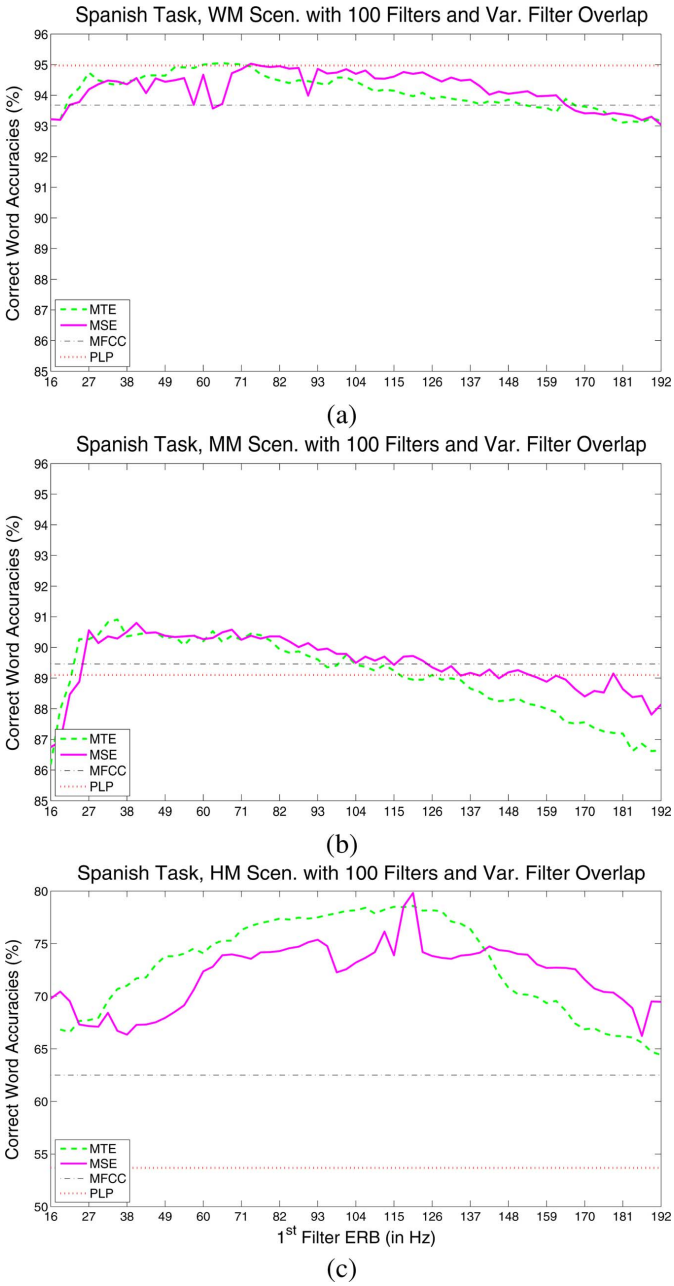

Fig. 5. Word accuracy for the cepstral MTE and MSE-based features (using CMS) for 100 Gammatone filters, in the Aurora-3 Spanish database. The horizontal axis displays the *ERB Values* of the first filter. These values are equalized (sequentially) to the first filter ERB values of the 25- and 50-filter filterbanks (when the filter overlap percent ranges in 30%–85%). Results for three training/testing mismatched scenarios are shown: (a) High mismatch (HM), (b) Medium mismatch (MM), and (c) Well matched (WM). The baseline MFCC and PLP results are shown as dashed lines.

the TIMIT + Noise database are shown in Fig. 6 for Gabor and Gammatone filterbanks, and MSE/MTE estimation. Finally, the PLP [16] and MFCC [20] features (extracted using the HTK platform [35]) provide the baseline performance. All features are normalized after removing their long-term cepstral means (CMS). Plots in Fig. 5 have different y-axis ranges to further enhance their readability.

According to the experimental results, moderate filter bandwidths, i.e., the middle-part of the graphs in Fig. 5 and the middle column in Fig. 6 seem to be more robust to different training/testing mismatches and yield the higher recognition rates across all noise scenarios. For the case of low and medium mismatch between training and testing conditions, i.e., the WM and MM scenarios, the MTE- and MSE-based features appear to always outperform the baseline MFCC features, providing enhanced immunity to noise. Both features perform similarly for reasonable values of the filter bandwidths. However, for the high mismatch task (HM), the performance of the MSE and MTE front-ends diverge significantly, especially when wider filters are employed (the right-most part of the plots or when 25-filter filterbanks are employed, Table II). The MSE-based features present an additional 12% relative improvement (for moderate filter bandwidths) compared to the MTE-based features and 30% improvement when compared to the baseline results (obtained by the ETSI WI007 front-end) Table II. These improvements are reached when the filter bandwidths assume reasonable values, i.e., the bandwidth of the first filter is less than 130 Hz. As detailed above, increased filter bandwidths lead to differences between the two energy estimation schemes. In the case where the filter bandwidths (as in the right-most part of the plots) take very large values,[16] the MSE-based features outperform the MTE-based ones due to the presence of the high-frequency noise components in the low SNR conditions, as shown in Fig. 2. On the other hand, ASR performance for both features (MTE- and MSE-based features) is similar, on average, for the case of narrow filters (the case of 100-filter filterbanks; see Table II).

Next, in Fig. 6, the performance of MSE/MTE is investigated as a function of both the number of employed filters and their shapes in the TIMIT + Noise task. The filter shape does not significantly affect the ASR performance, provided that the corresponding ERB bandwidths are normalized, comparing the plots in Fig. 6(a)–(b). It, also, appears that the number of filters employed is not an important factor, as well; similar results are obtained for different filterbanks employing 25–100 filters,

TABLE II
WORD ACCURACY (%) ON THE AURORA-3 (SPANISH TASK) DATABASE USING HTK. THE FILTERBANKS CONSIST OF 25- OR 100-FILTER GAMMATONE FILTERS. RESULTS FOR FOUR FEATURE SETS ARE PRESENTED: MFCC (BASELINE), PLP, MTE- AND MSE-BASED CEPSTRAL COEFS. IMPROVEMENT RELATIVE TO MFCC (WITH CMS) BASELINE

Word Accuracy (%) of Aurora-3, Spanish Task					
Features \ Scenario	WM	MM	HM	Aver.	Rel. Impr.
Aurora Frontend (WI007)	92.94	80.31	51.55	74.93	-37.75
MFCC†	93.68	89.46	62.50	81.80	-
PLP†	94.97	89.10	53.68	79.25	-14.01
MSE-Based† (25 Filt.)	94.20	89.52	78.05	87.26	30.00
MSE-Based† (100 Filt.)	94.65	90.95	70.29	85.30	19.23
MTE-Based† (25 Filt.)	94.22	89.21	71.94	85.12	18.24
MTE-Based† (100 Filt.)	94.75	90.89	71.76	85.80	20.33
† Features are Normalized using Cepstral Mean Subtraction (CMS)					

[16]The first filter bandwidth in the filterbank takes values greater than 140 Hz.

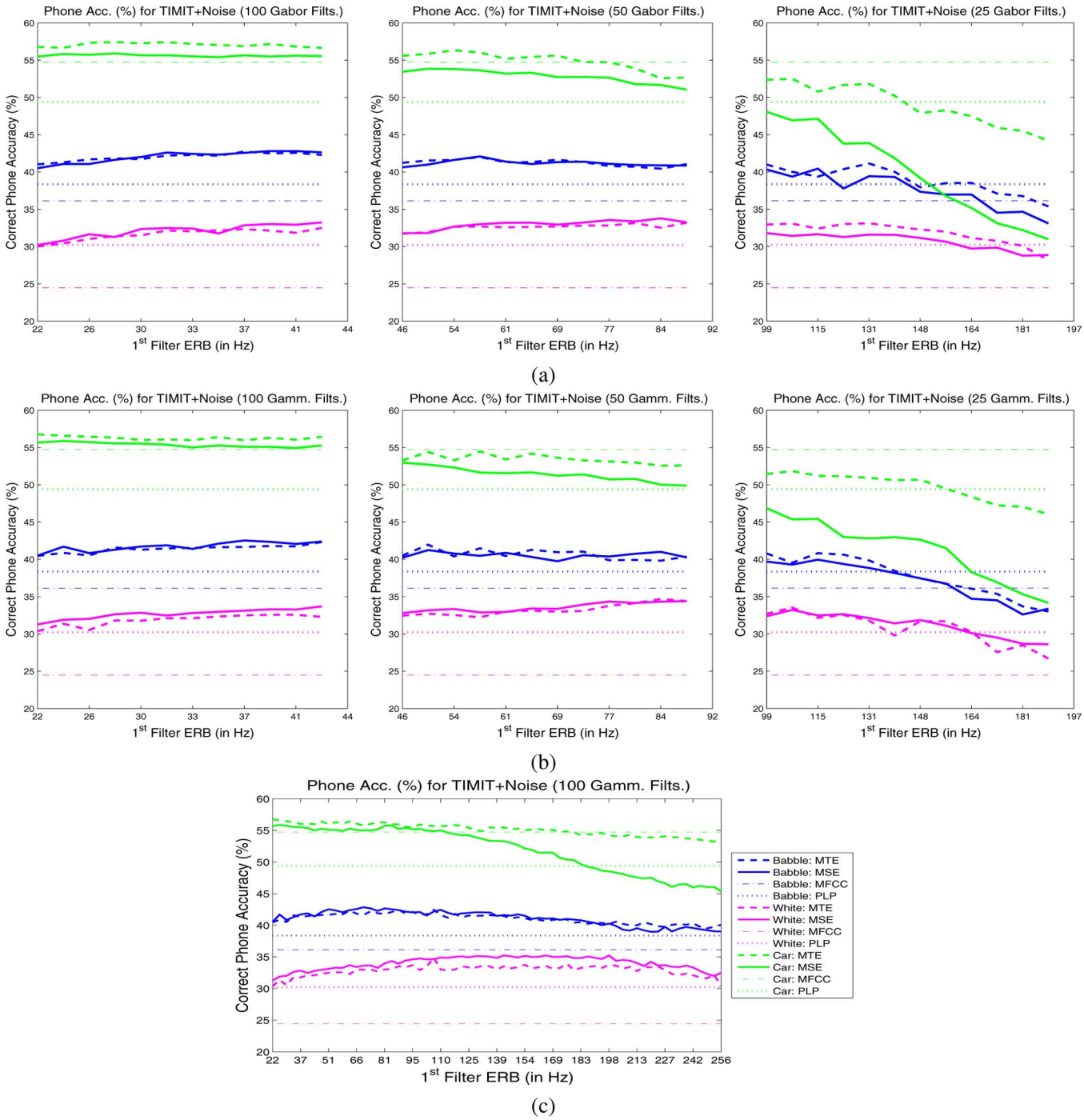

Fig. 6. Phone accuracy for cepstral features based on MTE/MSE schemes tested on the TIMIT + Noise database (with CMS). Three different scenarios are investigated: (a) Gabor Mel-Filterbank and (b) Gammatone Mel-Filterbank (for both cases the number of filters is 100, 50, or 25 for each of the three columns, respectively). The horizontal axis displays the *ERB Value* of the first filter, for ERB overlap ranging in 30%–85%. (c) Mel-spaced, 100-Filter Gammatone filterbank and ERB values proportional to those of the 25- and 50-filter Filterbanks (when the ERB percent ranges in 30%–85%). Results are shown for three noise types: babble, white, and Volvo (car).

when the corresponding filter bandwidths are equalized,[17] Fig. 6(b)–(c). Examining the relative performance of the MSE- and MTE-based features, the MTE clearly outperforms the other features for the case of Volvo (car) noise, especially when filters present large ERB bandwidths. For other noise types, the MSE- and MTE-based features display similar performance, Table III. The differences in performance are more pronounced in the case of wide filters, e.g., when using a 25-filter filterbank.

Overall, if we fix the energy estimation scheme, the parameter that mainly affects ASR performance is the filter bandwidth, rather than the bandwidth overlap percentage[18] or the shape of the filters (as long as their ERBs are normalized). There is also

[17]The overlap percentage has been altered accordingly to ensure wider filters in the case of the 100-filter filterbank. The first filter ERB values are equalized (sequentially) to the first filter ERB values of the 25- and 50-filter filterbanks (when the filter overlap percent ranges in 30%–85%), and the rest of the ERB values are increased proportionally.

[18]Note that the range of overlap remains the same for the 25, 50, 100 filter experiments, ranging from 35%–85%; see Fig. 6.

TABLE III
PHONE ACCURACIES (%) ON THE TIMIT + Noise (ADDITIVE BABBLE, WHITE OR CAR NOISES) DATABASE. THE FILTERBANKS CONSIST OF 25- OR 100-FILTER GAMMATONE FILTERS. IMPROVEMENT IS SHOWN RELATIVE TO THE MFCC (WITH CMS) BASELINE

Phone Accuracy (%) of TIMIT+Noise Task					
Scenario / Features	Babble	White	Car	Aver.	Rel. Impr.
MFCC	36.14	24.48	54.73	38.45	-
PLP	38.36	30.23	49.39	39.33	2.29
MSE-Based (25 Filt.)	39.95	33.22	46.86	40.01	4.06
MSE-Based (100 Filt.)	42.53	33.69	55.88	44.03	14.51
MTE-Based (25 Filt.)	40.83	33.52	51.84	42.06	9.38
MTE-Based (100 Filt.)	42.33	32.57	56.76	43.89	14.15
All Features are Normalized using Cepstral Mean Subtraction (CMS)					

a relatively wide range of filter ERBs (from approx. 50 Hz to 120 Hz) where good ASR performance is achieved. Thus, the word error rates seem to mostly depend on the ERB values, exhibiting a stable performance for a wide range of ERB values. Similar results were obtained when additive noise was added to the noise-corrupted TIMIT database.

VI. CONCLUSION—DISCUSSION

We have investigated four key parameters in the feature extraction process, namely: filter bandwidth, filter bandwidth overlap, number of filters, and the energy computation scheme. We have also examined their impact on ASR performance for three different recognition experiments. The presented results are supported by a theoretical analysis of the cepstral coefficients estimation error in noise. Overall, the equivalent rectangular filter bandwidths and the energy estimation scheme appear to be two of the most significant parameters determining ASR performance. According to the presented findings, ASR performance can be predicted for a particular choice of filter bandwidth range and energy estimation scheme when the relative spectral energy distributions of signal and noise are considered.

In more detail, the performance of the averaged energy estimation scheme is mainly a function of the relative spectral energy content of the noise versus the speech input signal, when examined within the filter passbands. The proposed generalized cepstral features are directly related to these energy distributions. Therefore, it is of great importance to ensure a robust and efficient energy computation process. Energy estimation errors propagate to the cepstral coefficients, as well. The proposed noisy cepstral coefficient deviations (deviations from the clean case) are, on average (RMS values), smaller than those of the MFCCs. This is due to the energy scheme and the wider filters employed.

In this context, it is shown that features using filters of different spectral shape present similar performance when their effective filter bandwidths are kept equal, regardless of their design parameters, for low and medium mismatch training/testing scenarios. For high mismatch, the energy computation scheme is usually the most important factor affecting performance; the signal versus noise spectral content should be first analyzed, selecting the most appropriate energy computation scheme.

Finally, similar trends and conclusions can be drawn when advanced signal denoising and feature equalization techniques are applied in combination with the feature extraction scheme, as shown in [36]. There, the performance improvements appear to be additive on top of the signal and feature enhancement techniques, such as Wiener filtering and parameter equalization (PEQ). This is particularly important in building robust ASR systems.

In future work, we plan to extend our work to the design of filterbanks that optimize ASR performance under adverse recording conditions and under time-varying noise.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and for bringing to our attention the Ph.D. dissertation of Dr. P. J. Moreno [9].

REFERENCES

[1] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition," in *Proc. Eurospeech'05*, Sep. 2005.
[2] A. de la Torre, J. C. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proc. ICASSP'02*, Apr. 2002, pp. 401–404.
[3] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.
[4] Y. M. Cheng and D. O'Shaugnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 1943–1954, Sep. 1991.
[5] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *Proc. ASRU'09*, Dec. 2009.
[6] Z. Tufekci, "Convolutional bias removal based on normalizing the filterbank magnitude," *IEEE Signal Process. Lett.*, vol. 15, no. 7, pp. 485–488, Jul. 2007.
[7] C. P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.
[8] C. Yang, F. K. Soong, and T. Lee, "Static and dynamic spectral features: Their noise robustness and optimal weights for ASR," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1087–1097, Mar. 2007.
[9] P. J. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie Mellon Univ., Pittsburgh, PA, 1996.
[10] F.-H. Liu, R. M. Stern, A. Acero, and P. J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *Proc. ICASSP'94*, Apr. 1994, pp. 61–64.
[11] B. Raj, E. B. Gouvea, P. J. Moreno, and R. M. Stern, "Cepstral compensation by polynomial approximation for environment-independent speech recognition," in *Proc. ICSLP'96*, Sep. 1996, pp. 2340–2343.
[12] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.
[13] J. W. Pitton, K. Wang, and B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proc. IEEE*, vol. 84, no. 9, pp. 1199–1215, Sep. 1996.
[14] A. Biem, S. Katagiri, E. McDermott, and B. H. Juang, "An application of discriminative feature extraction to filterbank-based speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 96–110, Feb. 2001.
[15] J. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, Oct. 1994.
[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
[17] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[18] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 258–261, Feb. 2004.
[19] B. Mak, Y. Cheung-Tam, and Q. Li, "Discriminative auditory-based features for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 27–36, Jan. 2004.

[20] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 3, pp. 357–366, Aug. 1980.
[21] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and Teager–Kaiser operators for short-term energy estimation in additive noise," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2569–2581, Jul. 2009.
[22] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The Gammachirp," *J. Acoust. Soc. Amer.*, vol. 101, pp. 412–419, 1997.
[23] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Comput. Tech. Rep. #35, 1993.
[24] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[25] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 3, pp. 196–200, Mar. 2001.
[26] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.
[27] M. L. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust recognition," in *Proc. Eurospeech'03*, Sep. 2003.
[28] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, 1946.
[29] M. D. Skowronski and J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 116, pp. 1774–1780, Sep. 2004.
[30] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.
[31] A. C. Bovik, J. P. Havlicek, M. D. Desai, and D. S. Harding, "Limits on discrete modulated signals," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 867–879, Apr. 1997.
[32] J. P. Havlicek, D. S. Harding, and A. C. Bovik, "Multidimensional quasi-eigenfunction approximations and multicomponent AM-FM models," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 227–242, Feb. 2000.
[33] L. Rigazio, P. Nguyen, D. Kryze, and J. C. Junqua, "Large vocabulary noise robustness on aurora-4," in *Proc. Eurospeech'03*, Sep. 2003.
[34] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.
[35] S. Young *et al., The HTK Book (for HTK Version 3.2)*. Cambridge, U.K.: Cambridge Research Lab: Entropics, 2002.
[36] D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, and V. Pitsikalis, "Robust features for speech recognition in extremely adverse environments," in *Proc. ICSLP'07*, Sep. 2007.

Dimitrios Dimitriadis (S'99–M'06) received the Diploma in electrical and computer engineering and the Ph.D. degree both from the National Technical University of Athens, Athens, Greece, in 1999 and 2005, respectively.

From 2001 to 2002, he was an Intern at the Multimedia Communications Lab, Bell Labs, Lucent Technologies, Murray Hill, NJ. From 2005 to 2009, he was a Postdoctoral Research Associate at the National Technical University of Athens. He is now Principal Member of Technical Staff with the Networking and Services Research Laboratory, AT&T Labs, Florham Park, NJ. His current research interests include speech processing, analysis, synthesis and recognition, multi-modal systems, and nonlinear and multi-sensor signal processing. He has authored or coauthored over 15 papers in professional journals and conferences.

Dr. Dimitriadis has been a member of the IEEE Signal Processing Society (SPS) since 1999 and he has served as a reviewer for the IEEE SPS.

Petros Maragos (S'81–M'85–SM'91–F'96) received the electrical engineering Diploma from the National Technical University of Athens (NTUA), Athens, Greece, in 1980, and the M.Sc.E.E. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1982 and 1985, respectively.

From 1985 to 1993, he was an Electrical Engineering Professor at the Division of Applied Sciences, Harvard University, Cambridge, MA. In 1993, he joined the electrical and computer engineering faculty at Georgia Tech. During parts of 1996–1998, he was on sabbatical working as Director of Research at the Institute for Language and Speech Processing, Athens. Since 1998, he has been working as an Electrical and Computer Engineering Professor at NTUA. His research and teaching interests include signal processing, systems theory, pattern recognition, and applications to image processing and computer vision, speech and language processing, multimedia, cognition, and robotics.

Prof. Maragos received a 1987 NSF Presidential Young Investigator Award, a 1988 IEEE SP Society's Young Author Paper Award, a 1994 IEEE SP Senior Award, the 1995 IEEE W. R. G. Baker Prize Award, a 1996 Pattern Recognition Society's Honorable Mention Award, the 2007 EURASIP Technical Achievements Award, and 2010 EURASIP Fellow.

Alexandros Potamianos (M'92–SM'10) received the Diploma in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1990 and the M.S. and Ph.D. degrees in engineering sciences from Harvard University, Cambridge, MA, in 1991 and 1995, respectively.

From 1991 to June 1993, he was a Research Assistant at the Harvard Robotics Lab., Harvard University. From 1993 to 1995, he was a Research Assistant at the Digital Signal Processing Lab, the Georgia Institute of Technology, Atlanta. From 1995 to 1999, he was a Senior Technical Staff Member at the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. From 1999 to 2002, he was a Technical Staff Member and Technical Supervisor at the Multimedia Communications Lab at Bell Labs, Lucent Technologies, Murray Hill, NJ. From 1999 to 2001, he was an Adjunct Assistant Professor at the Department of Electrical Engineering of Columbia University, New York. In the spring of 2003, he joined the Department of Electronics and Computer Engineering, Technical University of Crete, Chania, Greece, as an Associate Professor. His current research interests include speech processing, analysis, synthesis and recognition, dialog and multi-modal systems, nonlinear signal processing, natural language understanding, artificial intelligence, and multimodal child–computer interaction. He has authored or coauthored over 80 papers in professional journals and conferences. He is the coeditor of the book "*Multimodal Processing and Interaction: Audio, Video, Text*. He holds four patents.

Prof. Potamianos is the coauthor of the paper "Creating conversational interfaces for children" that received a 2005 IEEE Signal Processing Society Best Paper Award, He has been a member of the IEEE Signal Processing Society since 1992 and he is currently serving his second term at the IEEE Speech Technical Committee.