Histogram Equalization of Speech Representation for Robust Speech Recognition

Ángel de la Torre, Antonio M. Peinado, *Member, IEEE*, José C. Segura, *Senior Member, IEEE*,
José L. Pérez-Córdoba, *Member, IEEE*, Ma Carmen Benítez, *Member, IEEE*, and
Antonio J. Rubio, *Senior Member, IEEE*

***Abstract*—This paper describes a method of compensating for nonlinear distortions in speech representation caused by noise. The method described here is based on the histogram equalization method often used in digital image processing. Histogram equalization is applied to each component of the feature vector in order to improve the robustness of speech recognition systems. The paper describes how the proposed method can be applied to robust speech recognition and it is compared with other compensation techniques. The recognition experiments, including results in the AURORA II framework, demonstrate the effectiveness of histogram equalization when it is applied either alone or in combination with other compensation techniques.**

***Index Terms*—Cepstral mean normalization, histogram equalization, mean and variance normalization, Mel frequency cepstral coefficients, probability density function (pdf), robust speech recognition, vector Taylor series approach.**

I. Introduction

Noise strongly degrades the performance of speech recognition systems. For this reason, robust speech recognition is one of the focus areas of speech research [1]–[5]. Noise has two main effects on speech representation. First, it introduces a distortion in the feature space. Secondly, due to its random nature, it also causes a loss of information [6], [7]. The effect of the distortion depends on the speech representation and the type of noise, and it usually produces a nonlinear transformation of the feature space. For example, in the case of cepstral-based representations, additive noise causes a nonlinear transformation that has no significant effect on high-energy frames but a strong effect on those with energy levels in the same range or below that of the noise [6]–[9]. This distortion causes a mismatch between the training and recognition conditions. The acoustic models trained with speech acquired under clean conditions do not model speech acquired under noisy conditions accurately and this degrades the performance of speech recognizers. Compensation methods for robust speech recognition mainly focus on minimizing this mismatch. Some methods try to adapt the acoustic models to noisy conditions in order to allow them to represent noisy speech properly, whereas other methods try to determine the features of the clean speech from the observed noisy speech. In the former case, the noisy speech is recognized using noisy models. In the latter case, a clean version of the speech is recognized using the clean models. Finally, some methods include operations in the feature extraction module in order to minimize the effect of noise superimposed on the speech representation [1], [10], [11].

Thus, for example, cepstral mean normalization (CMN) [12] is usually applied as a part of the feature extraction in order to remove the global shift of the mean affecting the cepstral vectors. This normalization compensates for the main effect of channel distortion and some of the side effects of additive noise. However, the nonlinear effects of additive noise on cepstral-based representations cannot be treated by CMN and this makes this method effective only for moderate levels of additive noise. This method is improved by mean and variance normalization (MVN) [13] because normalization of the mean and the variance yields a better compensation of the mismatch caused by additive noise.

Other methods, such as spectral subtraction [14] or the vector Taylor series (VTS) approach [6], [8], [15]–[18] yield more effective compensation of additive noise since they can deal with the nonlinear effects of noise. Robust methods based on the adaptation of acoustic models include Statistical Re-estimation [6], [8], [19] and parallel model combination [9], [20], which apply independent corrections to each Gaussian pdf in the acoustic models. These are able to model nonlinear effects of the distortion caused by noise correctly. Most compensation methods are based on estimations of convolutional and additive noise and a statistical or analytical formulation describing the effect of noise superimposed on the speech representation.

This paper describes a method of compensating for the noise affecting speech representation. The method is based on the histogram equalization (HEQ) technique, which is often used in digital image processing [21], [22]. This technique has been adapted here for use with speech representation. This method provides a transformation mapping the histogram of each component of the feature vector onto a reference histogram. This compensates for the effect of noise processes distorting the feature space. The effectiveness of the method relies on estimating the histograms of the speech to be compensated correctly and the assumption that the effect of the noise distortion is a monotonic transformation of the representation space. The first assumption makes the method more effective as more speech frames are involved in estimating the histograms. Generally, the second assumption cannot be verified due to the random behavior of the noise process. The effect of the

Manuscript received October 30, 2001; revised January 30, 2004. This work was supported in part by the Spanish Government under CICYT Project TIC99-0583. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

The authors are with Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada, 18071 Granada, Spain (e-mail: atv@ugr.es).

Digital Object Identifier 10.1109/TSA.2005.845805

1063-6676/$20.00 © 2005 IEEE

noise can be considered to be a random transformation of the feature space. If both processes are considered separately, the transformation causes a mismatch between training and recognition conditions while the random behavior causes an irreversible loss of information. HEQ is able to compensate for the transformation causing the mismatch, but (like all other compensation methods) it is not able to compensate for the effect introduced by the random behavior of the noise. An important difference between the HEQ method and most of the other compensation methods is the fact that no analytic assumptions are made about the noise process or the way the noise affects the speech representation. Therefore, the HEQ method would be able to compensate for a wide range of noise processes affecting a wide variety of parameterizations of the speech signal.

This paper reviews the HEQ method and describes how it can be adapted to compensate for the noise superimposed on the speech representation. We have studied the relationship between HEQ and other compensation methods such as CMN and MVN, and more complex methods such as VTS. We have also proposed a combination of VTS and HEQ. The different noise compensation methods were evaluated using automatic speech recognition experiments performed under a variety of different noise conditions. The experimental results reveal the usefulness and limitations of HEQ for speech recognition in the presence of noise.

II. Histogram Equalization for Robust Speech Recognition

A. Review of the Histogram Equalization Procedures

HEQ was originated as a technique for digital image processing [21], [22]. Its aim is to provide a transformation $x_1 = F(x_0)$ that converts the probability density function $p_0(x_0)$ of the original variable into a reference probability density function $p_1(x_1) = p_{ref}(x_1)$. The transformation therefore converts the histogram of the original variable into the reference histogram, i.e., it equalizes the histogram. The formulation of the method is described below.

Let x_0 be a unidimensional variable following a distribution $p_0(x_0)$. A transformation $x_1 = F(x_0)$ modifies the probability distribution according to the expression

$$p_1(x_1) = p_0(G(x_1))\frac{\partial G(x_1)}{\partial x_1} \tag{1}$$

where $G(x_1)$ is the inverse transformation of $F(x_0)$. The relationship between the cumulative probabilities associated with these probability distributions is given by

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' \\ &= \int_{-\infty}^{F(x_0)} p_0(G(x_1'))\frac{\partial G(x_1)}{\partial x_1'}dx_1' \\ &= \int_{-\infty}^{F(x_0)} p_1(x_1')dx_1' \\ &= C_1(F(x_0)) \end{aligned} \tag{2}$$

and therefore, the transformation $x_1 = F(x_0)$, which converts the distribution $p_0(x_0)$ into the reference distribution $p_1(x_1) = p_{ref}(x_1)$ (and hence converts the cumulative probability $C_0(x_0)$ into $C_1(x_1) = C_{ref}(x_1)$), is obtained from (2) as

$$x_1 = F(x_0) = C_1^{-1}[C_0(x_0)] = C_{ref}^{-1}[C_0(x_0)] \tag{3}$$

where $C_{ref}^{-1}[C]$ is the inverse function of the cumulative probability $C_{ref}(x_1)$, providing the value x_1 that corresponds to a certain cumulative probability C. For practical implementations, a finite number of observations are considered and therefore cumulative histograms are used instead of cumulative probabilities. For this reason the procedure is referred to as histogram equalization rather than probability distribution equalization.

The HEQ method is frequently applied in digital image processing as a means of improving the brightness and contrast of digital images and to optimize the dynamic range of the grey-level scale. HEQ is a simple and effective method for automatically correcting images that are either too bright or too dark or that have a poor contrast.

B. Application of Histogram Equalization to the Speech Representation

HEQ allows accurate compensation of the effect of any nonlinear transformation of the feature space provided that 1) the transformation is monotonic (and hence does not cause an information loss) and 2) there are sufficient observations of the signal being compensated to allow an accurate estimate of the original probability distribution.

In the case of digital image processing, the brightness and contrast alterations are mainly due to incorrect lighting or nonlinearities in the receptors. These usually correspond to monotonic nonlinear transformations of the grey-level scale. On the other hand, an image typically contains from several thousand to several million pixels. All of them contribute to an accurate estimation of the original probability distributions. This makes HEQ very effective for image processing.

In the case of automatic speech recognition, the speech signal is segmented into frames, with a frame period of about 10 ms, and each frame is represented by a feature vector. The number of observations for the estimation of the histograms is much smaller than in the case of image processing (typically several hundred frames per sentence) and also an independent HEQ procedure needs to be applied to each component of the feature vector. If the method is applied for noise compensation, it should be borne in mind that the more frames that are considered when estimating histograms, the more accurate the transformation obtained for the noise compensation will be. Additionally, HEQ is intended to correct monotonic transformations but the random behavior of the noise makes the transformation nonmonotonic, resulting in a loss of information in addition to the mismatch. Therefore, like other noise compensation methods, HEQ can deal with the mismatch caused by the noise but not with the loss of information caused by the random behavior of the noise. This limits the effectiveness of HEQ based noise compensation.

We applied HEQ to each component of the feature vector representing each frame of the speech signal. In order to obtain the transformation for each component, the cumulative histogram was estimated by considering 100 uniform intervals between $\mu_i - 4\sigma_i$ and $\mu_i + 4\sigma_i$, where μ_i and σ_i are the mean and the standard deviation for the i^{th} component of the feature vector, respectively. The transformation was computed according to (3) for the points in the center of each interval and was applied to the parameters to be compensated as a linear interpolation using the closest pair of points for which the transformation was computed. Original histograms were estimated using the frames of each utterance. Thus, the HEQ method was applied on a sentence-by-sentence basis. The speech representation used was based on the Mel Frequency Cesptral Coefficients (MFCC) [23], [12], and included the logarithm of the energy, the cepstral coefficients and the first and second associated regression coefficients. A Gaussian probability distribution with zero mean and unity variance was used as the reference probability distribution for each component. HEQ was applied as a part of the speech signal parameterization process both during training of the acoustic models and during the recognition process.

Fig. 1 shows how the HEQ method compensates bfor the effect of noise on the speech representation. In this case, We contaminated the speech signal with additive Gaussian white noise at SNRs ranging from 60 dB to 5 dB. The figure shows the effect of the noise and HEQ on the energy coefficient and the 3rd cepstral coefficient. The plots in the first row show the original probability distributions[1] for these components and for the different noise levels. As may bee seen, the noise severely affects the probability distributions of the speech causing a considerable mismatch when the training and recognition SNRs differ. The plots in the second row show how these coefficients change over time. The speech signal corresponds to the pronunciation of the Spanish digit string “8089.” Again, the mismatch caused by noise can be observed. The plots in the third row show the transformations obtained in each case in order to convert the original histograms into the reference histogram, according to the procedure described above. The histograms of the transformed speech representation are shown in the following plots and, as may be observed, they approximate to the reference Gaussian probability distribution. Finally, the last plots show how the equalized components change over time. In this case, the mismatch caused by the noise is significantly reduced. However, HEQ cannot remove completely the noise effect due to its randomness. Similar plots would be observed for the other components.

C. Relationship Between Histogram Equalization and Other Methods

HEQ can be considered an extension of other well known methods, such as Cepstral Mean Normalization [12] or mean and variance normalization (MVN) [13]. CMN is usually applied in MFCC-based parameterizations and compensates for the global shift of the probability distributions caused by the presence of noise. In order to apply CMN, the mean μ_0 of the variable x_0 is estimated, and the compensated variable x_1 is computed as

$$x_1 = F(x_0) = x_0 - \mu_0. \tag{4}$$

CMN therefore makes the mean of the compensated variable x_1 zero and so equalizes the first moment of its probability distribution.

The MVN method equalizes the first two moments of the distribution (i.e., the mean and the variance) by applying the linear transformation

$$x_1 = F(x_0) = \frac{x_0 - \mu_0}{\sigma_0} \tag{5}$$

where μ_0 and σ_0 are the mean and the standard deviation of the variable x_0, respectively. After applying MVN, the equalized variable x_1 has zero mean and unity variance.

Both CMN and MVN are a useful means of compensating for channel distortion and also for some effects of additive noise. However, they only provide linear transformations of the original variable. Due to the nonlinear nature of the distortion caused by additive noise in the cepstral domain, they suffer from limitations when are used to compensate for the effect of additive noise.

The HEQ procedure described here equalizes all the moments of the probability distribution to those of the reference probability distribution. Therefore, this procedure can be considered to be an extension of CMN and MVN to all the moments of the pdf. In addition, it allows the estimation of nonlinear transformations. This makes HEQ more appropriate than CMN or MVN when dealing with additive noise.

Fig. 2 shows the effect of the CMN and MVN procedures when they are applied to compensate for additive noise. These plots correspond to the same speech signal as in Fig. 1. The transformations and histograms of the normalized variable and the time course of the normalized variable are shown for the energy coefficient. On the left side, the effect of CMN is shown. The effect of MVN is shown on the right. We can observe that both CMN and MVN methods provide a limited compensation of the mismatch caused by additive noise. Compared with these linear methods, HEQ provides more appropriate transformations to reduce the noise mismatch.

III. Combination of Histogram Equalization With Other Methods

One of the particularities of HEQ is that its formulation does not rest on any assumptions about the speech representation or the process causing the distortion. Other methods for robust speech recognition are formulated taking into account the nature of the noise and the mechanisms affecting the speech representation in a given domain. One could expect that such methods provide a compensation of the noise effects that is more accurate than that provided by HEQ. This absence of assumptions could be considered a drawback of the proposed HEQ method.

However, because of this absence of assumptions, HEQ is able to deal with distortions coming from different processes. In particular, one could expect a compensation of the residual noise after applying other methods such as spectral subtraction, Wiener filtering or VTS. An additional improvement could be expected from compensation of this residual noise provided by HEQ.

[1] These probability distributions have been estimated by smoothing and normalizing the histograms obtained from 30 s of speech.

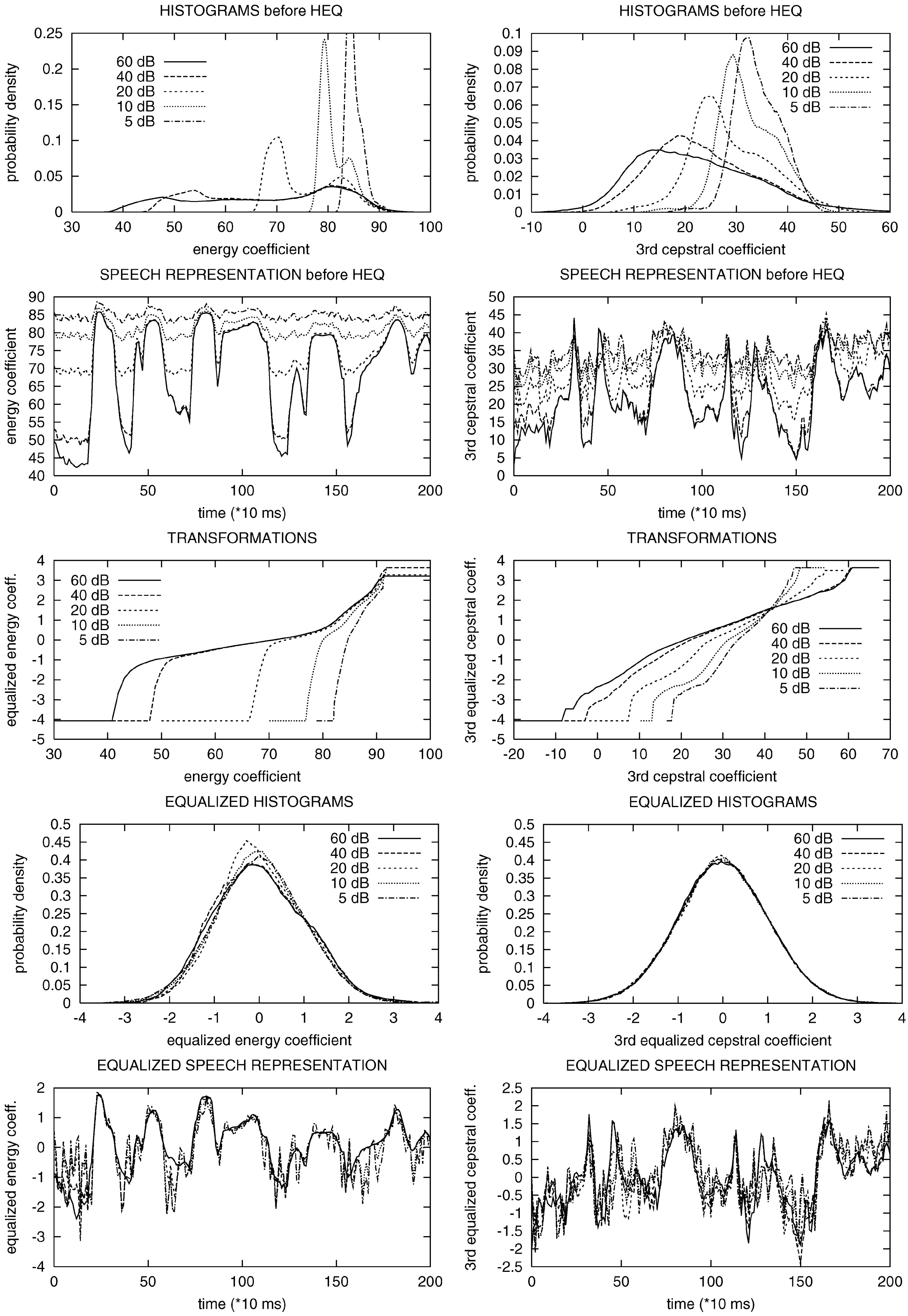

Fig. 1. Effect of HEQ over the speech representation for the energy coefficient (plots in the left side) and the third cepstral coefficient (in the right side).

In this paper we have considered the combination of HEQ with VTS. In the following sections we will review the VTS approach and propose a combination of VTS and HEQ as a way of improving the compensation provided by VTS.

A. Vector Taylor Series Approach

The VTS approach [6], [15], [8], [16]–[18] is a noise compensation method providing a clean speech representation by removing the additive and/or the convolutional noise. This noise compensation is performed in the logarithmically scaled filter-bank energy (log-FBE) domain. The method assumes a model describing the statistics of clean speech, and that the effect of the noise can be described as an additive term in the log-FBE domain

$$\mathbf{y}(\mathbf{x},\mathbf{n},\mathbf{h}) = \mathbf{x} + \mathbf{g}(\mathbf{x},\mathbf{n},\mathbf{h}) \tag{6}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors in the log-FBE domain representing the clean and noisy speech respectively, for a given frame, and $\mathbf{n}$ and $\mathbf{h}$ represent the additive noise and the channel distortion affecting this speech frame, respectively. For simplicity, we shall ignore channel distortion (more details can be found in the references concerning VTS). For the i^{th} channel, the relationship between the noisy speech, the clean speech and the additive noise can be written as

$$y(i) = x(i) + g(i) \tag{7}$$

where

$$g(i) = \alpha \log\left(1 + \exp\left(\frac{n(i)-x(i)}{\alpha}\right)\right) \tag{8}$$

and α is a constant whose value depends on the logarithmic compression applied to convert the filter-bank energies $X(i)$ into $x(i)$ in the log-FBE domain (if $x(i) = \log(X(i))$ then $\alpha = 1$; if $x(i) = 10\log_{10}(X(i))$ then $\alpha = 10\log_{10}(e) = 4.342\,944\,819\,032\,52$).

Two auxiliary functions $f(i)$ and $h(i)$ can be defined as

$$f(i) \equiv \frac{1}{1+\exp\left(\frac{x(i)-n(i)}{\alpha}\right)} \tag{9}$$

$$h(i) \equiv \frac{1}{\alpha}(1-f(i))f(i) \tag{10}$$

verifying that

$$\frac{\partial g(i)}{\partial x(j)} = -\frac{\partial g(i)}{\partial n(j)} = -f(i)\delta_{i,j} \tag{11}$$

$$\begin{aligned}\frac{\partial^2 g(i)}{\partial x(j)\partial x(k)} = \frac{\partial^2 g(i)}{\partial n(j)\partial n(k)} = -\frac{\partial^2 g(i)}{\partial x(j)\partial n(k)}\\ = h(i)\delta_{i,j}\delta_{i,k}\end{aligned} \tag{12}$$

where $\delta_{i,j}$ is the Kronecker's delta. The noisy speech $y(i)$ can be approached using a Taylor series around the values $x_0(i)$ and $n_0(i)$. The second-order approach is

$$\begin{aligned}y(i) \approx\, & x(i) + g_0(i) + f_0(i)[-(x(i)-x_0(i))\\ & + (n(i)-n_0(i))] + \frac{1}{2}h_0(i)[(x(i)-x_0(i))^2\\ & + (n(i)-n_0(i))^2 - 2(x(i)-x_0(i))(n(i)-n_0(i))]\end{aligned} \tag{13}$$

where $g_0(i)$, $f_0(i)$ and $h_0(i)$ are the functions $g(i)$, $f(i)$ and $h(i)$ evaluated at $x_0(i)$ and $n_0(i)$.

We can describe how a Gaussian pdf in the log-FBE domain is affected by additive noise using this Taylor series approach. Let us consider a Gaussian pdf representing clean speech, with mean $\mu_x(i)$ and covariance matrix $\Sigma_x(i,j)$ and let us assume a Gaussian noise process with mean $\mu_n(i)$ and covariance matrix $\Sigma_n(i,j)$. We can expand the Taylor series around $x_0(i) = \mu_x(i)$ and $n_0(i) = \mu_n(i)$. The mean and the covariance matrix of the pdf describing the noisy speech can be obtained as the expected values

$$\mu_y(i) = E[y(i)] \tag{14}$$

$$\Sigma_y(i,j) = E[(y(i)-\mu_y(i))(y(j)-\mu_y(j))] \tag{15}$$

and can be estimated as a function of $\mu_x(i)$, $\mu_n(i)$, $\Sigma_x(i,j)$ and $\Sigma_n(i,j)$ as

$$\mu_y(i) \approx \mu_x(i) + g_0(i) + \frac{1}{2}h_0(i)[\Sigma_x(i,i) + \Sigma_n(i,i)] \tag{16}$$

$$\begin{aligned}\Sigma_y(i,j) \approx\, & (1-f_0(i))(1-f_0(j))\Sigma_x(i,j)\\ & + f_0(i)f_0(j)\Sigma_n(i,j)\\ & + \frac{1}{2}h_0^2(i)(\Sigma_x(i,i)+\Sigma_n(i,i))^2\delta_{i,j}\end{aligned} \tag{17}$$

where $g_0(i)$, $f_0(i)$ and $h_0(i)$ are evaluated for $x_0(i) = \mu_x(i)$ and $n_0(i) = \mu_n(i)$. Thus, the Taylor series approach gives a Gaussian pdf describing the noisy speech from the Gaussian pdf describing the clean speech and the Gaussian pdf describing the noise.

If the clean speech is modeled as a mixture of K Gaussian pdfs, the Vector Taylor Series approach provides an estimate of the clean speech $\hat{\mathbf{x}}$ given the observed noisy speech $\mathbf{y}$ and the statistics of the noise (μ_n and Σ_n) as

$$\hat{\mathbf{x}} \approx \mathbf{y} - \sum_{k=1}^{K} P(k \mid \mathbf{y})\mathbf{g}(\mu_{x,k}, \mu_n) \tag{18}$$

where $\mu_{x,k}$ is the mean of the k^{th} clean Gaussian pdf and $P(k \mid \mathbf{y})$ is the probability of the noisy Gaussian k generating the noisy observation $\mathbf{y}$, given by

$$P(k \mid \mathbf{y}) = \frac{P(k)\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})}{\sum_{k'=1}^{K} P(k')\mathcal{N}(\mathbf{y}, \mu_{y,k'}, \Sigma_{y,k'})} \tag{19}$$

where $P(k)$ is the a-priori probability of the k^{th} Gaussian and $\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})$ is the k^{th} noisy Gaussian pdf (with mean $\mu_{y,k}$ and covariance matrix $\Sigma_{y,k}$) evaluated at $\mathbf{y}$. The mean and covariance matrix of the k^{th} noisy Gaussian pdf can be estimated from the noise statistics (μ_n and Σ_n) and the k^{th} clean Gaussian pdf ($\mu_{x,k}$ and $\Sigma_{x,k}$) using (16) and (17). In the experiments described here we have considered a first-order Taylor series. Only additive noise has been compensated for and no compensation of channel distortion has been considered.

B. Combination of Histogram Equalization With VTS

Taking into account the fact that HEQ and VTS compensation methods are based on different formulations one could expect that different noise effects could be compensated by the application of each. For this reason, we have also carried out exper-

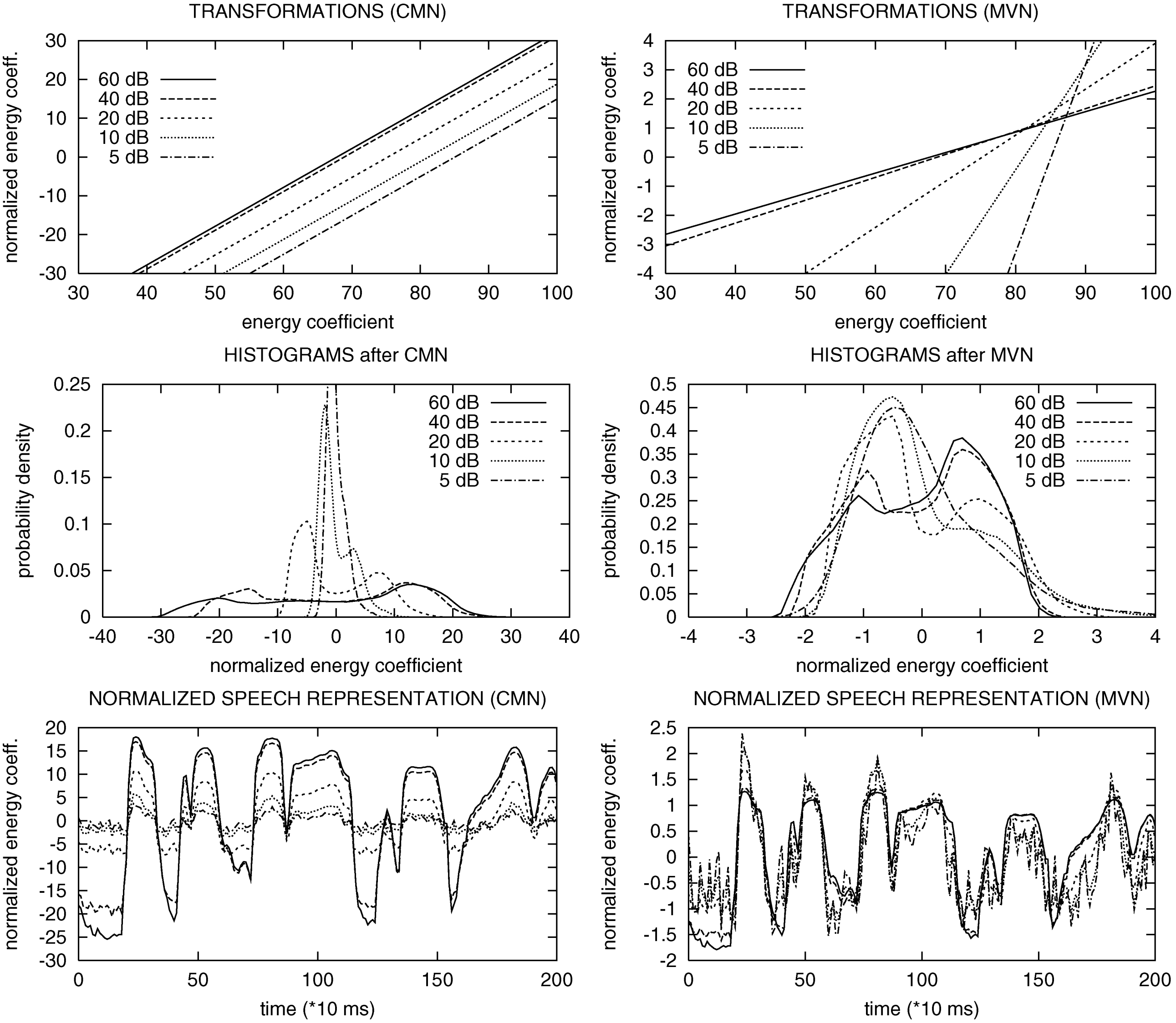

Fig. 2. Effect of CMN (left side) and MVN (right side) over the representation of speech for the energy coefficient.

iments where both compensation methods are combined. The VTS approach provides an appropriate compensation of additive noise provided that (A) the combination of the speech and the noise is additive in the FBE domain and (B) the standard deviation of the noise process is small for the different components in the log-FBE domain. However, the noise compensation yielded by VTS is limited for several reasons:

- noise transforms each Gaussian pdf $\mathcal{N}(\mathbf{x}, \mu_{x,k}, \Sigma_{x,k})$ into a non-Gaussian pdf;
- even for a high order in the VTS expansion, (18) is no more than an approximation;
- a limited order in the VTS expansion assumes that standard deviations in the clean speech Gaussian pdfs and the noisy Gaussian pdf are small.

These facts mean that the VTS-based estimation of the clean speech is affected by residual noise [24]. Modeling the residual noise is difficult in the log-FBE domain or in the cepstral domain, because it is not stationary (as it depends on the local signal to noise ratio) and because of its spectral distribution.

HEQ can be used in order to reduce this residual noise. After VTS is applied in the log-FBE domain and the feature vectors (including the logarithmic Energy, the cepstrum and the regression coefficients) are obtained, HEQ can be applied to each component of the feature vector. The experiments in which both compensation methods have been applied are labeled VTS+HEQ. As in the case of other compensation methods, VTS+HEQ is applied sentence-by-sentence during both training and recognition.

IV. Experiments and Results

The proposed HEQ method was evaluated with two different groups of continuous speech recognition experiments. In the first group, stationary noise processes were considered and recognition experiments were carried out with two Spanish speech databases. The second group of experiments was performed within the AURORA II framework. In this case, the speech was contaminated with several kinds of noise recorded in real conditions.

A. Experiments With Stationary Noise

Two different recognition tasks, based on two Spanish speech databases were used to test HEQ with stationary noise. The task labeled MGEO consisted of recognizing of sentences in a geographical context, with a vocabulary of 203 words. For this task we used the MINIGEO Spanish database [25], [26], which contains 600 sentences and 5655 words. The task labeled CONN-DIG consisted of recognizing connected digits (10 words in the vocabulary) and for these recognition experiments a Spanish connected digits database was used. This database contains 600 sentences and 4800 words. For the MGEO task, an appropriate bigrammar (estimated from the set of sentences allowed in this recognition task) was used as the language model [26]. The perplexity estimated for MGEO task using this bigrammar is 5.9. A connected-digit language model was used for the CONN-DIG task (ten word vocabulary; unrestricted length of sentences; after each digit, all digits are equally probable).

A Semi-Continuous Hidden Markov Model (SCHMM) recognition system [27], [28] was trained for both tasks. This recognizer used 256 Gaussian pdfs common to all the states in all the models. Each of the 24 main Spanish phonemes was modeled as a 3-state left-to-right HMM. A special HMM with 1 state was included in order to model silence. The acoustic models were trained with the EUROM1 database [29] (containing 803 sentences and 8648 words). For both training and recognition, versions of the cited databases decimated to 8 kHz were used. The speech representation includes pre-emphasis and segmentation of the signal into frames, with a frame length of 30 ms and a frame period of 10 ms. Each frame is represented as a feature vector including a logarithmic energy coefficient, 14 MFCC coefficients and the associated delta and acceleration coefficients, thus totaling 45 components.

In these experiments the speech signal was contaminated with 2 different kinds of additive noise. The first was recorded near a computer and included stationary sound from several sources, the most important of which came from the spinning of the hard disk. The second was Additive Gaussian White Noise (AGWN). The spectrum of both noise sources is represented in Fig. 3. These noise signals were used to artificially contaminate the speech signals. In order to reduce the mismatch between the training and the recognition conditions, the recognition system was trained with speech contaminated with AGWN at a SNR of 30 dB. The recognition experiments were performed for both recognition tasks and by contaminating the speech signals with both kinds of noise with SNRs ranging between 30 and 0 dB.

For each recognition condition we carried out 5 different experiments: CMN, MVN, HEQ, VTS, and VTS+HEQ. CMN and MVN were extended to all the components of the feature vector and used as a reference to compare the HEQ compensation method. The HEQ results correspond to the application of the histogram equalization method to each component of the feature vector. CMN, MVN and HEQ were applied to each sentence independently, i.e., the estimation of the mean, variance or the histograms are based on the observations corresponding to the sentence to be compensated. VTS was also implemented as a sentence-by-sentence compensation method. The noise affecting each band of the filter bank was estimated from the frames labeled as silence by a speech activity detector. The estimation of clean speech was based on a first-order VTS development (including terms associated with the speech and the noise) and made use of a 128 Gaussian mixture in the log-FBE domain estimated using the training data-base. Finally, VTS+HEQ was a combination of the last two methods. In this case, compensation of additive noise was first performed by applying VTS in the log-FBE domain. HEQ was then applied to the resulting speech representation in the MFCC domain. In all the cases, the compensation method was applied during both training and recognition.

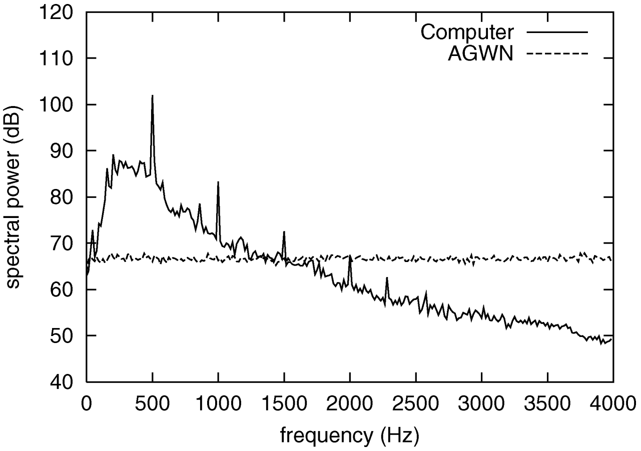

Fig. 3. Power spectrum of the noise signals used to contaminate the speech signals.

As a reference for the upper limit of performance that can be achieved by the compensation methods, we also included the results labeled Retrain. In this case the recognition system was trained with speech contaminated with noise under the same conditions as those in which recognition was performed. This minimizes the mismatch between the training and the recognition conditions.

The plots in Fig. 4 show the recognition performance (Word Error Rate) as a function of the SNR for both tasks and both kinds of noise when each of the different compensation methods was used. These plots show how, for a given SNR, the effect of the computer noise is less important than the effect of the AGWN, because the latter produces a degradation of all the spectral components while the former degrades a few spectral components more severely but leaves the others only slightly affected. The information loss caused by noise can be seen in the degradation of the performance in the case of the retrained recognizer. In this case, the mismatch is minimized and the performance is degraded mainly because of the random behavior of the noise and the loss of information it causes.

The results given by MVN are better than those of CMN. This is due to the equalization of more moments in the pdf of the noisy speech, which provides a better compensation of the mismatch caused by noise. Similarly, a significant improvement was given by HEQ compared with CMN and MVN. The linear transformations provided by CMN and MVN were not enough to compensate for the nonlinear effects of additive noise, while HEQ was able to deal with them.

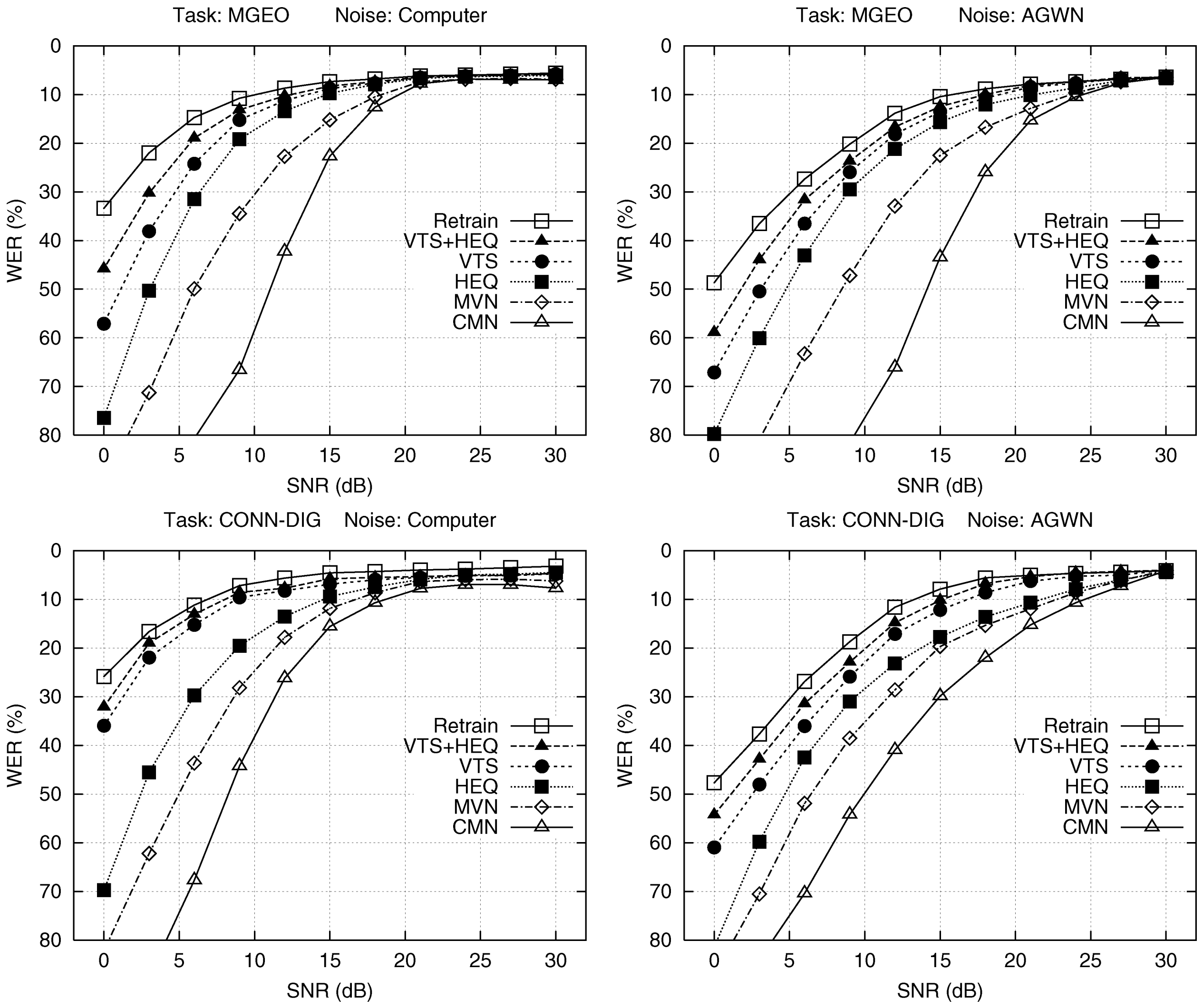

Fig. 4. Recognition results under stationary additive noise. Word error rate as a function of the SNR obtained by applying the different noise compensation methods.

The VTS compensation method provides more accurate compensation of the noise than HEQ since 1) it is based on an estimate of the noise statistics; 2) it makes use of information about the distribution of the vectors representing clean speech; and 3) it also makes use of an analytical description of the process by which the speech signal and the noise are combined in the representation. When the VTS and HEQ methods are combined, the results are better than those achieved using each technique in isolation. This fact shows that each compensation method is able to exploit some independent information. When both compensation methods are combined, considerable improvements are achieved under noisy conditions and the recognition results are close to those obtained when the recognizer is retrained.

B. Experiments With the AURORA II Database

The different noise compensation methods were evaluated within the AURORA II experimental framework [30]. According to the recommendations suggested during the AURORA session at the ICSLP-2002 conference, a re-endpointed version of the database was used. The database was accurately endpointed leaving 200 ms of silence at the beginning and at the end of each sentence. The AURORA II database is a subset of the TI-DIGITS, and contains connected digits spoken in English and recorded in a clean environment. Utterances were contaminated by adding several noise types at different SNR levels. Three test sets were defined. Set A and Set B contained only additive noise, whereas set C included additive noise and a simulated channel mismatch.

For this task, continuous density left-to-right HMMs were used as acoustic models. Each digit was modeled with 16 emitting states and a three-Gaussian mixture per state. Two additional models are defined for the silence. The first one models the silence at the beginning and at the end of each utterance. It consists of three states with a six Gaussian mixture per state. The other one models the inter-digit pauses and has only one state tied to the central state of the silence model. The recognizer is based on HTK and it uses a 39 component feature vector including 12 MFCCs, the logarithm of the energy and the corresponding delta and acceleration coefficients. Feature vectors are extracted at a frame rate of 100 Hz (more details can be found in [30]).

TABLE I
RECOGNITION RESULTS (WER %) OF MGEO AND CONN-DIGIT TASKS, AVERAGED ACROSS THE SNR BETWEEN 3 AND 21 dB, AND FOR AURORA II TASK, AVERAGED FOR SNR BETWEEN 0 AND 20 dB

Task	MGEO			Conn-Dig.			AURORA II			
Noise	Computer	AGWN	Average	Computer	AGWN	Average	Set A	Set B	Set C	Average
Baseline	-	-	-	-	-	-	39.51	44.48	32.56	**40.11**
CMN	45.98	59.14	**52.56**	37.11	45.26	**41.18**	31.58	27.12	33.14	**30.11**
MVN	30.21	39.48	**34.84**	25.52	33.79	**29.66**	21.38	20.54	24.86	**21.74**
HEQ	19.83	27.39	**23.61**	18.73	28.35	**23.54**	19.19	17.87	19.30	**18.68**
VTS	15.97	23.41	**19.69**	10.48	22.02	**16.25**	18.41	18.10	21.08	**18.82**
VTS+HEQ	13.55	20.93	**17.24**	9.26	19.18	**14.22**	13.50	13.49	15.31	**13.86**
Multicondition	-	-	-	-	-	-	11.93	12.45	15.80	**12.91**
Retrain	10.94	17.87	**14.41**	7.63	16.19	**11.91**	-	-	-	-

For each noise type and each SNR, we have carried out 6 different recognition experiments: Baseline, CMN, MVN, HEQ, VTS and VTS+HEQ. In all these experiments, the recognizer is trained using clean speech. An additional experiment, labeled Multicondition, was also included. In this case the baseline representation was used and the recognizer was trained with sentences contaminated with different kinds and levels of noise, as defined in the AURORA II framework.

As in the previous experiments, the different compensation methods were used during both training and recognition. The estimates of the means and variances (for CMN and MVN) or the estimates of the histograms (for HEQ and VTS+HEQ) were obtained from the sentence to be recognized. VTS was also applied on a sentence-by-sentence basis. The statistics of the noise affecting each band of the filter-bank were estimated from the silence at the beginning of each utterance. The estimate of the clean speech was based on a first-order VTS approach (including the terms associated with the speech and noise) and used a 128 Gaussian mixture in the log-FBE domain obtained from the training database.

Fig. 5 shows the recognition performance (Word Error Rate) as a function of the SNR for the AURORA II experiments. These results are the average for sets A, B and C, i.e., average results for all the types of noise considered. In the figure, successive improvements may be observed as more moments are equalized in the pdfs. Thus, CMN improves on the Baseline results, MVN improves on CMN and HEQ improves on MVN. Again, the linear transformations provided by CMN or MVN seem to be insufficient to compensate for the nonlinear effects caused by the noise.

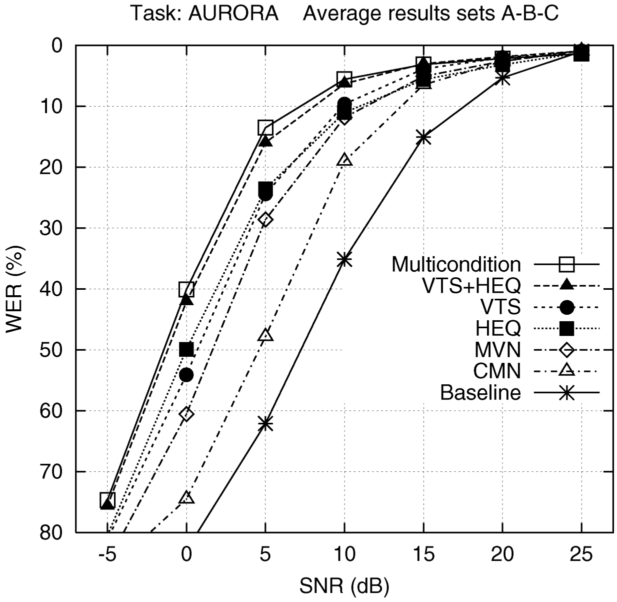

Fig. 5. AURORA II recognition results. Word error rate as a function of the SNR obtained by applying the different noise compensation methods.

If we compare the results provided by HEQ and VTS, the behavior is different from that observed in the MGEO and CONN-DIG tasks. With AURORA II, the plots are significantly closer, and HEQ outperforms VTS at SNR levels below 5 dB. This difference could be associated with the nonstationarity of the noise used for AURORA II experiments, which makes it harder to estimate the noise statistics and also increases the values of the covariance matrix of the noise. Both effects make VTS less effective than HEQ. When VTS and HEQ methods are combined, significant improvements are observed compared with using each method independently. This shows again that HEQ compensated for some distortions that VTS did not, and vice-versa. The results obtained for VTS+HEQ are very close to those obtained with the baseline system trained in Multicondition mode.

Table I shows the results of each recognition experiment, averaged across the different SNR levels. For tasks MGEO and CONN-DIG, the results are averaged between 3 and 21 dB. AURORA II results are averaged between 0 and 20 dB. In this table it is easy to compare the performance of the different compensation methods.

For the first two tasks, the quality of the results increases in the following order: CMN, MVN, HEQ, VTS, VTS+HEQ, and Retrain. All the improvements are statistically significant: the probability of each method being better than that preceding it was greater than 99.99% in all the experiments. In the AURORA II task, the order in which the quality of the results increases is: Baseline, CMN, MVN, VTS, HEQ, VTS+HEQ and Multi-condition. Even though HEQ outperforms VTS, the difference in the performance is slight and it is not statistically significant (the probability of HEQ being better than VTS is 84.4%). In the average results, all the other differences are significant (with a probability of improvement greater than 99.99%).

The small difference between VTS and HEQ in AURORA II experiments may be partly explained by the fact that Set C involved some channel distortion in addition to additive noise. The VTS implemented did not include compensation of the channel distortion while HEQ was able to compensate for it. This could explain why HEQ improves on the results of VTS for the Set C experiments. However, the results of HEQ and VTS are also very close for Set A and Set B. Since sets A and B only involve additive noise, the small difference between VTS and HEQ can be attributed to the nonstationarity of the noise added in the AURORA II task. In all the sets in the AURORA II, there was a significant improvement of VTS+HEQ with respect to VTS and HEQ. This improvement was greater in this case than in that of the MGEO and CONN-DIG tasks.

V. Summary and Conclusions

This paper describes an adaptation of the HEQ method to robust speech recognition. Based on an estimation of the histograms for the different components of the feature vectors in the sentence to be recognized, the method provides the transformations (one for each component) that convert the original histograms into a reference one. This method is able to compensate for the nonlinear distortions caused by noise. HEQ compensates for the effect of noise without relying on any prior assumptions about the nature of the components in the feature vector or the effect that the different noise processes affecting the speech signal produce on those components.

The compensation technique put forward here has been evaluated with continuous speech recognition experiments in which the signal has been contaminated at different SNRs with different types of noise. The HEQ method has yielded significant improvements in recognition performance under noisy conditions with respect to the baseline recognizer and with respect to linear methods such as CMN and MVN. HEQ can be considered as an extension of CMN and MVN to all the moments of the pdf. This way, HEQ provides appropriate transformations to compensate for the nonlinear effects caused by noise. The HEQ method has also been compared with the VTS compensation method, the formulation of which is based on an estimate of the noise statistics and an analytical description of the effect of the noise superimposed on the speech representation. In the case of stationary noise, VTS-based compensation improves the results given by HEQ due to its more accurate description of the contamination mechanism. In the AURORA II recognition experiments, HEQ and VTS provide very similar results. Both methods were also combined in order to obtain a more powerful technique. When VTS and HEQ methods were combined, significant improvements were achieved with respect to the independent application of each method. Thus, the compensation method based on HEQ is able to recover some information that VTS cannot and vice-versa. This is mainly due to the fact that each of the procedures is based on different assumptions. The experimental results show the utility of the method described in compensating for the effect of noise. One of the advantages of HEQ compared with other compensation methods is that it is not based on explicit models describing the contamination process and does not make any assumptions about the components in the feature vector or how these components are affected by noise. Therefore, the method can be applied to different kinds of speech parameterizations and be effective in the presence of different noise processes.

In the experiments described here, a sentence-by-sentence compensation for the noise was performed. Since the method we used relies on an accurate estimate of the original histograms, a sufficient number of observations is necessary for correct compensation. This fact needs to be taken into account for practical implementations, for example, in order to apply it in dialog systems. In the case of a very short sentence (including just a few frames) HEQ would not be able to compensate for the noise accurately. In this case, the use of several sentences when estimating the histograms should be considered (if the application or the dialog system allow it), in order to improve the efficiency of HEQ method. On the other hand, in the case of nonstationary noise, a segmental implementation of HEQ could be considered. This way, the compensation procedure could be adapted to a changing environment.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, 1995.

[2] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Splitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue, "The challenge of spoken language systems: Research directions for the nineties," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 1–21, Jan. 1995.

[3] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1996.

[4] A. de la Torre, D. Fohr, and J. P. Haton, "Compensation of noise effects for robust speech recognition in car environments," in *Proc. ICSLP 2000*, Oct. 2000.

[5] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garundadri, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivadas, "Robust ASR front-end using spectral based and discriminant features: Experiments on the Aurora tasks," in *Proc. EuroSpeech 2001*, Sep. 2001, pp. 429–432.

[6] P. J. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1996.

[7] A. de la Torre, D. Fohr, and J. P. Haton, "On the comparison of front-ends for robust speech recognition in car environments," in *Proc. ISCA ITR Workshop Adaptation Methods Speech Recognition*, Aug. 2001, pp. 109–112.

[8] R. M. Stern, B. Raj, and P. J. Moreno, “Compensation for environmental degradation in automatic speech recognition,” in *ESCA-NATO Tutorial Res. Workshop Robust Speech Recognition Unknown Communication Channels*, Apr. 1997, pp. 33–42.

[9] M. F. J. Gales, “‘Nice’ model-based compensation schemes for robust speech recognition,” in *ESCA-NATO Tutorial Res. Workshop Robust Speech Recognition Unknown Communication Channels*, Apr. 1997, pp. 55–64.

[10] S. Furui, “Recent advances in robust speech recognition,” in *ESCA-NATO Tutorial Res. Workshop Robust Speech Recognition Unknown Communication Channels*, Apr. 1997, pp. 11–20.

[11] C. H. Lee, “On stochastic feature and model compensation approaches to robust speech recognition,” *Speech Commun.*, vol. 25, pp. 29–47, 1998.

[12] C. R. Jankowski, Hoang-Doan Jr., and R. P. Lippmann, “A comparison of signal processing front ends for automatic word recognition,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 286–293, Jul. 1995.

[13] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Commun.*, vol. 25, pp. 133–147, 1998.

[14] S. V. Vaseghi and B. P. Milner, “Noise compensation methods for hidden Markov model speech recognition in adverse environments,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 11–21, Jan. 1997.

[15] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proc. ICASSP 96*, Atlanta, GA, 1996, pp. 733–736.

[16] N. S. Kim, D. Y. Kim, B. G. Kong, and S. R. Kim, “Application of VTS to environment compensation with noise statistics,” in *ESCA-NATO Tutorial Res. Workshop Robust Speech Recognition Unknown Communication Channels*, Apr. 1997, pp. 99–102.

[17] D. Y. Kim, C. K. Un, and N. S. Kim, “Speech recognition in noisy environments using first order vector Taylor series,” *Speech Commun.*, vol. 24, no. 1, pp. 39–49, 1998.

[18] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, “Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora-II database and tasks,” in *Proc. of EuroSpeech 2001*, Sep. 2001, pp. 221–224.

[19] P. J. Moreno, B. Raj, and R. M. Stern, “Data-driven environmental compensation for speech recognition: A unified approach,” *Speech Commun.*, vol. 24, no. 4, pp. 267–288, Jul. 1998.

[20] M. J. F. Gales, “Predictive model-based compensation schemes for robust speech recognition,” *Speech Commun.*, vol. 25, no. 1-3, pp. 49–74, 1998.

[21] R. C. Gonzalez and P. Wintz, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1987.

[22] J. C. Russ, *The Image Processing Handbook*. Boca Raton, FL: CRC, 1995.

[23] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuosly spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[24] J. C. Segura, M. C. Benítez, A. de la Torre, S. Dupont, and A. J. Rubio, “VTS residual noise compensation,” in *Proc. ICASSP*, Orlando, FL, 2000, pp. 409–412.

[25] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J. M. Pardo, and A. Rubio, “Development of Spanish corpora for speech research (Albayzin),” in *Proc. Workshop Int. Cooperation Standarization Speech Databases Speech I/O Assessment Methods*, Sep. 1991, pp. 26–28.

[26] J. E. Díaz-Verdejo, A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, and F. Casacuberta, “Albayzin: A task-oriented Spanish speech corpus,” in *Proc. 1st Int. Conf. Language Resources Evaluation (LREC 98)*, vol. 1, 1998, pp. 497–501.

[27] X. D. Huang and M. A. Jack, “Unified techniques for vector quantization and hidden Markov modeling using semi-continuous models,” in *Proc. ICASSP 89*, Glasgow, U.K., May 1989, pp. 639–642.

[28] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[29] J. Llisterri, L. Aguilar, B. Blecua, M Machuca, C. Mota, A. Ríos, and A. Moreno, “Spanish EUROM.1,” in *ESPRIT PROJECT 6819 (SAM-A)*, 1993.

[30] H. G. Hirsch and D. Pierce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions,” in *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millenium*, Paris, France, Sep. 2000.

Ángel de la Torre received the M.Sc and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1994 and 1999, respectively.

Since 1994, he has been working with the Research Group on Signals, Networking, and Communications, Department of Electronics and Computer Technology, University of Granada. In 2000, he joined the PAROLE Group, Laboratoire RFIA du LORIA, Nancy, France, as a Postdoctoral Researcher in the field of robust speech recognition, and the Institut für Angewandte Physik, Innsbruck, Austria, as a Postdoctoral Researcher in the field of cochlear implants. Since 2003, he has been an Associate Professor with the University of Granada. His research interests are in the field of signal processing, and particularly robust speech recognition, speech processing in noise conditions, and signal processing for cochlear implants. He is a reviewer for several scientific journals.

Antonio M. Peinado (M’95) was born in Granada, Spain, in 1963. He received the Licentiate, Grado, and Doctor degrees in physics from the University of Granada in 1987, 1989, and 1994, respectively. His Ph.D. dissertation was on HMM parameter estimation.

Since 1988, he has been working with the Research Group on Signals, Networking, and Communications, Department of Electronics and Computer Technology, University of Granada, on several topics related to speech recognition, coding, and transmission. In 1989, he was a Consultant in the Speech Research Department, AT&T Bell Labs, Murray Hill, NJ. Since 1996, he has been Associate Professor in the Department of Electronics and Computer Technology, University of Granada, and has also taught several national and international courses. His research interests are in distributed and robust speech recognition and speech and audio coding and transmission.

Dr. Peinado is a member of ISCA and AERFAI.

José C. Segura (M’93–SM’03) was born in Alicante, Spain, in 1961. He received the M.S. and Ph.D. degrees in physics from the University of Granada, Spain, in 1984 and 1991, respectively. His Ph.D. dissertation was on a variant of HMM modeling.

Since 1986, he has been working with the Research Group on Signals, Networking and Communications (GSTC), Department of Electronics and Computer Technology, University of Granada. Since January 2004, he has been the Coordinator of this research group. He has been the director of three Ph.D. dissertations on topics related to speech recognition. From 1987 to 1993, he was Assistant Professor, and since 1993 has been Associate Professor with the Department of Electronics and Computer Technology, University of Granada, and has also taught several national and international courses. His research interests are in robust speech recognition, distributed speech recognition, and signal processing.

Dr. Segura is a member of ISCA and AERFAI.

José L. Pérez-Córdoba (M’94) received the M.Sc. and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1987 and 2000, respectively.

Since 1990, he has been with the Department of Electrónica y Tecnología de Computadores, Faculty of Sciences, University of Granada, first as a Research and Assistant Professor and since 2004 as Associate Professor. From September 1993 to February 1994, he was a Visiting Researcher at System Research Center, University of Maryland. His research interests include speech processing (speech coding, distributed and robust speech recognition) and joint source-channel coding.

Ma Carmen Benítez (M'91) received the M.Sc. and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1988 and 1998, respectively.

Since 1990, she has been with the Department of Electrónica y Tecnología de Computadores, Faculty of Sciences, University of Granada, first as a Research and Assistant Professor and since 2003 as Associate Professor. From 2001 to 2002, she was a Visiting Researcher at the International Computer Science Institute, Berkeley, CA. Her research interests include speech processing, with a specific goal of speech recognition, confidence measures, and robust parameterization for speech recognition.

Dr. Benítez is a member of ISCA.

Antonio J. Rubio (SM'03) received the M.S. degree in physics from the University of Sevilla, Sevilla, Spain, in 1972 and the Ph.D degree from the University of Granada, Granada, Spain, in 1978.

He has been the Director of the Research Group on Signals, Networking, and Communications of the University of Granada since its creation. He is Full Professor with the Department of Electrónica y Tecnología de Computadores, University of Granada, in the area of signal theory and communications. His investigation is centered in the field of signal processing and, in particular, in the field of automatic speech recognition, in which he has directed several research projects. He has been the director of ten Ph.D. dissertations on topics related to speech recognition. He is a reviewer for several international journals.