Modeling, Estimating, and Compensating Low-Bit Rate Coding Distortion in Speech Recognition

Néstor Becerra Yoma, *Member, IEEE*, Carlos Molina, Jorge Silva, and Carlos Busso

***Abstract*—A solution to the problem of speech recognition with signals distorted by low-bit rate coders is presented in this paper. A model for the coding-decoding distortion, a HMM compensation method to include this model, and an EM-based adaptation algorithm to estimate this distortion are proposed here. Medium vocabulary continuous-speech speaker-independent recognition experiments with 8 kbps G.729(CS-CELP), 13 kbps RPE-LTP (GSM), 5.3 kbps G723.1, 4.8 kbps FS-1016 and 32 kbps G.726(ADPCM) coders show that the approach described in this paper is able to dramatically reduce the effect of the coding distortion and, in some cases, gives a word accuracy higher than the baseline system with uncoded speech. Finally, the EM estimation algorithm requires only one adapting utterance and the approach described is certainly suitable for dialogue systems where just a few adapting utterances are available.**

***Index Terms*—Coding distortion, EM estimation algorithm, HMM compensation, low-bit rate coders, speech recognition.**

I. Introduction

THE evolution and popularity of cellular and TCP/IP networks has created the problem of improving the recognition accuracy for speech distorted by low-bit rate coders. The distortion of coding schemes in speech recognizers is difficult to model and is an open problem that cannot be solved by applying conventional noise cancelling techniques [1] such as spectral subtraction [2], cepstral mean subtraction [3] and RASTA [4]. For instance, the effect of GSM coding in the cepstral domain leads to a spreading and displacing of the means of the Gaussians [5]. In [6] the application of probabilistic optimal filtering to various coding topologies scenarios was studied. The effect of coding and the relationship between bit rate and tandeming on small vocabulary isolated word and phone based speech recognition systems were reported in [7]. The effect of coding on speech recognition was also addressed in [8] where the performance of different cepstral features was analyzed. In [9], HMMs were trained with speech signals processed by several coders, and the most suitable acoustic model was selected using one utterance in the testing procedure. This method reduced the recognition error rate in cellular phone speech by 33%. The estimation of the instantaneous distortion introduced by GSM coding was able to lead to a reduction of 70% in the error rate introduced by the compression technique [5]. In [10], a front-end for speech recognition over IP networks was proposed to extract the feature vectors directly from the encoded speech instead of extracting the parameters after decoding. However, this approach requires access to the bit stream and does not avoid the distortion problem in the coding process.

This paper addresses the problem of the distortion produced by coding—decoding schemes employed in cellular systems and VoIP. For this purpose, the cepstral coefficients from uncoded and coded-decoded speech signals are linearly aligned to estimate a model for the distortion introduced by coding schemes. As a result, this distortion can be approximated with a Gaussian distribution whose mean and variance do not depend on the phonetic class. Then, a HMM compensation method is proposed by considering the original and unseen uncoded cepstral parameter as a random variable and by estimating the expected value of the output probability density function. Moreover, an Expectation-Maximization (EM) based algorithm to compute the coding distortion is proposed. This estimation algorithm requires no information about the coding scheme. The approach described in this paper has not been found in the literature, and dramatically eliminates the additional speech recognition errors introduced by the coder giving, in some scenarios, a word accuracy (WAC) even higher than the speaker-independent (SI) baseline system with uncoded speech. It is worth highlighting the fact that the method employs as few as one adapting utterance. Finally, the proposed scheme could also allow unsupervised retraining of the models using data which may even be slightly biased when compared to the original data.

II. Modeling Distortion Due to Coding

In order to model the distortion caused by coding algorithms, samples of clean speech were coded and decoded with the following coding schemes: 8 kbps CS-CELP [11] 13 kbps GSM [12], 5.3 kbps G723.1 [13], 4.8 kbps FS-1016 [14] and 32 kbps ADPCM [15]. After that, the original and coded-decoded speech signals, which were sampled at a rate of 8000 samples/s, were divided in 25 ms frames with 12.5 ms overlapping. Each frame was processed with a Hamming window, the band from 300 to 3400 Hz was covered with 14 Mel DFT filters, at the output of each channel the energy was computed and the log of the energy was estimated. The frame energy plus ten static cepstral coefficients, and their first and second time derivatives were estimated. Then, the parameterized original and coded-decoded utterances were linearly aligned to generate Figs. 1–6. The points (O_n^o, O_n^d), where O_n^o and O_n^d are the cepstral coefficient n estimated with the original and coded-decoded signals, respectively, are symmetrically distributed with

Manuscript received September 24, 2002; revised February 16, 2005. This work was supported by Fondecyt Grant 1030956 and Fondef Grant DO2I-1089, Chile. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ramesh A. Gopinath.

The authors are with the Electrical Engineering Department, University of Chile, Santiago, Chile (e-mail: nbecerra@cec.uchile.cl).

Digital Object Identifier 10.1109/TSA.2005.852994

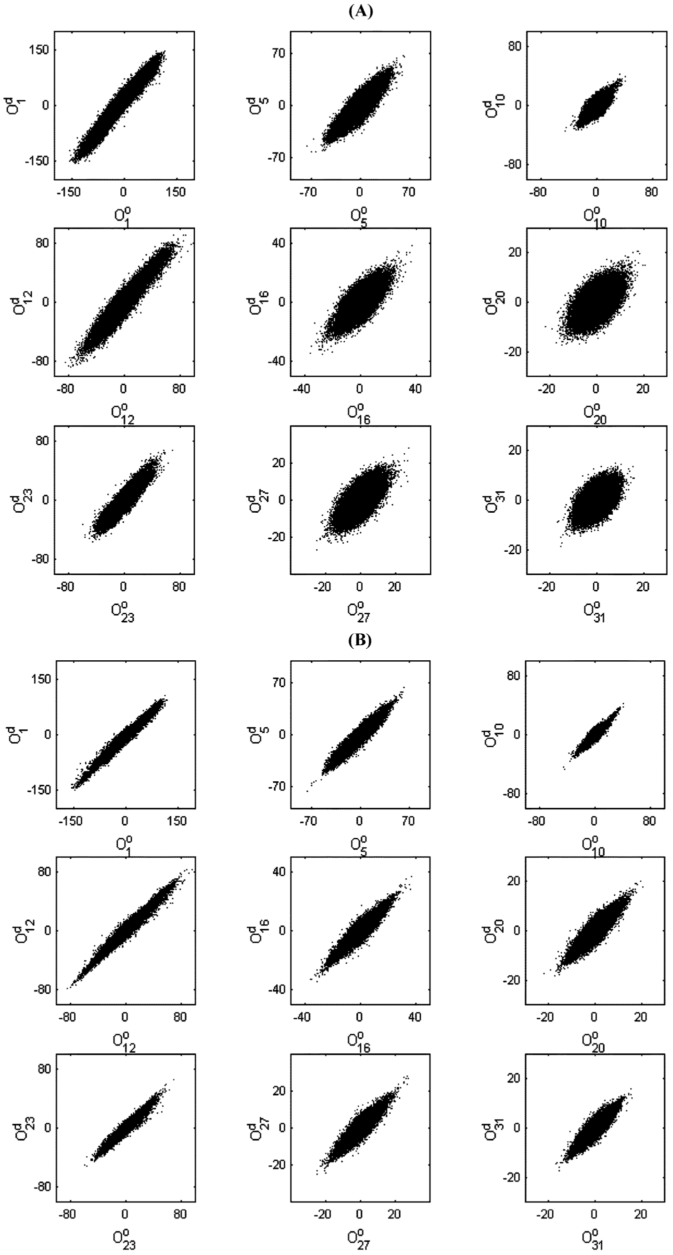

Fig. 1. Cepstral coefficients from uncoded (O^o) versus coded-decoded (O^d) speech signals. The coders correspond to (a) the 8 kbps CS-CELP from the ITU-T standard G.729 and (b) the 32 kbps ADPCM from the ITU-T standard G.726 The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The pairs (O^o, O^d) were generated by linearly aligning uncoded with coded-decoded speech. (a) 8 kbps CS-CELP and (b) 32 kbps ADPCM.

respect to the diagonal axis in the 8 kbps CS-CELP [Fig. 1(a)] and in the 32 kbps ADPCM [Fig. 1(b)]. This suggests that the coding-decoding distortion, defined as $D_n = O^o_n - O^d_n$, presents a reasonably constant dispersion around the mean that seems to be close to zero. As a consequence, the distribution of the coding-decoding distortion does not show a strong dependence on O^o_n in those cases. However, the same behavior is not observed in the 13-kbps GSM coder (Fig. 2) where the pairs (O^o_n, O^d_n) seems to be symmetrically distributed around a center near (0, 0).

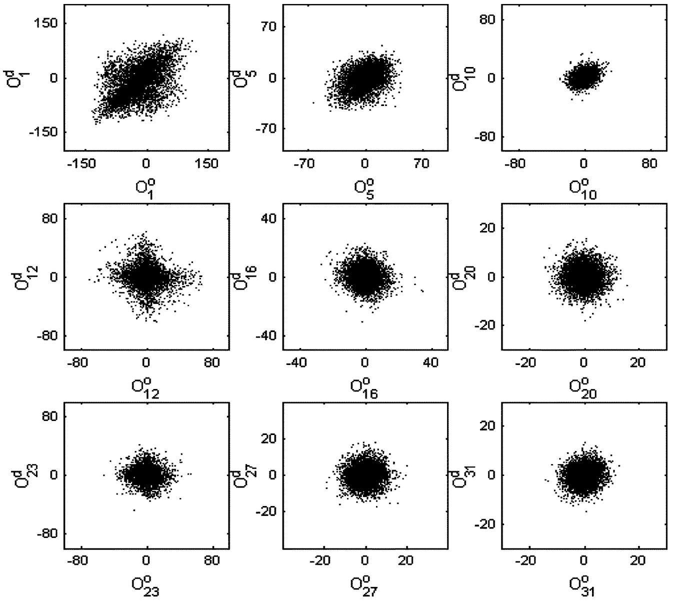

Fig. 2. Cepstral coefficients from uncoded (O^o) versus coded-decoded (O^d) speech signals. The coder is the 13 kbps GSM from the ETSI GSM-06.10 Full Rate Speech Transcoding. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The pairs (O^o, O^d) were generated by linearly aligning uncoded with coded-decoded speech.

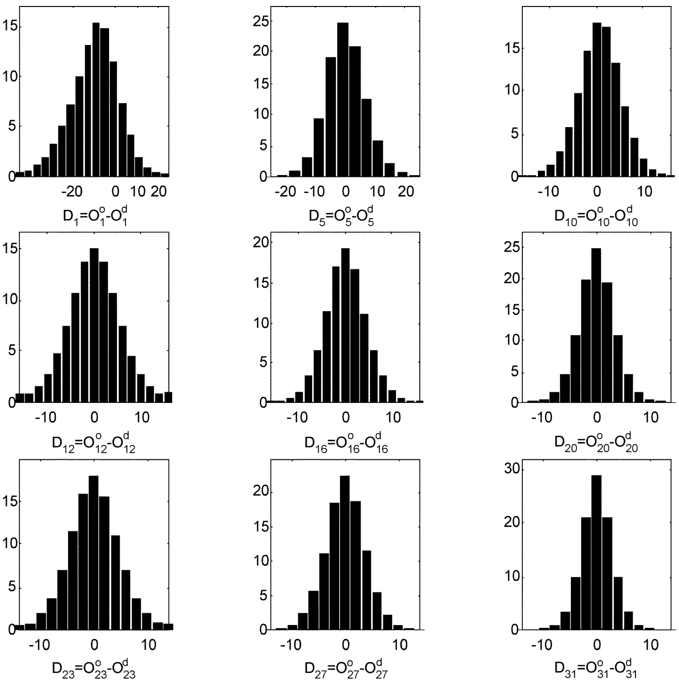

Fig. 3. Distribution of coding distortion $(O^o - O^d)$ with signals processed by 8 kbps CS-CELP from the ITU-T standard G.729. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated with the same data employed in Fig. 1.

The histograms presented in Fig. 3 (8 kbps CS-CELP) and Fig. 4 (5.3 kbps G723.1) strongly suggest that the coding-decoding distortion could be modeled as a Gaussian p.d.f., although the 5.3 kbps G723.1 coder provides (O^o_n, O^d_n) patterns similar to those observed with the 13 kbps GSM coder in Fig. 2.

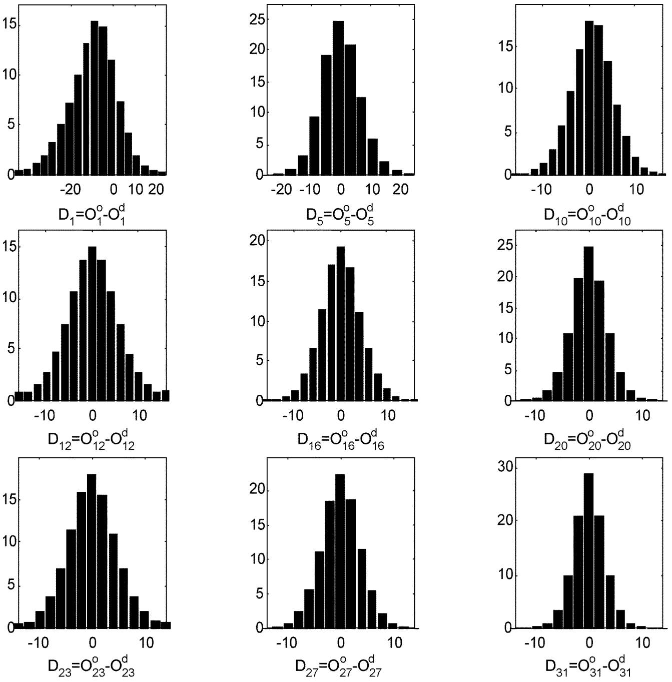

Fig. 4. Distribution of coding distortion $(O^o - O^d)$ with signals processed by 5.3 kbps G723-1 from the ITU-T standard G.723.1. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated with the same data employed in Fig. 4.

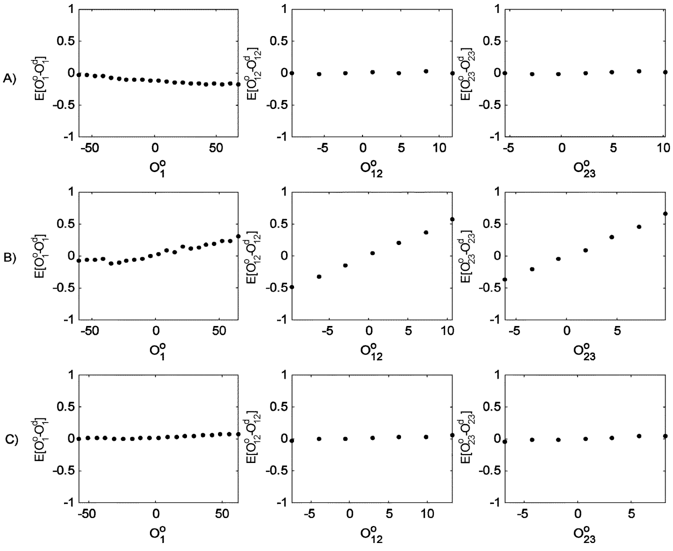

Fig. 5. Expected value of the coding-decoding error, $E[O_n^o - O_n^d] = m_n^d$, versus O^o. The expected value is normalized with respect to the range of observed O^o. The following coders are analyzed: (a) 8 kbps CS-CELP; (b) 13 kbps GSM; and (c) 32 kbps ADPCM. The cepstral coefficients correspond to a static (1), a delta (12), and a delta-delta (23).

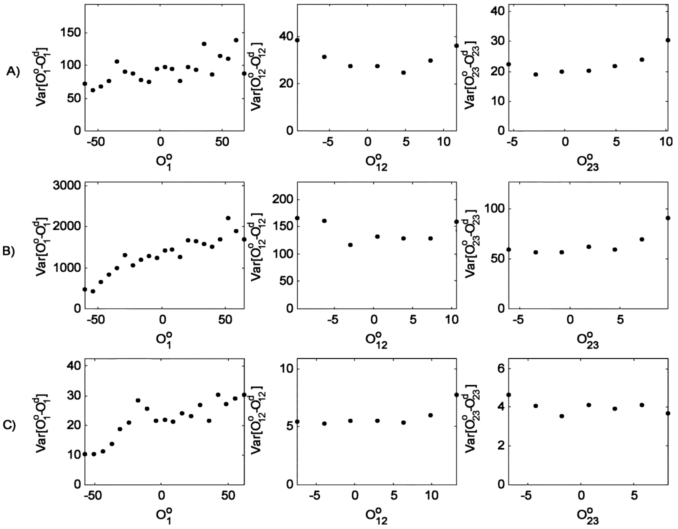

Fig. 6. Variance of the coding-decoding error, $Var[O_n^o - O_n^d] = v_n^d$, versus O^o. The following coders are analyzed: (a) 8 kbps CS-CELP; (b) 13 kbps GSM; and (c) 32 kbps ADPCM. The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23).

The expected value, normalized with respect to the range of the observed O_n^o, of the coding-decoding distortion versus O_n^o is shown in Fig. 5. Notice that the dependence of the expected value on O_n^o is weak for the 8 kbps CS-CELP and the 32 kbps ADPCM. Nevertheless, in the case of the 13 kbps GSM scheme this dependence is more significant, although the expected value is low compared to O_n^o itself and displays an odd symmetry. It is interesting to emphasize that the fuzzy circular-like (O_n^o, O_n^d) patterns observed with the 13 kbps GSM (Fig. 2) and the 5.3 kbps G723.1 coders are the result of this odd symmetry presented by the expected value of the distortion. The variance of the coding-decoding distortion versus O_n^o is shown in Fig. 6. According to Fig. 6, the assumption related to the independence of the variance with respect O_n^o does not seem to be unrealistic. Moreover, this assumption is strengthen by the fact that the distribution of O_n^o tends to be concentrated around $O_n^o = 0$.

From the previous analysis based on empirical observations and comparisons of the uncoded and coded-decoded speech signals, it is possible to suggest that the cepstral coefficient n in frame t of the original signal, $O_{t,n}^o$, could be given by

$$O_{t,n}^o = O_{t,n}^d + D_n \tag{1}$$

where $O_{t,n}^d$ is the cepstral coefficient corresponding to the coded-decoded speech signal; D_n is the distortion caused by the coding-decoding process with p.d.f. $f_{D_n}(D_n) = N(m_n^d, v_n^d)$ that does not depend on the value of the cepstral coefficient n, and therefore the phonetic class; $N(m_n^d, v_n^d)$ is a Gaussian distribution with mean m_n^d and variance v_n^d. The assumption related to the independence of D_n with respect to the value of a cepstral coefficient or the phonetic class is rather strong but seems to be a realistic model in several cases, despite the odd symmetry shown by the expected value of the coding-decoding distortion with some coders. Notice that this analysis takes place in the log-cepstral domain that is not linear. Moreover, as discussed later, this model is able to lead to dramatic improvements in WER with all the coding schemes considered in this paper.

In a real situation, $O_{t,n}^d$ is the observed cepstral parameter and $O_{t,n}^o$ is the hidden information of the original speech signal. From (1), the expected value of $O_{t,n}^o$ is given by

$$E\left[O_{t,n}^o\right] = O_{t,n}^d + m_n^d. \tag{2}$$

Concluding, according to the model discussed in this section, the distortion caused by the coding-decoding scheme is represented by the mean vector

$M^d = [m_1^d, m_2^d, m_3^d, \ldots m_n^d \ldots, m_N^d]$ and the variance vector $V^d = [v_1^d, v_2^d, v_3^d, \ldots v_n^d \ldots, v_N^d]$. Moreover, this distortion could be considered independent of the phonetic class and is consistent with the analysis presented in [5].

III. HMM Compensation

In this section a HMM compensation scheme is proposed. The training database is composed of clean speech signals that were not processed by any coding scheme. This compensation takes place in the Viterbi decoding procedure, and employs the mean m_n^d and variance v_n^d that model the distortion due to the coding-decoding process.

In the ordinary HMM topology the output probability of observing the frame O_t at state s, $b_s(O_t)$, is computed considering O_t as being a vector of constants. In the experiments reported here the observation vector is composed of static, delta and delta-delta cepstral coefficients, and according to Section II the distortion caused by the coding-decoding schemes can be modeled as (1). As a consequence, the cepstral parameters of the original signal, which are not available, should be considered as being random variables with normal distributions. Therefore, to counteract this incompatibility, this paper proposes to replace, in the decoding Viterbi algorithm, $b_s(O_t)$ with $E[b_s(O_t)]$ that denotes the expected value of the output probability, as suggested in [16] where the problem of speaker verification with additive noise was addressed.

In most HMM systems, the output probability is modeled with a mixture of Gaussians with diagonal covariance matrices [17]

$$b_s(O_t) = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} \frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}} \cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}^o - E_{s,g,n}\right)^2}{Var_{s,g,n}}} \tag{3}$$

where s, g, n are the indices for the states, the Gaussian components and the coefficients, respectively; p_g is a weighting parameter; $O_t = [O_{t,1}^o, O_{t,2}^o, O_{t,3}^o, \ldots, O_{t,N}^o]$ is the parameter vector composed of N coefficients that corresponds to the original uncoded signal; and,$E_{s,g,n}$ and $Var_{s,g,n}$ are the HMM mean and variance, respectively. Assuming that the coefficients $O_{t,n}^o$ are uncorrelated, which in turn results in the diagonal covariance matrices, the expected value of $b_s(O_t)$ is given by

$$E\left[b_s(O_t)\right] = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} E\left[\frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}} \cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}^o - E_{s,g,n}\right)^2}{Var_{s,g,n}}}\right]. \tag{4}$$

Considering (4), the expected value of $b_s(O_t)$ can be written as [16]

$$E\left[b_s(O_t)\right] = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} \frac{1}{\sqrt{2\cdot\pi\cdot Vtot_{s,g,n}^d}} \cdot e^{-\frac{1}{2}\cdot\frac{\left(E\left[O_{t,n}^o\right] - E_{s,g,n}\right)^2}{Vtot_{s,g,n}^d}} \tag{5}$$

where $E[O_{t,n}^o]$ is given by (2), and

$$Vtot_{s,g,n}^d = Var_{s,g,n} + v_n^d. \tag{6}$$

The coding-decoding corruption is modeled as an additive process in the cepstral domain that implies the introduction of additive correction terms in the mean, M^d, and variance, V^d, that could be modeled as being independent of the phonetic class and of the output probability densities $b_s(O_t)$. Additive correction in the mean and variance parameters has been applied in the context of additive/convolutional noise and Lombard effect [18], and speaker adaptation [19]. In those cases the compensation depends on the phonetic class and requires more correction parameters, which in turn increases the amount of adaptation samples. Moreover, considering the original signal as a random variable to take the expected value of the output p.d.f. is not always equivalent to estimate the bias mean and variance in HMMs as in [18] and [19].

IV. Estimation of Coding Distortion

In this section the coding-decoding distortion as modeled in Section II is evaluated employing the maximum likelihood criteria. Estimating the coding distortion in the HMM acoustic modeling is equivalent to find the vectors M^d and V^d defined in Section II. In this paper these parameters are estimated with the Expectation-Maximization (EM) algorithm using a code-book, where every code-word corresponds to a multivariate Gaussian, built with uncoded speech signals. The use of a code-book to represent the pdf of the features of the clean speech is due to the fact that M^d and V^d are considered independent of the phonetic class. Inside each code-word cw_j the mean $\mu_j^o = [\mu_{j,1}^o, \mu_{j,2}^o, \ldots, \mu_{j,N}^o]$ and variance $(\sigma_j^o)^2 = [(\sigma_{j,1}^o)^2, (\sigma_{j,2}^o)^2, \ldots, (\sigma_{j,N}^o)^2]$ are computed, and the distribution of frames in the cells is supposed to be Gaussian

$$f\left(O_t^o/\phi_j^o\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\left|\sum_j^o\right|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\cdot(O_t^o-\mu_j^o)^t\left(\sum_j^o\right)^{-1}(O_t^o-\mu_j^o)} \tag{7}$$

where N is the number of cepstral coefficients and also the dimension of the code-book; $\sum_j^o$ is the N-by-N covariance matrix that is supposed diagonal; and, $\phi_j^o = (\mu_j^o, \sum_j^o)$. In this case the speech model is composed of J code-words. Consequently, the p.d.f. associated to the frame O_t^o given the uncoded speech signal model is

$$f\left(O_t^o/\Phi^o\right) = \sum_{j=1}^{J} f\left(O_t^o|\phi_j^o\right) \cdot \Pr(cw_j) \tag{8}$$

where $\Phi^o = \{\phi_j^o \mid 1 \le j \le J\}$ denotes all the means and variances of the code-book. Equation (8) is equivalent to modeling the speech signal with a Gaussian mixture with J components.

If the coded-decoded distortion is independent of the code-word or class, it is possible to show that the coded-decoded speech signal is represented by the model whose parameters are denoted by $\Phi^d = \{\phi_j^d \mid 1 \le j \le J\}$, where $\phi_j^d = (\mu_j^d, \sum_j^d)$ and

$$\mu_j^d = \mu_j^o - M^d \tag{9}$$

$$\left(\sigma_{j,n}^d\right)^2 = \left(\sigma_{j,n}^o\right)^2 + v_n^d. \tag{10}$$

Consequently, the code-book that corresponds to the coded-decoded speech signal can be estimated from the original code-book by means of adding the vectors $-M^d$ and V^d, which model the compression distortion, to the mean and variance vectors, respectively, within each code-word.

In this paper M^d and V^d are estimated with the maximum likelihood (ML) criterion using adaptation utterances. Due to the fact that the maximization of the likelihood does not lead to analytical solutions, the EM algorithm [17], [20] was employed. Given an adaptation utterance O^d distorted by a coding-decoding scheme and composed of T frames

$$O^d = \left[O_1^d, O_2^d, O_3^d, \ldots, O_t^d, \ldots, O_T^d\right]$$

O^d is also called observable data. In the problem addressed here, the unobserved data is represented by

$$Y^d = \left[y_1^d, y_2^d, y_3^d, \ldots, y_t^d, \ldots, y_T^d\right]$$

where y_t^d is the hidden number that refers to the code-word or density of the observed frame O_t^d. The function $Q(\Phi, \hat{\Phi})$ is expressed as

$$Q(\Phi, \hat{\Phi}) = E\left[\log\left(f\left(O^d, Y^d/\hat{\Phi}\right)\right)|O^d, \Phi\right] \tag{11}$$

where $\hat{\Phi} = \{\hat{\phi}_j \; | 1 \leq j \leq J\}$, where $\hat{\phi}_j = (\mu_j^d, \sum_j^d)$ denotes the parameters that are estimated in an iteration by maximizing $Q(\Phi, \hat{\Phi})$. It can be shown that (11) can be decomposed in two terms

$$A = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr\left(cw_j|O_t^d, \hat{\Phi}\right) \cdot \log\left(\hat{\Pr}(cw_j)\right) \tag{12}$$

and

$$B = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr\left(cw_j|O_t^d, \Phi_j\right) \cdot \log\left(f\left(O_t^d|cw_j, \hat{\Phi}_j\right)\right). \tag{13}$$

The probabilities $\hat{\Pr}(cw_j)$ are estimated by means of maximizing A with the Lagrange method

$$\hat{\Pr}(cw_j) = \frac{1}{T}\sum_{t=1}^{T} \Pr\left(cw_j|O_t^d, \phi_j\right). \tag{14}$$

The distortion parameters defined in (1) could be estimated by applying to B the gradient operator with respect to M^d and V^d, and setting the partial derivatives equal to zero. However, this procedure does not lead to an analytical solution for V^d. In order to overcome this problem, the following algorithm is proposed.

1) Start with $\Phi = \Phi^o$, where $\Phi = \{\phi_j | 1 \leq j \leq J\}$ and $\phi_j = (\mu_j, \sum_j)$.

2) Compute $\Pr(cw_j|O_t^d, \phi_j)$

$$\Pr\left(cw_j|O_t^d, \phi_j\right) = \frac{f\left(O_t^d|\phi_j\right) \cdot \Pr(cw_j)}{\sum_{k=1}^{J} f\left(O_t^d|\phi_k\right) \cdot \Pr(cw_k)}. \tag{15}$$

3) Estimate $\hat{\Pr}(cw_j)$ with (14).

4) Estimate $\Delta\mu_n$ with

$$\Delta\mu_n = \frac{\sum\limits_{t=1}^{T}\sum\limits_{j=1}^{J}\left(\hat{Pr}\left(cw_j|O_t^d, \phi_j\right) \cdot \frac{\left(O_{t,n}^d - \mu_{j,n}\right)}{\sigma_{j,n}^2}\right)}{\sum\limits_{t=1}^{T}\sum\limits_{j=1}^{J}\left(\frac{\hat{Pr}\left(cw_j|O_t^d, \phi_j\right)}{\sigma_{j,n}^2}\right)}. \tag{16}$$

5) Estimate $\hat{\mu}_{j,n}$, $1 < j < J$ and $1 < n < N$

$$\hat{\mu}_{j,n} = \mu_{j,n} + \Delta\mu_n. \tag{17}$$

6) Estimate $\hat{\sigma}_{j,n}^2$ for each code-book

$$\hat{\sigma}_{j,n}^2 = \frac{\sum\limits_{t=1}^{T} \hat{Pr}\left(cw_j|O_t^d, \phi_j\right) \cdot \left(O_{t,n}^d - \hat{\mu}_{j,n}\right)^2}{\sum\limits_{t=1}^{T} \hat{Pr}\left(cw_j|O_t^d, \phi_j\right)}. \tag{18}$$

7) Estimate likelihood of the adaptation utterance O^d with the re-estimated parameters

$$f\left(O^d/\hat{\Phi}\right) = \sum_{\mathrm{t}=1}^{\mathrm{T}}\sum_{j=1}^{J} f\left(O_t^d|\widehat{\phi}_j\right) \cdot \widehat{P}r(cw_j). \tag{19}$$

8) Update parameters

$$\Phi = \hat{\Phi}$$
$$\Pr(cw_j) = \hat{\Pr}(cw_j)^{\cdot}$$

9) If convergence was reached, stop iteration; otherwise, go to step 2.

10)Estimate M^d and V^d

$$m_n^d = -\left(\mu_{j,n} - \mu_{j,n}^o\right) \tag{20}$$

for any $1 < j < J$ and

$$v_n^d = \frac{\sum\limits_{j=1}^{J}\left[\sigma_{j,n}^2 - \left(\sigma_{j,n}^o\right)^2\right] \cdot \Pr(\mathrm{cw_j})}{\sum\limits_{j=1}^{J} \Pr(\mathrm{cw_j})} \tag{21}$$

where $1 < n < N$. If $v_n^d < 0$, v_n^d is made equal to 0. It is worth observing that (16) was derived with $(\partial B)/(\partial(\Delta\mu_n)) = 0$, where B is defined in (13), $\hat{\mu}_{j,n} = \mu_{j,n} + \Delta\mu_n$ corresponds to the re-estimated code-word mean in an iteration. Expression (18) was derived by $(\partial B)/(\partial\hat{\sigma}_{j,n}^2) = 0$. Moreover, expressions (20) and (21) assume that the coding-distorting is independent of the code-word or class, and (21) attempts to weight the information provided by code-words according to the *a priori* probability $\Pr(cw_j)$.

The EM algorithm is a maximum likelihood estimation method based on a gradient ascent algorithm and considers the parameters M^d and V^d as being fixed but unknown. In contrast, maximum a posteriori (MAP) estimation [19] would assume the parameters M^d and V^d to be random vectors with a given prior distribution. MAP estimation usually requires less adaptation data, but the results presented here show that the proposed EM algorithm can lead to dramatic improvements with as few as one adapting utterance. Nevertheless, the proper use of an *a priori* distribution of M^d and V^d could lead to reductions in the computational load required by the coding-decoding distortion evaluation. When compared to MLLR [21], the proposed computation of the coding-decoding distortion requires fewer parameters to estimate, although it should still lead to improvements in word accuracy as a speaker adaptation method.

Finally, the method proposed here to estimate the coding-decoding distortion is similar to the techniques employed in [22]–[24] to compensate additive/convolutional noise and estimate the unobserved clean signal. In those papers the pdf for the features of clean speech is also modeled as a summation of multivariate Gaussian distributions, and the EM algorithm is applied to estimate the mismatch between training and testing conditions. However, this paper proposes a model of the low bit rate coding-decoding distortion that is different from the model of the additive and convolutional noise, although they are similar to some extent. The mean and variance compensation is code-word dependent in [22]–[24]. In contrast, M^d and V^d are considered independent of the code-word in this paper. This assumption is very important because it dramatically reduces the number of parameters to estimate and the amount of adaptation data required. Despite the fact that (16) to estimate M^d is the same expression employed to estimate convolutional distortion [22] if additive noise is not present [25], the methods in [22]–[24] do not compensate the HMMs. Notice that the effect of the transfer function that represents a linear channel is supposed to be an additive constant in the log-cepstral domain. On the other hand, additive noise corrupts the speech signal according to the local SNR [16], which leads to a variance compensation that clearly depends on the phonetic class and code-word.

V. Experiments

The estimation and compensation of the coding-decoding distortion proposed in this paper was tested with SI continuous speech recognition experiments using the LATINO database [26]. This database is composed of speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish. The training utterances were 4500 uncoded sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary is composed of almost 6000 words. The testing database was composed of 500 utterances provided by 4 testing speakers (two females and two males). Static, delta and delta-delta cepstral parameters were estimated as described in Section II. Each context-dependent phoneme was modeled with a 3-state left-to-right topology without skip transition, with eight multivariate Gaussian densities per state and diagonal covariance matrices. The HMMs were trained by means of the uncoded signal utterances using HTK and trigram language model was employed during recognition. Finally, it is worth mentioning that the speakers have different Latin American accents, which in turn makes the task more difficult.

The code-book to model the nondistorted speech process was composed of 256 code-words and was generated with the uncoded training utterances. The techniques that are proposed here are indicated as follows: *HMM-Comp*, with HMM compensation where M^d and V^d are estimated with the training utterances by directly aligning original and coded-decoded speech signals; and, *HMM-Comp-EM*, with HMM compensation where M^d and V^d are estimated according to the EM-based algorithm proposed in Section IV. Observe that *Baseline* indicates that no HMM compensation was applied. The word error rate (WER) was computed as $(S + D + I)/W \cdot 100$ where S, D, and I are the number of substitution, deletion and insertion errors, respectively, and W is the total number of words in the testing utterances. The results are shown in Tables I–V, and Fig. 7. The baseline system with nondistorted speech and without any compensation gave a WER equal to 5.9%.

VI. Discussions

According to Tables I, the ADPCM, GSM, CS-CELP, G723-1, and FS-1016 coders increased the error rate from 5.9% (baseline system) to 6.2%, 6.9%, 11.2%, 11.9%, and 15.2%, respectively. Also in Table I, it is possible to observe that the HMM compensation led to a reduction as high as 37% or 71% in the error rate introduced by the coding schemes when the average coding-decoding distortion was estimated by directly aligning the training uncoded and coded-decoded speech, *HMM-Comp*. This result clearly shows the validity of the method to model the coding distortion and to compensate the HMMs. However, it is worth mentioning that in *HMM-Comp* all the training speakers were employed to compute the average M^d and V^d. Notice that *HMM-Comp* gave a WER lower than the one achieved by the baseline system with uncoded speech (i.e. 5.9%) in some cases. This result could suggest that the HMMs are slightly under trained, so V^d could also tend to compensate this effect.

According to Fig. 7, the EM algorithm described here can lead to a reasonable approximation of M^d and V^d when compared to the average coding-decoding distortion computed with the training database. The difference between the EM estimation and the average M^d and V^d (Fig. 7) could be due to fact that the coding-decoding distortion depends on the speaker. As can be seen in Table I, the EM estimation of M^d and V^d with only one adaptation utterance dramatically reduced the effect of the ADPCM, GSM, CS-CELP, G723-1 and FS-1016 coding distortion, and gave a WER lower than *HMM-Comp* and than the one achieved by the baseline system with uncoded speech. A reasonable hypothesis could be the fact that the approach proposed here also provides an adaptation to testing condition beyond the type of codification because the estimation of the vectors M^d and V^d may also account for a speaker adaptation effect. Actually, Table II shows that the EM estimation algorithm applied to uncoded signal reduces in 56% the WER when compared to the baseline system. In fact, this result would be consistent with [27] where additive bias compensation in the cepstral domain for speaker adaptation was studied. Also according to Table II, it is possible to observe that the reduction in WER compared to the baseline system is as high as 52% or 78%, which in turn suggests that the approach proposed here is effective to model, estimate and compensate the coding-decoding distortion. It is worth emphasizing the fact that the reduction in WER increases when the bit-rate decreases. Finally, when compared to the baseline system, *HMM-Comp-EM* reduces the averaged difference between WER with distorted speech and clean signal from 4.4% to 0.4%.

The training database is composed of utterances from just 36 speakers. Consequently, the fact that that the proposed EM compensation method also introduces a speaker adaptation effect would be consistent with the size of the database. Most of

TABLE I
WER (%) WITH SIGNAL PROCESSED WITH THE FOLLOWING CODERS: 32 kbps ADPCM, 13 kbps GSM, 8 kbps CS-CELP, 5.3 kbps G723-1 AND 4.8 kbps FS-1016. THE BASELINE SYSTEM WITHOUT ANY COMPENSATION GIVES A WER EQUAL TO 5.9% WITH UNCODED UTTERANCES

Coder	*Bit rate*	*Baseline WER(%)*	*HMM-Comp. WER(%)*	*HMM-Comp-EM WER(%).*
ADPCM	32 kbps	6.2	3.9	2.8
GSM	13 kbps	6.9	3.8	3.3
CS-CELP	8 kbps	11.2	3.3	2.6
G723-1	5.3 kbps	11.9	5.8	2.6
FS-1016	4.8 kbps	15.2	7.4	3.6

TABLE II
WER (%) WITH UNCODED SIGNAL AND SIGNAL PROCESSED WITH THE FOLLOWING CODERS: 32 kbps ADPCM, 13 kbps GSM, 8 kbps CS-CELP, 5.3 kbps G723-1 AND 4.8 kbps FS-1016. THE CODING-DECODING DISTORTION IS ESTIMATED WITH THE EM ALGORITHM. THE REDUCTION IN WER IS ESTIMATED WITH RESPECT TO THE WER PROVIDED BY THE BASELINE SYSTEM WITHOUT ANY COMPENSATION

Coder	*HMM-Comp-EM WER(%)*	*Reduction(%) in WER*
Uncoded	2.6	56%
ADPCM	2.8	55%
GSM	3.3	52%
CS-CELP	2.6	77%
G723-1	2.6	78%
FS-1016	3.6	76%

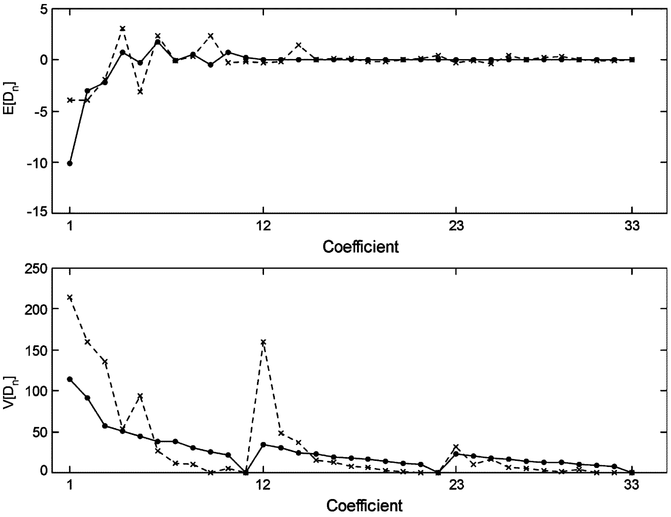

Fig. 7. M^d (top) and V^d (bottom) estimated with the EM based algorithm (— — x — —) and computed with the training database by directly aligning uncoded and coded-decoded speech samples (————). The signals were processed by the 8 kbps CS-CELP from the ITU-T standard G.729.

TABLE III
WER (%) WITH SIGNAL PROCESSED WITH THE FOLLOWING CODERS: 32 kbps ADPCM, 13 kbps GSM AND 8 kbps CS-CELP. THE EM ESTIMATION METHOD IS APPLIED IN THREE CASES: ADAPTATION OF MEANS AND VARIANCES AS ORIGINALLY PROPOSED; ADAPTATION OF MEANS ONLY; AND, ADAPTATION OF VARIANCES ONLY

Coder	*Baseline WER(%)*	*HMM-Comp-EM WER(%)*	*HMM-Comp-EM Mean adaptation only WER(%)*	*HMM-Comp-EM Variance adaptation only WER(%).*
ADPCM	6.2	2.8	5.8	2.9
GSM	6.9	3.3	6.9	3.5
CS-CELP	11.2	2.6	9.5	2.7
CLEAN	5.9	2.6	6.3	2.3

the compensation methods for HMMs attempt to adapt means or variances of the observation probability density functions. Moreover, it is to be expected that a canceling/compensation technique proposed to address a given distortion also helps to reduce the error introduced by another type of distortion. For instance, RASTA filtering was initially proposed to cancel convolutional noise but it also reduces the effect of additive noise. It is also hard to believe that a speaker adaptation scheme could not compensate or reduce convolutional noise. Finally, as shown in this paper, a speaker adaptation should also be useful for diminishing coding-decoding distortion, although this reduction would depend on the model adopted to estimate the means and variances. However, in additional speaker-dependent (SD) experiments with all the coders tested here, *HMM-Comp-EM* was able to lead to an average reduction in WER as high as 54% when compared to the baseline system. Those SD experiments were done by training the HMMs with both the training and testing databases. Consequently, the mismatch was restricted to the coding decoding distortion. This result strongly suggests that: first, the speaker adaptation effect in *HMM-Comp-EM*, if there is any, is not the most important mechanism in the reduction of WER provided by the proposed technique; and second, the improvement in word accuracy given by the method presented here is not due to under trained conditions.

To compare the improvement due to the compensation of means and variances, the EM estimation algorithm was modified to operate in two modes: compute M^d and make $V^d = 0$; and, compute V^d and make $M^d = 0$. According to Table III, the reduction in WER due to the estimation of V^d only is much higher than the one achieved with the computation of M^d only. As a consequence, this result suggests that the adaptation of variances plays a more important role in the compensation of coding-decoding distortion than the adaptation of means. Actually, when compared with the baseline system, the estimation of M^d only gives no improvement in WER with GSM coding. This result should be due to the odd symmetry presented by the expected value of the GSM distortion (Fig. 5), which in turn would make the assumption related to the independence of M^d with respect O_n^o less realistic than in ADPCM and CS-CELP. Also in Table III, the estimation of M^d only slightly degrades word accuracy with uncoded speech. This must be due to the fact that the model proposed here for M^d does not provide speaker adaptation and would require more adapting utterances if the mismatch between training and testing condition is low.

As shown in Table III, the estimation and compensation of V^d is very stable and leads to large reductions in WER. V^d tends to flatten the observation density function and to increase the number of hypotheses within the beam-width [28]. Consequently, the optimal search in the Viterbi decoding becomes more exhaustive. Actually, significant reductions in WER are

TABLE IV
WER (%) WITH SIGNAL PROCESSED WITH THE FOLLOWING CODERS: 32 kbps ADPCM, 13 kbps GSM AND 8 kbps CS-CELP. THE COMPENSATION IS MADE WITH BLIND RATZ. THE RESULTS ARE COMPARED WITH THE METHOD PROPOSED HERE. N_{AU} DENOTES THE NUMBER OF ADAPTING UTTERANCES

Coder	***Baseline WER(%)***	***HMM-Comp-EM WER(%)***	***RATZ*** $N_{AU}=1$	***RATZ*** $N_{AU}=4$	***RATZ*** $N_{AU}=10$	***RATZ*** $N_{AU}=100$	***RATZ*** $N_{AU}=500$
ADPCM	6.2	2.8	14.6	7.4	7.5	7.5	7.6
GSM	6.9	3.3	18.0	8.0	6.9	7.3	7.1
CS-CELP	11.2	2.6	20.4	10.4	9.7	8.6	10.3
CLEAN	5.9	2.6	13.5	7.0	6.5	7.1	7.6

observed when v_n^d is made equal to a constant, independently of n. The optimal WER achieved with $v_n^d = constant$ is 40% or 60% higher than *HMM-Comp-EM* with G.729, G.723 and FS-1016, and 20% lower than *HMM-Comp-EM* with GSM, ADPCM, and uncoded speech. However, $v_n^d = constant$ employs a computational load 60% or 70% higher than *HMM-Comp-EM* to obtain the optimal improvements in word accuracy.

The proposed EM adaptation method is unsupervised and requires only one adaptation utterance. Table IV presents results with blind RATZ without variance compensation. According to [23], blind RATZ jointly compensates for additive and convolutional noise by employing the EM algorithm and a summation of multivariate Gaussian distributions to model the p.d.f. for the features of clean speech. The mean and variance compensation is code-word dependent, so the coding distortion is now represented by the set of mean and variance vectors M_j^d and V_j^d, respectively, where j denotes the code-word cw_j and $1 \leq j \leq J$ as defined in Section IV. First, blind RATZ computes M_j^d and V_j^d. Then, it estimates the cepstral coefficient n in frame t of the original unobserved signal, $O_{t,n}^o$, by using M_j^d, $Pr(cw_j|O_t^d, \phi_{\rm j})$ and a modified minimum mean square error (MMSE) method [23]. The number of adapting utterances, N_{AU}, was made equal to 1, 4, 10, 100 and 500. The adopted procedure is described as follows to cover all the testing data: first, RATZ was applied on N_{AU} utterances; then, the compensated N_{AU} utterances were recognized. Notice that blind RATZ is an unsupervised method. As can be seen in Table IV, RATZ without variance compensation was not able to lead to significant improvements except with CS-CELP. Word accuracy given by RATZ strongly depends on the number of adapting utterances employed to compute $O_{t,n}^o$. When compared to the baseline system, RATZ could provide an improvement in WER if the number of adapting utterances is higher than 4 or 10. If the method employs only one adaptation utterance, it always gives a WER even higher than the one achieved with the baseline system. This must be due to the fact that RATZ estimates $2 \times J$ mean and variances vectors M_j^d and V_j^d. In contrast, the proposed method *HMM-Comp-EM* requires less adaptation data because it needs to estimate only two vectors: M^d and V^d. It is worth highlighting that *HMM-Comp-EM* provides higher recognition accuracy even when the whole testing data was employed by RATZ. This must result of the fact that the speaker adaptation effect is less important when the adaptation utterances come from more than one speaker. Notice that RATZ does not directly employ the variance vector V_j^d to compensate HMMs or estimate the original signal, so it could be considered a special case of the proposed EM adaptation method. Also, no improvement was observed when N_{AU} increases from 10 or 100 to 500. This should be a result of the fact that the coding-decoding distortion is speaker-dependent and all the testing speakers are employed to estimate M_j^d and V_j^d, which in turn also reduces any speaker adaptation effect.

TABLE V
WER (%) WITH SIGNAL PROCESSED WITH THE FOLLOWING CODERS: 32 kbps ADPCM, 13 kbps GSM AND 8 kbps CS-CELP. THE CODING-DECODING DISTORTION IS COMPUTED WITH SUPERVISED ML ESTIMATION BASED ON FORCED VITERBI ALIGNMENT. THE RESULTS ARE COMPARED WITH THE METHOD PROPOSED HERE. N_{AU} DENOTES THE NUMBER OF ADAPTING UTTERANCES

Coder	***Baseline WER(%)***	***HMM-Comp-EM WER(%)***	***Superv-ML*** $N_{AU}=1$	***Superv-ML*** $N_{AU}=4$	***Superv-ML*** $N_{AU}=10$	***Superv-ML*** $N_{AU}=100$	***Superv-ML*** $N_{AU}=500$
ADPCM	6.2	2.8	4.6	3.7	4.2	3.0	3.2
GSM	6.9	3.3	5.8	4.6	4.1	4.2	3.7
CS-CELP	11.2	2.6	6.0	4.5	4.5	3.8	3.8
CLEAN	5.9	2.6	5.5	3.4	3.9	3.1	3.2

Table V presents results with supervised ML estimation, *Superv-ML*, based on forced Viterbi alignment to estimate M^d and V^d. This supervised adaptation algorithm is similar to the one presented in [18] except for the fact that the Forward-Backward procedure was replaced with the Viterbi algorithm. Moreover, instead of employing one or more Gaussian bias per HMM, Superv-ML makes use of the proposed coding distortion model and estimates only one set of vectors M^d and V^d per adaptation utterances. The number of adapting utterances, N_{AU}, was also made equal to 1, 4, 10, 100, and 500. The following procedure was applied to cover all the testing data: first, M^d and V^d were evaluated with N_{AU} utterances; then, the next N_{AU} utterances were recognized employing the previously estimated coding distortion. According to Table V, the improvement in WER given by *Superv-ML* also depends on N_{AU}. The stochastic model employed by the proposed EM unsupervised algorithm is more robust than the one provided by the supervised ML method, which in turn is composed of only the HMMs corresponding to the adapting utterances. Consequently, the requirement with respect to the amount of adaptation data to achieve the highest reduction in WER is more severe in *Superv-ML*. Finally, as can be seen in Table V, when $N_{AU} = 500$ the supervised algorithm could give improvements in WER slightly better than *HMM-Comp* and worse than *HMM-Comp-EM* with GSM and ADPCM, despite the fact that the proposed EM unsupervised estimation algorithm employed only one adaptation utterance and *Superv-ML* made use of the whole testing database. This should be due to: *HMM-Comp* used the training database and *Superv-ML* employed the testing utterances to compute the coding-decoding distortion; second, the coding-decoding distortion should be speaker-dependent.

VII. CONCLUSION

This paper proposes a solution to the problem of speech recognition with signals distorted by low-bit rate coders. The solution includes: a model for the coding-decoding distortion; a HMM compensation method to include this model; and an EM-based adaptation algorithm to estimate this distortion that corresponds to a mean and a variance vector. Medium vocabulary continuous-speech SI recognition experiments with ADPCM, GSM, CS-CELP, G723-1 and FS-1016 coders show that the approach presented here is able to substantially reduce the effect of the coding distortion without no information about the coding scheme, and can give a WAC even higher than the baseline system with uncoded speech. The approach could also be analyzed in the context of optimally increasing the hypothesis density within the beam search, which in turn has not been addressed in the specialized literature. Moreover, the EM estimation algorithm needs only one adapting utterance and the approach described here is certainly suitable for dialogue systems where just a few adapting utterances are available. As a result, the method could be used where many short calls come from different coders. The problem of reducing the computational load of the coding distortion estimation is considered out of the scope of this paper and is proposed as a future work. Finally, the HMM compensation strategy employed here could also be applied as a framework to address the problem of joint additive/convolutional noise and coding distortion canceling.

Acknowledgment

The authors wish to thank Dr. S. King, CSTR/University of Edinburgh, Edinburgh, U.K., for having proofread this manuscript, and S. Ferraz de Campos Neto, International Telecommunication Union (ITU), Geneva, Switzerland, for the very productive discussions on speech coding technology and standards.

References

[1] Y. Gong, "Speech recognition in noise environments: a survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.

[2] M. Berouti *et al.*, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verication," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, 1981.

[4] H. Hermansky *et al.*, "Compensation for the effect of the communication chennel in auditory-like analysis of speech (RASTA-PLP)," in *Proc. Eurospeech 91*, 1991, pp. 1367–1370.

[5] J. M. Huerta, "Speech recognition in mobile environments," Ph.D dissertation, Dept. Elect. Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.

[6] T. Salonidis and V. Digalakis, "Robust speech recognition for multiple topological scenarios of the GSM mobile phone system," in *Proc. ICASSP 98*, 1998.

[7] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. ICSLP 96*, 1996.

[8] S. Euler and J. Zinke, "The influence of speech coding algorithms on automatic speech recognition," in *Proc. ICASSP 94*, vol. I, 1994, pp. 621–624.

[9] M. Naito, S. Kuroiwa, T. Kato, T. Shimizu, and N. Higuchi, "Rapid CODEC adaptation for cellular phone speech recognition," in *Proc. Eurospeech 2001*, Alborg, Denmark, 2001.

[10] C. Peláez, A. Gallardo, and F. Díaz-de-María, "Recognizing voice over IP: a robust front-end for speech recognition on the world wide web," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 209–218, 2001.

[11] ITU-T, Recommendation G.729-Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-CELP), Mar. 1996.

[12] ETSI, GSM-06.10 Full Rate Speech Transcoding. RPE-LTP (Regular Pulse Excitation, Long Term Predictor), ETSI, France, Oct. 1992.

[13] ITU-T, Recommendation G.723.1 Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbps, Marzo 1996.

[14] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Process.*, vol. 1, no. 3, pp. 145–155, 1991.

[15] ITU-T, Recommendation G.726, "40-,32-,24-, and 16-Kb/s adaptive differential pulse code modulation", Dec. 1990.

[16] N. B. Yoma and M. Villar, "Speaker Verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 3, pp. 158–166, Mar. 2002.

[17] X. D. Huang *et al.*, *HMM for Speech Recognition*. Edinburgh, U.K.: Edinburgh Univ. Press, 1990.

[18] M. Afify, Y. Gong, and J. Haton, "A general joint additive and convolutive bias compensation approach applied to noise Lombard speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 6, pp. 524–538, Nov. 1998.

[19] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[20] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[21] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[22] A. Acero and R. Stern, "Environmental robustness in automatic speech recognition," in *Proc. ICASSP*, 1990.

[23] P. J. Moreno, B. Raj, E. Govea, and R. M. Stern, "Multivariate Gaussian based cepstral normalization for robust speech recognition," in *Proc. ICASSP '95*, 1995.

[24] R. M. Raj, E. B. Gouvea, P. J. Moreno, and R. M. Stern, "Cepstral compensation by polynomial approximation for environment-independent speech recognition," in *Proc. ICSLP*, 1996.

[25] N. B. Yoma, "Speech recognition in noise using weighted matching algorithms," Ph.D. dissertation, Univ. Edinburgh, Edinburgh, U.K., 1998.

[26] Linguistic Data Consortium (LDC). (1995) Latino database. Univ. Pennsylvania, Philadelphia, PA. [Online] Available: http://www.ldc.upenn.edu/Catalog/LDC95S28.html

[27] Y. Zhao, "An acoustic-phonetic based speaker adaptation technique for improving speaker independent continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 380–394, Jul. 1994.

[28] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Process. Mag.*, pp. 64–81, Sep. 1999.

Néstor Becerra Yoma (S'96–A'98–M'04) was born in Santiago, Chile, in 1964. He received the B.Sc. and M.Sc. degrees from Campinas State University (UNICAMP), Sao Paulo, Brazil, in 1986 and 1993, respectively, and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 1998, respectively, all in electrical engineering.

In 1998 and 1999, he was a Postdoctoral Researcher at UNICAMP and a full-time Professor at Mackenzie University, Sao Paulo, Brazil. From 2000 to 2002, he was an Assistant Professor with the Department of Electrical Engineering, Universidad de Chile, Santiago, where he is currently lecturing on telecommunications and speech processing, and working on robust speech recognition/speaker verification, dialogue systems, and voice over IP. At the Universidad de Chile he has set up the Speech Processing and Transmission Laboratory (LPTV) to study speech technology applications on the Internet and telephone line. He has been an Associate Professor since 2003 and is the first author of 12 journal articles and 30 conference papers. His research interests include speech processing, real-time Internet protocols, QoS, and usability evaluation of interfaces.

Dr. Yoma is a member of the International Speech Communication Association.

Carlos Molina was born in Santiago, Chile, in 1980. He received the B.Sc. degree in engineering sciences from Universidad de Chile in 2003. Since 2004, he has been a postgraduate student in the Speech Processing and Transmission Laboratory at the Department of Electrical Engineering, Universidad de Chile, where he has worked on noise canceling and speaker adaptation techniques for speech recognition.

He has been a co-author in three journal papers and three conference articles in the last two years. He has also worked on the implementation of speech recognition based dialogue systems.

Mr. Molina is a student member of the International Speech Communication Association.

Jorge Silva was born in Santiago, Chile, in 1977. He received the B.Sc. and engineering degree in electrical engineering with highest distinction from the Universidad de Chile, Santiago, in 2002. He is pursuing the Ph.D. degree in the Department of Electrical Engineering at the University of Southern California, Los Angeles.

He holds a faculty position in the Department of Electrical Engineering, Universidad de Chile. From 2000 to 2003, he was a Research Assistant in the Speech Processing and Transmission Laboratory (LPTV), Department of Electrical Engineering, Universidad de Chile. His general research interests include speech recognition, statistical signal processing, multiresolution analysis, and information theory applied to signal processing.

Carlos Busso was born in Santiago, Chile. He received his M.Sc. degree from the Department of Electrical Engineering, Universidad de Chile, in 2003, and he is currently pursuing the Ph.D. degree at the University of Southern California, Los Angeles.

From 2000 to 2003, he was a Research Assistant in the Speech Processing and Transmission Laboratory (LPTV) at the Department of Electrical Engineering where he worked on speech coding, voice over IP, QoS, low bit rate coding distortion in speech recognition and real-time protocols for the Internet. His current research interests include digital signal processing, speech and video signal processing, and multimodal interfaces.