Feature-dependent compensation of coders in speech recognition

Néstor Becerra Yoma*, Carlos Molina

Electrical Engineering Department, Universidad de Chile, Av.Tupper 2007, P.O. Box 412-3, Santiago, Chile

Abstract

A solution to the problem of speech recognition with signals corrupted by coders is presented. The coding-decoding distortion is modelled as feature dependent. This model is employed to propose an unsupervised expectation-maximization (EM) estimation algorithm of the coding–decoding distortion that is able to cancel the effect of coders with as few as one adapting utterance. No knowledge about the coder is required. The feature-dependent adaptation can give a word error rate (WER) 21% lower than the feature-independent model. Finally, when compared to the baseline system, the reduction in WER can be as high as 70%.

Keywords: Speech recognition; Coding distortion; HMM compensation

1. Introduction

The mismatch between training and testing condition is certainly the most important problem to be solved to make speech recognition successful in real applications. This mismatch can be due to additive and/or convolutional noise, to speaker and to the distortion resulted from low-bit-rate coders. The adaptation to new testing corrections is usually achieved by estimating corrections to the HMM (hidden Markov models) means and variances. These corrections may or may not depend on the HMM state. However, all the adaptation methods found in the specialized literature do not explore the relationship between the distortion that needs to be compensated and the value of the feature parameters. This is certainly due to the fact that the observed feature parameters are those already distorted, while the original clean signal is unobserved and hidden information.

The rapid growth of mobile networks and Internet around the world has created the problem of improving the recognition accuracy for speech

*Corresponding author. Tel: +56 2 678 4205; fax: +56 2 695 3881.

E-mail address: nbecerra@cec.uchile.cl (N. Becerra Yoma).

distorted by low-bit-rate coders. GSM (global system for mobile communications) and G.729 standards are certainly the most popular and this paper focuses on estimating and compensating the distortion of these coders in speech recognition. This distortion cannot be solved by applying cancelling/compensation techniques that employ a model for additive and/or convolutional noise [1], such as, spectral subtraction [2], Rasta [3], PMC (parallel model combination) [4], and (CMN) cepstral mean normalization. In [1] weighted acoustic modelling based on average distortion statistics provided a reduction of up to 70% in word error rate (WER) introduced by GSM if the coder is known and fixed. MLLR (maximum likelihood linear regression) has successfully been applied to compensate for additive noise and speaker adaptation [5]. However, there is not any evidence in the specialized literature suggesting that MLLR can lead to substantial improvements with low-bit-rate coders. MLLR builds a transform for the HMM parameters (the Gaussian means and variances) using linear regression so that the adapted parameters better represent the test scenario. In blind RATZ [6] the p.d.f. for the features of clean speech is modelled as a summation of multivariate Gaussian distributions, and the EM algorithm is applied to estimate the mismatch between training and testing conditions. First, blind RATZ adapts the means and variances within every Gaussian. Then, it estimates the cepstral coefficients of the original unobserved signal by using a modified minimum mean square error. In [6] RATZ was applied to WSJ database corrupted with additive noise leading to significant reductions in WER by employing as many as 10, 40 or even 100 adapting utterances. In [7] a unified approach to the acoustic mismatch problem was proposed. A maximum likelihood state-based additive bias compensation algorithm was employed in the context of the continuous density HMM. This technique estimates the correction parameters to compensate the means and variances by applying the EM algorithm in combination with a parallel model combination transform. This approach was applied on a supervised basis and led to reductions between 50% and 70% in WER using 42 or even 84 adapting utterances (isolated words) to compensate for additive/convolutional noise and Lombard effect. In contrast, the feature-dependent technique proposed in the current paper adapts the HMM on an unsupervised basis by considering the original uncoded signal as a random variable, and by assuming the coding–decoding distortion as independent of the state and model, but dependent on the value of the original uncoded cepstral coefficient. This assumption dramatically reduces the number of parameters to estimate. Compared to the baseline system, the proposed approach can lead to reductions in WER as high as 70%.

Empirical observations suggested that the coding–decoding distortion in cepstral coefficient n in frame t could be modelled as

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

where $O^{o}_{t,n}$ and $O^{d}_{t,n}$ are the cepstral coefficients corresponding to the original and coded–decoded speech signal, respectively; D_n is the distortion caused by the coding–decoding process with p.d.f. $f_{D_n}(D_n) = N(m^{d}_n, v^{d}_n)$, and is modelled as a Gaussian distribution with mean $m^{d}_n = E[D_n] = E[O^{o}_{t,n} - O^{d}_{t,n}]$ and variance $v^{d}_n = \mathrm{Var}[D_n]$. The HMM compensation is achieved by replacing the output probability by its expected value in the Viterbi algorithm [8], which in turn leads to replacing the observed $O^{d}_{t,n}$ and variance with

$$E\left[O^{o}_{t,n}\right] = O^{d}_{t,n} + E[D_n], \tag{2}$$

$$\mathrm{Var}^{d}_{h,s,g,n} = \mathrm{Var}_{h,s,g,n} + v^{d}_n, \tag{3}$$

where $E[O^{o}_{t,n}]$ is the expected value of the unseen cepstral coefficient, considered as a random variable, in the original speech signal according to (1); $\mathrm{Var}^{d}_{h,s,g,n}$ and $\mathrm{Var}_{h,s,g,n}$ are, respectively, the compensated and original variances in HMM h, state s, Gaussian component g and coefficient n. According to Fig. 1 $m^{d}_n = E[D_n]$ could be considered independent of $O^{o}_{t,n}$ when the speech signal is processed with G.729 CS-CELP (conjugate structure-code excited linear prediction). However, this model is not always accurate and this paper

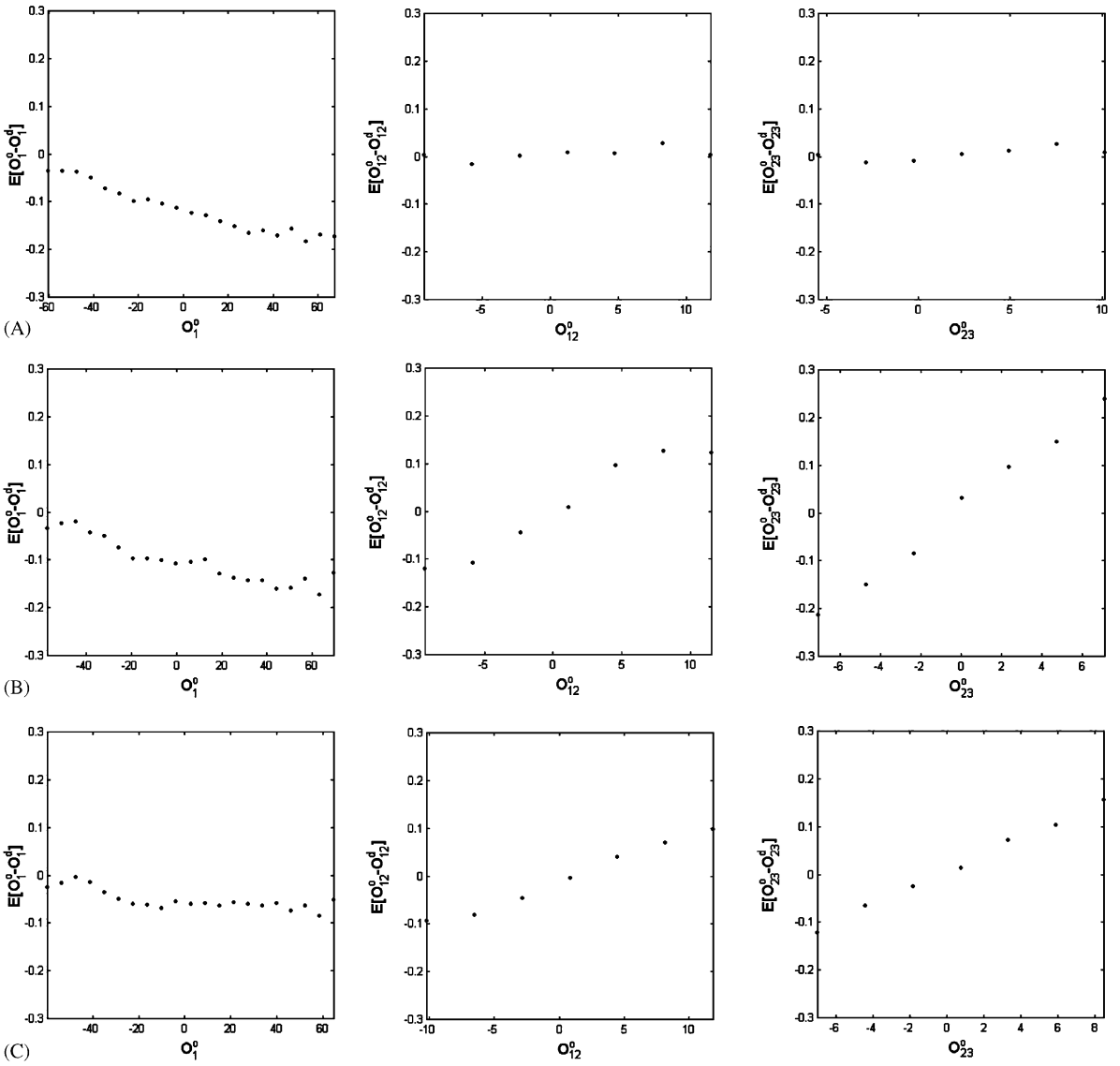

Fig. 1. Expected value of the coding–decoding error $E[O_n^o - O_n^d] = E[D_n]$ vs. O^o. The expected value is normalized with respect to the range of O^o. The following coders are analysed: (A) 8 kbps CS-CELP; and, (B) 5.3 kbps G.723 and (C) 13 kbps GSM-FR. The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23). The curves were obtained with 4500 utterances from 36 speakers.

proposes that the coding–decoding distortion should be modelled as feature dependent. This approach has not been found in the specialized literature, is interesting from both the theoretical and practical points of view, and could be applied to other adaptation problems.

2. Modelling the coding–decoding distortion as a function of cepstral parameters

The model described by (1), (2) and (3) describes well the distortion caused by the G.729 CS-CELP coder. However, according to Figs. 2 and 3, the

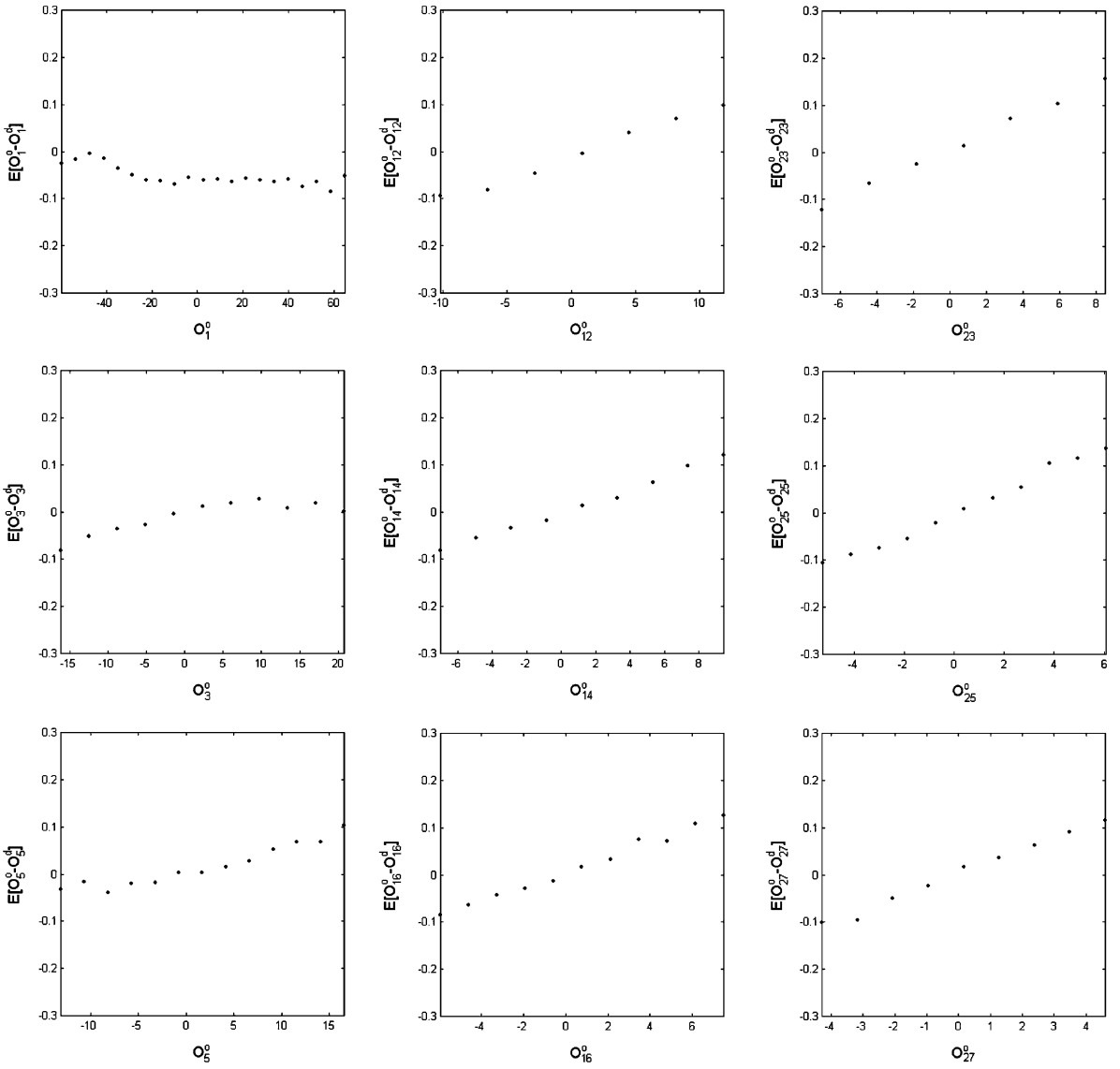

Fig. 2. Expected value of the coding–decoding error $E[O_n^{\mathrm{o}} - O_n^{\mathrm{d}}] = E[D_n]$ vs. O^{o} for the 13 kbps GSM-FR codec. The expected value is normalized with respect to the range of O^{o}. The cepstral coefficients correspond to a static (1, 3, 5), a delta (12, 14, 16) and a delta-delta (23, 25, 27). The curves were obtained with 4500 utterances from 36 speakers.

mean distortion $E[D_n] = E[O_{t,n}^{\mathrm{o}} - O_{t,n}^{\mathrm{d}}]$ in the GSM-FR (full-rate) and GSM-EFR (enhanced full-rate) coders clearly depends on the value of the cepstral coefficient. This dependence could be modelled as

$$E[D_n] = E\left[O_{t,n}^{\mathrm{o}} - O_{t,n}^{\mathrm{d}}\right] = B_n O_{t,n}^{\mathrm{o}} + A_n, \tag{4}$$

where B_n and A_n are constants. The expected value of the GSM (FR and EFR) coded–decoded distortion depends on the uncoded speech feature. Replacing (4) in (2) and applying the expected value, $E[O_{t,n}^{\mathrm{o}}]$ can be expressed as (see the Appendix):

$$E\left[O_{t,n}^{\mathrm{o}}\right] = \frac{O_{t,n}^{\mathrm{d}} + A_n}{1 - B_n}. \tag{5}$$

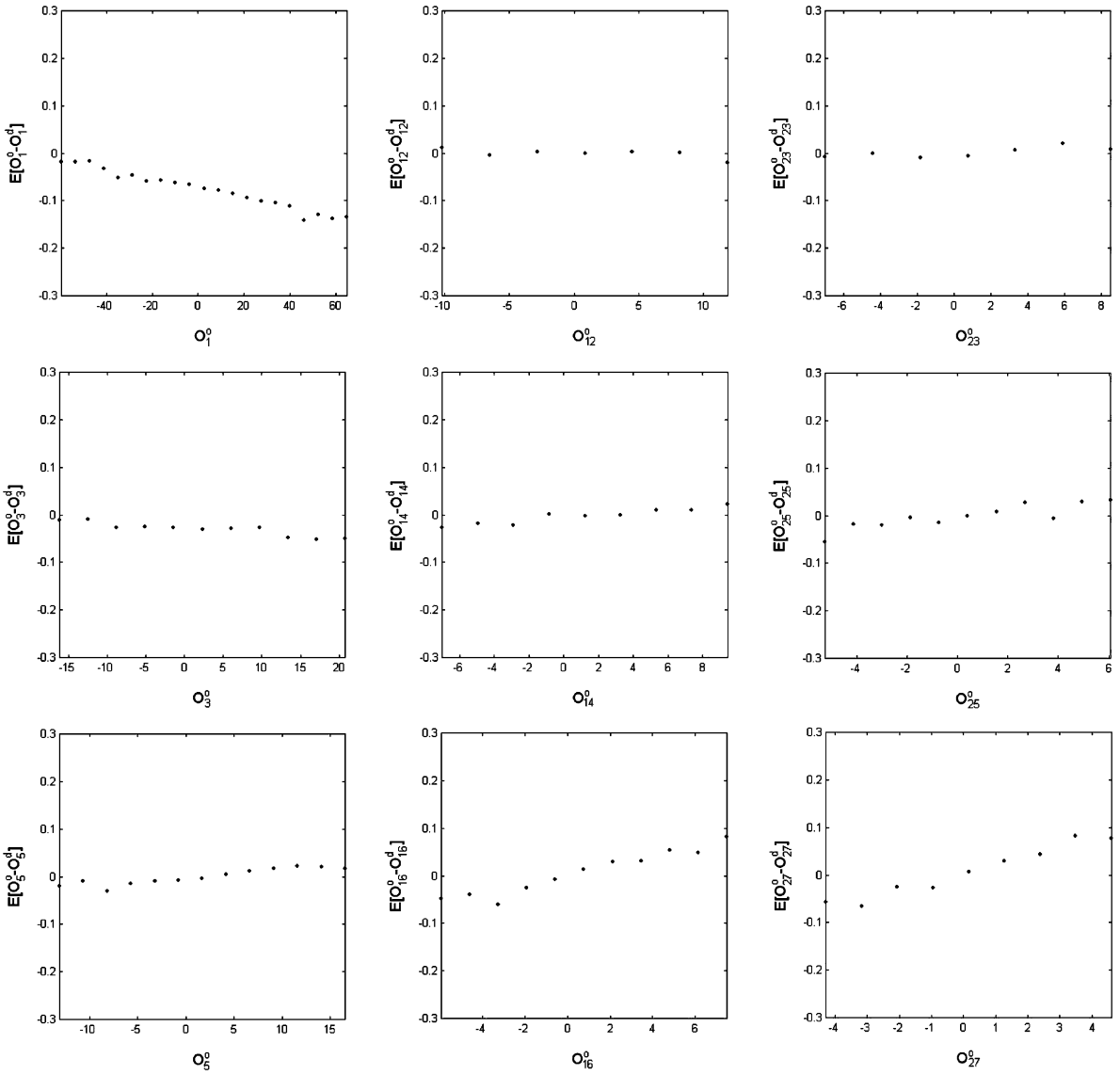

Fig. 3. Expected value of the coding–decoding error $E[O_n^{\mathrm{o}} - O_n^{\mathrm{d}}] = E[D_n]$ vs. O^{o} for the 12.2 kbps GSM-EFR codec. The expected value is normalized with respect to the range of O^{o}. The cepstral coefficients correspond to a static (1, 3, 5), a delta (12, 14, 16) and a delta–delta (23, 25, 27). The curves were obtained with 4500 utterances from 36 speakers.

Observe that $E[O_{t,n}^{\mathrm{d}}] = O_{t,n}^{\mathrm{d}}$ because $O_{t,n}^{\mathrm{d}}$ is an observed value and is a constant. The coding–decoding distortion variance, v_n^{d}, also depends on the value of the uncoded speech feature as seen in Fig. 4, although it seems to be more constant than $m_n^{\mathrm{d}} = E[D_n]$. Nevertheless, the introduction of this dependence does not result in analytical solutions in the EM estimation algorithm employed here. Consequently, v_n^{d} was supposed to be a constant as in (3), but the HMM was compensated with (5) instead of (2).

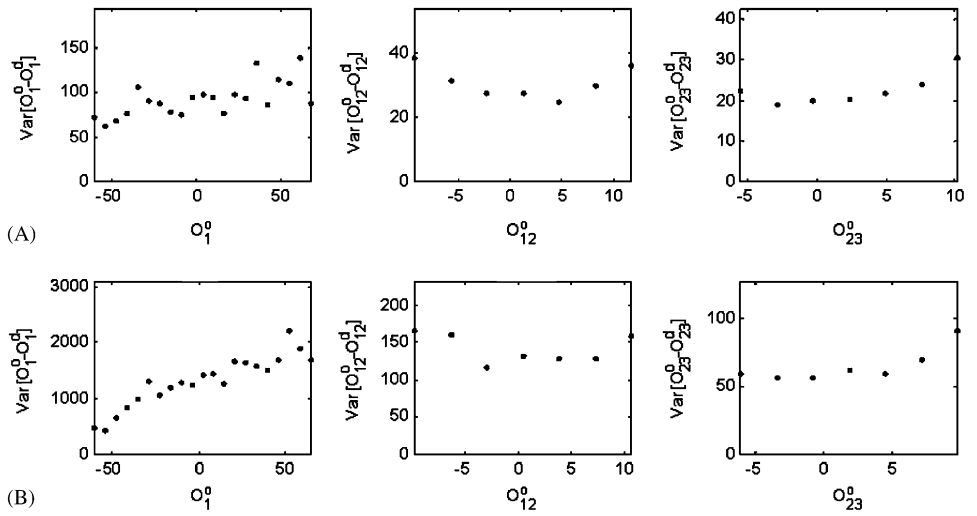

Fig. 4. Variance of the coding–decoding error $\mathrm{Var}[O_n^{\mathrm{o}} - O_n^{\mathrm{d}}] = \mathrm{Var}[D_n]$ vs. O^{o}. The following coders are analyzed: (A) 8 kbps CS–CELP; and, (B) 13 kbps GSM-FR. The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23). The curves were obtained with 4500 utterances from 36 speakers.

3. Estimation of feature-dependent coding–decoding distortion

Estimating the feature-dependent coding–decoding distortion is equivalent to find the vectors A_n, B_n and v_n^{d}. In this paper these parameters are estimated with the EM algorithm using a code-book, where every code-word corresponds to a multivariate Gaussian, built with uncoded speech signals. Inside each code-word cw_j the mean $\mu_j^{\mathrm{o}} = [\mu_{j,1}^{\mathrm{o}}, \mu_{j,2}^{\mathrm{o}}, \ldots, \mu_{j,n}^{\mathrm{o}}, \ldots, \mu_{j,N}^{\mathrm{o}}]$ and variance $(\sigma_j^{\mathrm{o}})^2 = [(\sigma_{j,1}^{\mathrm{o}})^2(, \sigma_{j,2}^{\mathrm{o}})^2, \ldots, (\sigma_{j,n}^{\mathrm{o}}), \ldots (\sigma_{j,N}^{\mathrm{o}})^2]$ are computed, where N is the number of cepstral coefficients and the dimension of the code-book. If there are J code-words, the p.d.f. associated with the frame $O_t^{\mathrm{o}} = [O_{t,1}^{\mathrm{o}}, O_{t,2}^{\mathrm{o}}, \ldots, O_{t,n}^{\mathrm{o}}, \ldots, O_{t,N}^{\mathrm{o}}]$ given the uncoded speech signal model is [6]:

$$f(O_t^{\mathrm{o}}/\Phi^{\mathrm{o}}) = \sum_{j=1}^{J} g(O_t^{\mathrm{o}}/\phi_j^{\mathrm{o}}) Pr(cw_j), \tag{6}$$

where $\Phi^{\mathrm{o}} = \{\phi_j^{\mathrm{o}} | 1 \leqslant j \leqslant J\}$ and $\phi_j^{\mathrm{o}} = (\mu_j^{\mathrm{o}}, \sum_j^{\mathrm{o}})$; $\Pr(\mathrm{cw}_j)$ is the a priori probability of code-word j; and $\sum_j^{\mathrm{o}}$ is the N-by-N covariance matrix that is supposed diagonal. $g(O_t^{\mathrm{o}}/\phi_j^{\mathrm{o}})$ is defined by

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

If A_n and B_n are considered 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 $\Phi^{\mathrm{d}} = \{\phi_{j,t}^{\mathrm{d}} | 1 \leqslant j \leqslant J\}$, where $\phi_{j,t}^{\mathrm{d}} = (\mu_{j,t}^{\mathrm{d}}, \sum_j^{\mathrm{d}})$ and,

$$\mu_{j,n,t}^{\mathrm{d}} = \mu_{j,n}^{\mathrm{o}} - E[D_n] = \mu_{j,n}^{\mathrm{o}} - B_n O_{t,n}^{\mathrm{o}} - A_n \tag{8}$$

$$(\sigma_{j,n}^{\mathrm{d}})^2 = (\sigma_{j,n}^{\mathrm{o}})^2 + v_n^{\mathrm{d}}. \tag{9}$$

Applying the expected value to (8) and replacing $E[O_{t,n}^{\mathrm{o}}]$ with (5), $\mu_{j,n,t}^{\mathrm{d}}$ can be written as

$$\mu_{j,n,t}^{\mathrm{d}} = \mu_{j,n}^{\mathrm{o}} - Z_n\left(A_n + O_{t,n}^{\mathrm{d}}\right) - A_n, \tag{10}$$

where $Z_n = B_n/(1 - B_n)$. In this paper A_n, B_n and v_n^{d} are estimated with the maximum likelihood criterion using adaptation utterances. The maximization of the likelihood does not lead to analytical solutions, so the EM algorithm was employed. Given an adaptation utterance O^{d} distorted by a coding–decoding scheme and composed of T frames, $O^{\mathrm{d}} = [O_1^{\mathrm{d}}, O_2^{\mathrm{d}}, \ldots, ; O_t^{\mathrm{d}}, \ldots, O_T^{\mathrm{d}}]$, O^{d} is also called observable data. The unobserved data is $Y^{\mathrm{d}} = [y_1^{\mathrm{d}}, y_2^{\mathrm{d}}, \ldots, y_t^{\mathrm{d}}, \ldots, y_T^{\mathrm{d}}]$ where y_t^{d} is the hidden number that refers to the code-word or density of the observed frame O_t^{d}. The EM algorithm defines the function $Q(\Phi, \hat{\Phi})$:

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

where $\hat{\Phi} = \{\hat{\phi}_{j,t}|1 \leqslant j \leqslant J, 1 \leqslant t \leqslant T\}$ and $\hat{\phi}_{j,t} = (\mu_{j,t}^{\mathrm{d}}, \sum_j^{\mathrm{d}})$ denotes the parameters that are estimated in an iteration by maximizing $Q(\Phi, \hat{\Phi})$. The maximization procedure corresponds to equalling to zero the partial derivatives of $Q(\Phi, \hat{\Phi})$ with respect to $\hat{\Pr}(\mathrm{cw}_j)$, $\hat{A}_n$ and $\hat{B}_n$, which in turn leads to the following algorithm (see the Appendix):

0. Initialization: $A_n = 0$, $B_n = 0$ $(Z_n = 0)$ and $v_n^{\mathrm{d}} = 0$.
1. Start with $\Phi = \Phi^{\mathrm{o}}$, where $\Phi = \{\phi_{j,t}|1 \leqslant j \leqslant J, 1 \leqslant t \leqslant T\}$ and $\phi_{j,t} = (\mu_j^{\mathrm{o}}, \sum_j^{\mathrm{o}})$. $\Pr(\mathrm{cw}_j)$ is initialized with the a priori probability of codeword j in the uncoded speech model defined in (6).
2. Compute $\Pr(\mathrm{cw}_j|O_t^{\mathrm{d}}, \phi_j)$,

$$\Pr(\mathrm{cw}_j|O_t^{\mathrm{d}}/\phi_j) = \frac{g(O_t^{\mathrm{d}}/\phi_{j,t}) \cdot \Pr(\mathrm{cw}_j)}{\sum_{k=1}^{J} g(O_t^{\mathrm{d}}/\phi_{k,t}) \cdot \Pr(\mathrm{cw}_k)}. \tag{12}$$

3. Estimate $\hat{\Pr}(\mathrm{cw}_j)$ with

$$\hat{\Pr}(\mathrm{cw}_j) = \frac{1}{T}\sum_{t=1}^{T} \Pr\left(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t}\right). \tag{13}$$

4. Estimate $\hat{A}_n$ with

$$\hat{A}_n = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})\frac{\left(\mu_{j,n}^{\mathrm{o}} - O_{t,n}^{\mathrm{d}} - Z_n O_{t,n}^{\mathrm{d}}\right)}{\sigma_{j,n}^2}\right)}{(Z_n + 1)\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\frac{\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})}{\sigma_{j,n}^2}\right)}. \tag{14}$$

5. Estimate $\hat{Z}_n$ with

$$\hat{Z}_n = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})\frac{\left(\mu_{j,n}^{\mathrm{o}} - A_n - O_{t,n}^{\mathrm{d}}\right)(A_n + O_{t,n}^{\mathrm{d}})}{\sigma_{j,n}^2}\right)}{\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\frac{\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})(A_n + O_{t,n}^{\mathrm{d}})^2}{\sigma_{j,n}^2}\right)} \tag{15}$$

and make $\hat{B}_n = \hat{Z}_n/(1 + \hat{Z}_n)$.

6. Update $\hat{\mu}_{j,n,t}^{\mathrm{d}}$ $1 < j < J$, $1 < n < N$, and $1 < t < T$ with (10).
7. Estimate $\hat{\sigma}_{j,n}^2$ for each code-book

$$\hat{\sigma}_{j,n}^2 = \frac{\sum_{t=1}^{T}\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})\left(O_{t,n}^{\mathrm{d}} - \hat{\mu}_{j,n,t}^{\mathrm{d}}\right)^2}{\sum_{t=1}^{T}\hat{\Pr}(\mathrm{cw}_j/O_t^{\mathrm{d}}, \phi_{j,t})}. \tag{16}$$

8. Estimate $\hat{v}_n^{\mathrm{d}}$:

$$\hat{v}_n^{\mathrm{d}} = \frac{\sum_{j=1}^{J}\left[\hat{\sigma}_{j,n}^2 - (\sigma_{j,n}^{\mathrm{o}})^2\right]\hat{\Pr}(\mathrm{cw_j})}{\sum_{j=1}^{J}\hat{\Pr}(\mathrm{cw_j})}. \tag{17}$$

9. Estimate the following convergence rates:

$$\Delta A_n = \frac{\left|A_n - \hat{A}_n\right|}{\hat{A}_n}, \tag{18}$$

$$\Delta B_n = \frac{\left|B_n - \hat{B}_n\right|}{\hat{B}_n}, \tag{19}$$

$$\Delta v_n^{\mathrm{d}} = \frac{\left|v_n^{\mathrm{d}} - \hat{v}_n^{\mathrm{d}}\right|}{\hat{v}_n^{\mathrm{d}}}. \tag{20}$$

10. Update parameters:

$$\Pr(\mathrm{cw}_j) = \hat{\Pr}(\mathrm{cw}_j)$$

$A_n = \hat{A}_n$; $B_n = \hat{B}_n$; $Z_n = \hat{Z}_n$, $\sigma^2_{j,n} = \hat{\sigma}^2_{j,n}$, and $v^{\mathrm{d}}_n = \hat{v}^{\mathrm{d}}_n$.

11. If convergence was reached, stop iteration; otherwise, go to step 2. Convergence is defined by the following condition:

$$A_n \leqslant \mathrm{CT} \text{ and } B_n \leqslant \mathrm{CT} \text{ and } v^{\mathrm{d}}_n \leqslant \mathrm{CT}, \qquad (21)$$

where CT is the convergence threshold and $1 < n < N$.

Note that the maximization of $Q(\Phi, \hat{\Phi})$ does not lead to an analytical solution for $\hat{v}^{\mathrm{d}}_n$, which in turn was estimated with (17). According to (17) $\hat{v}^{\mathrm{d}}_n$ is estimated as the averaged difference between the adapted and original code-word variance, $\sigma^2_{j,n}$ and $(\sigma^{\mathrm{o}}_{j,n})^2$, respectively. This averaged difference is weighted by the code-word probability $\Pr(\mathrm{cw_j})$. Finally, if $Z_n = 0$, the distortion model does not incorporate the relationship with the original cepstral uncoded parameters and the compensation is feature independent.

4. Experiments

The compensation method was tested with speaker-independent continuous speech recognition experiments using LATINO-40 database [9]. This database is composed of 40 Latin American native speakers, each reading 125 sentences from newspapers in Spanish. The training utterances were 4500 uncoded clean sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary has almost 6000 words. The testing database was composed of 500 utterances provided by 4 testing speakers (two females and two males). Each context-dependent phoneme was modelled with a 3-state left-to-right topology, with 8 multivariate Gaussian densities per state and diagonal covariance matrices. Trigram language modelling was employed. The frame energy plus ten Mel frequency cepstral coefficients (MFCC), and their first and second time derivatives were computed. The 500 testing uncoded signals were coded and decoded with the

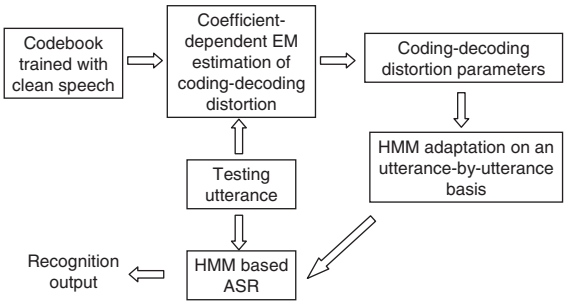

Fig. 5. Block diagram of the coding–decoding distortion compensation and recognition process proposed in this paper.

8 kHz G.729 CS-CELP, 13 kHz GSM-FR, 12.2 khz GSM-EFR and 5.3 khz G.723 coders to create the corrupted database. Fig. 5 shows a block diagram with the proposed compensation scheme. The techniques that were employed are indicated as follows: *baseline*, unmatched HMM trained with uncoded speech signals and without compensation; *feature-independent compensation*, unmatched HMM compensated with the EM algorithm presented here that estimates only A_n and v^{d}_n by making $B_n = Z_n = 0$; *feature-dependent compensation*, unmatched HMM compensated with the proposed EM algorithm that computes A_n, B_n and v^{d}_n; *matched-HMM*, HMM trained with speech signals that were coded and decoded; *CMN*, unmatched HMM with cepstral mean normalization; *CVN/CMN*, unmatched HMM with cepstral variance normalization (CVN) [10] and CMN; *RATZ*, unmatched HMM compensated with blind RATZ according to [6]; *supervised-ML,* unmatched HMM where the means and variances of the coding–decoding distortion, m^{d}_n and v^{d}_n, respectively, were estimated with the maximum likelihood criterion employing forced Viterbi alignment. *feature-independent compensation* and *feature-dependent compensation* computed the coding–decoding distortion on a sentence-by-sentence basis. The code-book was composed of 256 code-words and was generated with the uncoded training signals. The default value of the convergence threshold, *CT*, in (21) was made equal to 0.01. Due to practical restrictions, the maximum number of iterations in the EM algorithm was 20. The results are shown in Tables 1–3.

Table 1
WER (%) with signals processed with 8 kbps G.729 CS-CELP, 13 kbps GSM-FR, 12.2 kbps GSM-EFR, 5.3 kbps G.723 coders

CODER	Baseline	Matched-HMM	CMN	CVN/CMN	Feature-independent compensation	Feature-dependent compensation
GSM-FR	7.0	6.5	6.4	5.9	3.5	2.7
GSM-EFR	8.9	7.7	8.6	5.6	2.7	2.2
G.729 CS-CELP	11.2	7.5	10.3	5.9	3.4	3.0
G.723	12.0	9.5	10.0	7.3	3.4	2.7
Clean	5.9	5.9	5.9	5.8	2.6	2.0

Table 2
Comparison of the real time performance of *feature-dependent compensation* with *feature-independent compensation*. The codebook size was made equal to 8, 32 and 256 code-words or Gaussians. The signals were processed with the 12.2 kbps GSM-EFR coder

	Codebook size: 8 code-words	Codebook size: 32 code-words	Codebook size: 256 code-words
WER (%) feature-independent compensation	4.5	4.4	2.7
WER (%) feature-dependent-compensation	3.9	3.8	2.2
Times real time feature-independent compensation	1.0	3.8	26.4
Times real time feature-dependent compensation	1.3	4.9	45.0
Times real time feature-dependent compensation (CT in (21) was increased to make WER approximately equal to the one given by feature-independent compensation)	0.2	1.6	21.5

Table 3
Comparison of *feature-dependent compensation* with *RATZ* and *supervised-ML*. *AdU* denotes the number of adapting utterances. The signals were processed with the 12.2 kbps GSM-EFR coder

	Feature-dependent compensation $AdU = 1$	RATZ $AdU = 10$	Supervised_ML $AdU = 10$
WER (%)	2.7	7.1	8.3
Times real time (CT in (21) was increased to make WER approximately equal to the one given by feature-independent compensation)	21.5	25.2	0.4

The baseline system with uncoded speech gave a WER equal to 5.9%.

As can be seen in Table 1, the GSM-FR, GSM-EFR, G.729 CS-CELP and G723 coders gave WER equal to 7.0%, 8.9%, 11.2% and 12.0%, respectively, in experiments with the baseline system (*baseline*). In matched conditions, *matched-HMM*, WER was reduced by 19% on average when compared to *baseline*. This result is consistent with the one presented in [1]. *CMN* gave an average reduction of 9% in WER. This low improvement is due to the fact that the coding–decoding distortion model is different from the one for convolutional noise, which can be substantially

suppressed with CMN. *CVN/CMN* provided a higher improvement than *matched-HMM* and *CMN*: the average reduction in WER was 35% when compared to *baseline*. This must result from the fact that *CVN/CMN* attempts to normalize the mean and variance, which is more consistent with the coding–decoding distortion model employed here. Moreover, the combination *CVN/CMN* was applied on a sentence-by-sentence basis and was more effective than *matched-HMM* to cancel the coding–decoding distortion that is speaker dependent.

Also in Table 1, *feature-dependent compensation* gave an average reduction in WER equal to 72% and 57% when compared to *Baseline* and *CVN/CMN*, respectively. *CVN/CMN* strongly depends on the utterance length and does not make use of any model. On the other hand, *feature-dependent compensation* employs the information provided by the adapting utterance itself, the p.d.f. of the uncoded cepstral coefficients given by (6), and the model for coding–decoding distortion expressed in (4). Actually, the highest improvement was achieved with *feature-dependent compensation* that led to WER equal to 2.7%, 2.2%, 3.0% and 2.7% with the GSM-FR, GSM-EFR, G.729 CS-CELP and G723 coders, respectively. When compared to *feature-independent compensation*, *feature-dependent compensation* gave a reduction in WER of 23%, 19%, 12% and 21% with the GSM-FR, GSM-EFR, G.729 CS-CELP and G723 coders, respectively. Observe that the lowest improvement took place in experiments with G.729 CS-CELP where the difference between the feature-dependent and -independent models is less relevant according to Section 2 and Fig. 1. These reductions in WER strongly validate the model proposed in this paper. Significance analysis with the McNamar's test [11] shows that the improvements presented in Table 1 due to *feature-dependent compensation*, when compared to *feature-independent compensation*, are significant ($p<0.1$).

As can be seen in Table 1, *feature-dependent compensation* and *feature-independent compensation*, with only one adaptation utterance, dramatically reduced the effect of the GSM-FR, GSM-EFR, G.729 CS-CELP and G723 coders, and gave a WER lower than the baseline system with uncoded speech. This result is probably due to the fact that the approach proposed here also provides an adaptation to testing condition beyond the type of codification. The estimation of the vectors A_n, B_n, and v_n^{d} may also provide a speaker adaptation effect, for instance. When compared to the baseline system, *feature-dependent compensation* reduces the averaged difference between WER with distorted speech and clean signal from 3.9% to 0.7%. In other words, the proposed method substantially improves the robustness to coding–decoding distortion.

The EM algorithm employed in *feature-independent compensation* and *feature-dependent compensation* demands a high computational load. An exhaustive analysis to reduce the computational load of the EM estimation procedure is out of the scope of this paper. However, Table 2 presents results when the number of code-words or Gaussians in (6) was reduced from 256 to 8. The experiments were done with a PC Pentium IV 2.4 GHz.When compared to 256 code-words, the model with 8 Gaussians or code-words gave a WER 77% higher, but the computational load was 35 times lower. This result suggests that the algorithm proposed here can be optimized from the computational complexity point of view. Notice that, despite the fact that WER provided by *feature-dependent compensation* with 8 code-words is higher than the one with 256 Gaussians, this WER is still lower than the one given by the techniques in Table 1. Observe also that *feature-independent compensation* has a lower computational load than *feature-dependent compensation* with the same convergence criterion. However, *feature-dependent compensation* has a lower computational load if the convergence criterion, *CT* in (21), is softened to achieve the same WER as the one given by *feature-independent compensation*. This is due to the fact that the model employed by *feature-dependent compensation* is more accurate and the adaptation algorithm converges with fewer iterations.

Table 3 presents a comparison with *RATZ* and *supervised-ML*. Observe that both methods require a high number of adapting utterances, *AdU*. In the experiments reported here, *AdU* was made equal to 10 with *RATZ* and *supervised-ML* in order to achieve significant reductions in WER when

compared to *baseline*. This is consistent with [6] in the case of *RATZ* and with [7] in the case of *Supervised-ML*, although in [7] the adaptation procedure employed the forward–Backward algorithm. When compared to *RATZ* and *supervised-ML*, *feature-dependent compensation* in Table 1 gave a reduction in WER equal to 70% and 74%, respectively, with only one adapting utterance. This must be due to the fact that *RATZ* attempts to estimate the original cepstral coefficient, does not compensate the HMM by considering the original signal as a random variable, and does not employ a model for coding–decoding distortion. In the case of *supervised-ML*, the estimation is done by directly computing the means and variances of the coding–decoding distortion, m_n^{d} and v_n^{d}, respectively, without any assumption about the distribution of cepstral coefficients in uncoded speech as in (6) and about the coding–decoding distortion as in (4). As far as computational load is concerned, *feature-dependent compensation* can be more efficient than *RATZ* and also has a WER that is 62% lower. This results from the more accurate model employed by *feature-dependent compensation* that allows the estimation algorithm to converge faster than *feature-independent compensation* and *RATZ*. *Supervised-ML* is the technique that requires the lowest computational load, but it provides the poorest results and is a supervised approach, which in turn imposes serious restrictions on its applicability.

5. Conclusion

The feature-dependent compensation method proposed here dramatically compensates for the coding–decoding distortion, and can give a reduction in WER of 23%, 19%, 12% and 21% with GSM-FR, GSM-EFR, G.729 CS-CELP and G723 coders, respectively, when compared to the feature independent approach. When compared to the baseline system, the reduction in WER is as high as 70%. Moreover, the model employed by the feature-dependent compensation is more accurate than the one used by the feature-independent technique, and makes the adaptation algorithm converge in fewer iterations and give the same WER. As a consequence, the computational load requirement is reduced. This reduction depends on the number of code-words or Gaussians employed to model the distribution of cepstral coefficients in uncoded speech. It is worth emphasizing that the exhaustive analysis to decrease the computational load of the proposed method is out of the scope of the current paper.

In contrast to other methods already published in the specialized literature, the proposed technique is suitable for telephone dialogue systems because it needs only one adapting utterance. This is consistent with the fact that the coding–decoding distortion is speaker dependent. Moreover, a speaker adaptation effect may also take place, which in turn contributes to reduce WER in some circumstances. Finally, reducing the computational load of the estimation algorithm, the feature-dependent computation of the coding–decoding distortion variance, and the joint compensation of additive noise and coding–decoding distortion are proposed as future work.

Acknowledgement

The authors would like to thank Dr. Simon King, CSTR/University of Edinburgh, UK, for having proofread this manuscript. This work was supported by Conicyt, Fondecyt Proj. Nº 1030956 and Fondef Proj. Nº D02I-1089, Chile.

Appendix A

A.1. Given the model for the coding–decoding distortion:

$$E\left[O_{t,n}^{\mathrm{o}}\right] = O_{t,n}^{\mathrm{d}} + E[D_n]. \tag{A.1}$$

According to the model proposed in this paper, the expected value of the coding–decoding distortion is given by

$$E[D_n] = E\left[O_{t,n}^{\mathrm{o}} - O_{t,n}^{\mathrm{d}}\right] = B_n \cdot O_{t,n}^{\mathrm{o}} + A_n \tag{A.2}$$

Replacing (A.2) in (A.1)

$$E\left[O_{t,n}^{\mathrm{o}}\right] = O_{t,n}^{\mathrm{d}} + B_n O_{t,n}^{\mathrm{o}} + A_n \tag{A.3}$$

Taking the expected value of (A.3)

$$E\left[O^{o}_{t,n}\right] = O^{d}_{t,n} + B_n E\left[O^{o}_{t,n}\right] + A_n. \quad (A.4)$$

As a consequence, $E[O^{o}_{t,n}]$ can be written as

$$E\left[O^{o}_{t,n}\right] = \frac{O^{d}_{t,n} + A_n}{1 - B_n}. \quad (A.5)$$

A.2. Given the function $Q(\Phi, \hat{\Phi})$ expressed by (11) it can be shown that $Q(\Phi, \hat{\Phi})$ can be decomposed in two terms [12]

$$G = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr(\mathrm{cw}_j|O^{d}_t, \hat{\Phi})\log\left(\hat{\Pr}(\mathrm{cw}_j)\right) \quad (A.6)$$

and

$$H = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr(\mathrm{cw}_j|O^{d}_t, \Phi_j)\log\left(f(O^{d}_t|\mathrm{cw}_j, \hat{\Phi}_j)\right). \quad (A.7)$$

The probabilities $\hat{\Pr}(\mathrm{cw}_j)$ are estimated by means of maximizing G with the Lagrange method [12]

$$\hat{\Pr}(\mathrm{cw}_j) = \frac{1}{T}\sum_{t=1}^{T} \Pr\left(\mathrm{cw}_j|O^{d}_t, \phi_j\right). \quad (A.8)$$

The expressions (14) and (15) to estimate the distortion parameters A_n and B_n (or $Z_n = B_n/(1 - B_n)$) defined in (4) are derived by applying to H the gradient operator with respect to $\hat{A}_n$ and $\hat{Z}_n$, and setting the partial derivatives equal to zero

$$\frac{\partial H}{\partial(\hat{A}_n)} = 0, \quad (A.9)$$

$$\frac{\partial H}{\partial(\hat{Z}_n)} = 0 \quad (A.10)$$

where $\hat{A}_n$ and $\hat{Z}_n$ are the estimated parameters after one iteration of the EM algorithm. However, the coding–decoding distortion variance, v^{d}_n, cannot be estimated as in (A.9) and (A.10). This procedure does not lead to an analytical solution for v^{d}_n, and (17) described in Section 3 was adopted. Observe that (16) to estimate $\hat{\sigma}^2_{j,n}$ is also derived by equalling the partial derivative to zero

$$\frac{\partial H}{\partial(\hat{\sigma}^2_{j,n})} = 0 \quad (A.11)$$

References

[1] Huerta , J.M. Speech recognition in mobile environments, Ph.D. Thesis, Depaetment of Elec. and Comp. Engineering, Carnegie Mellon University, April 2000.

[2] S.V. Vaseghi, B.P. Milner, Noise compensation methods for Hidden Markov Model speech recognition in adverse environments, IEEE Trans. on Speech and Audio Processing 5 (1) (1997) 11–21.

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

[4] M.J.F. Gales, S.J. Young, HMM recognition in noise using parallel model combination, Proceedings of Eurospeech 93 (1993) 837–840.

[5] M.J.F. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Computer Speech and Language, Vol. 12, 1998.

[6] P. Moreno, Speech recognition in noisy environments, Ph.D. Thesis, Department of Elec. and Comp. Engineering, Carnegie Mellon University, 1996.

[7] M. Afify, Y. Gong, J.-P. Haton, A general joint additive and convolutive bias compensation approach applied to noisy lombard speech recognition, IEEE Trans. Speech Audio Process. 1998.

[8] N.B. Yoma, M. Villar, Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm, IEEE Trans. Speech and Audio Processing, 10 (3) (2002) 158–166.

[9] LDC, Latino-40 database provided by linguistic data consortium (LDC), University of Pennsylvania.

[10] R. Haeb-Umbach, Investigation on inter-speaker variability in the feature space. Proceedings of ICASSP 99.

[11] L. Gillik, S.J. Cox, Some statistical issues in the comparison of speech recognition algorithms. Proceedings of ICASSP' 89, 1989, pp. 532-535.

[12] X.D. Huang, et al., Hidden Markov Models for Speech Recognition, Edinburgh Univerity Press, 1990.