12

Uncertainty in Signal Estimation and Stochastic Weighted Viterbi Algorithm: A Unified Framework to Address Robustness in Speech Recognition and Speaker Verification

N. Becerra Yoma, C. Molina, C. Garreton and F. Huenupan
Speech Processing and Transmission Laboratory
Department of Electrical Engineering
Universidad de Chile
Chile

1. Introduction

Robustness to noise and low-bit rate coding distortion is one of the main problems faced by automatic speech recognition (ASR) and speaker verification (SV) systems in real applications. Usually, ASR and SV models are trained with speech signals recorded in conditions that are different from testing environments. This mismatch between training and testing can lead to unacceptable error rates. Noise and low-bit rate coding distortion are probably the most important sources of this mismatch. Noise can be classified into additive or convolutional if it corresponds, respectively, to an additive process in the linear domain or to the insertion of a linear transmission channel function. On the other hand, low-bit rate coding distortion is produced by coding – decoding schemes employed in cellular systems and VoIP/ToIP. A popular approach to tackle these problems attempts to estimate the original speech signal before the distortion is introduced. However, the original signal cannot be recovered with 100% accuracy and there will be always an uncertainty in noise canceling.

Due to its simplicity, spectral subtraction (SS) (Berouti et al., 1979; Vaseghi & Milner, 1997) has widely been used to reduce the effect of additive noise in speaker recognition (Barger & Sridharan, 1997; Drygajlo & El-Maliki, 1998; Ortega & Gonzalez, 1997), despite the fact that SS loses accuracy at low segmental SNR. Parallel Model Combination (PMC) (Gales & Young,1993) was applied under noisy conditions in (Rose et. al.,1994) where high improvements with additive noise were reported. Nevertheless, PMC requires an accurate knowledge about the additive corrupting signal, whose model is estimated using appreciable amounts of noise data which in turn imposes restrictions on noise stationarity, and about the convolutional distortion that needs to be estimated a priori (Gales, 1997). Rasta filtering (Hermansky et al., 1991) and Cepstral Mean Normalization (CMN) can be very useful to cancel convolutional distortion (Furui, 1982; Reynolds, 1994; van Vuuren, 1996) but, if the speech signal is also corrupted by additive noise, these techniques lose

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

effectiveness and need to be applied in combination with methods such as SS (Hardt & Fellbaum, 1997).

The idea of uncertainty in noise removal was initially proposed by the first author of this chapter in (Yoma et al., 1995; 1996-A; 1996-B; 1997-A; 1997-B; 1998-A; 1998-B; 1998-C; 1999) to address the problem of additive noise. The main idea was to estimate the uncertainty in noise canceling using an additive noise model and to weight the information provided by the signal according to the local SNR. As a consequence, Weighted DTW and Viterbi algorithms were proposed. Then, it was shown that convolutional noise could also be addressed in the framework of weighted matching algorithms. In (Yoma & Villar, 2001), the uncertainty in noise or distortion removal was modeled from the stochastic point of view. As a result, in the context of HMM, the original signal was modeled as a stochastic variable with normal distribution, which in turn leads to consider the expected value of the observation probability. If the observation probability is a Gaussian mixture, it is proved that its expected value is also a Gaussian mixture. This result, known as Stochastic Weighted Viterbi (SWV) algorithm, makes possible to address the problems of additive/convolutional (Yoma & Villar, 2001; 2002; Yoma et al., 2003-B), noise and low-bit rate coding distortion (Yoma et al., 2003-A; 2004; 2005; Yoma & Molina, 2006) in ASR and SV in a unified framework.

It is worth highlighting that SWV allows the interaction between the language and acoustic models in ASR just like in human perception: the language model has a higher weight in those frames with low SNR or low reliability (Yoma et al., 2003-B). Finally, the concept of uncertainty in noise canceling and weighted recognition algorithms (Yoma et al., 1995; 1996-A; 1996-B; 1997-A; 1997-B; 1998-A; 1998-B; 1998-C; 1999) have also widely been employed elsewhere in the fields of ASR and SV in later publications (Acero et al, 2006-A; 2006-B; Arrowood & Clements, 2004; Bernard & Alwan, 2002; Breton, 2005; Chan & Siu, 2004; Cho et al., 2002; Delaney, 2005; Deng, et al., 2005; Erzin et al., 2005; Gomez et al., 2006; Hung et al., 1998; Keung et al., 2000; Kitaoka & Nakagawa, 2002; Li, 2003; Liao & Gales, 2005 ; Pfitzinger, 2000; Pitsikalis et al., 2006; Tan et al., 2005; Vildjiounaite et al., 2006; Wu & Chen, 2001).

2. The model for additive noise

Given that $s(i)$, $n(i)$ and $x(i)$ are the clean speech, the noise and the resulting noisy signal, respectively, the additiveness condition in the temporal domain is expressed as:

$$x(i) = s(i) + n(i) \tag{1}$$

In the results discussed here, the signals were processed by 20 DFT mel filters. If inside each one of these DFT filters the phase difference between $s(i)$ and $n(i)$, and the energy of both signals are considered constant, the energy of the noisy signal at the output of the filter m, $\overline{x_m^2}$, can be modeled as (Yoma et al., 1998-B):

$$\overline{x_m^2} = \overline{s_m^2} + \overline{n_m^2} + 2 \cdot \sqrt{c_m} \sqrt{\overline{s_m^2}} \cdot \sqrt{\overline{n_m^2}} \cdot \cos(\phi) \tag{2}$$

where $\overline{s_m^2}$ and $\overline{n_m^2}$ are the energy of the clean speech and noise signals at the output of the filter m, respectively; ϕ is the phase difference, which is also considered constant inside

each one of the DFT mel filters, between the clean and noise signals; and c_m is a constant that was included due to the fact that these assumptions are not perfectly accurate in practice (Yoma et al., 1998-B); the filters are not highly selective, which reduces the validity of the assumption of low variation of these parameters inside the filters; and, a few discontinuities in the phase difference may occur, although many of them are unlikely in a short term analysis (i.e. a 25 ms frame). Nevertheless, this model shows the fact that there is a variance in the short term analysis and defines the relation between this variance and the clean and noise signal levels. Due to the approximations the variance predicted by the model is higher than the true variance for the same frame length, and the correction c_m had to be included. In (Yoma et al., 1998-B), this coefficient c_m was estimated with clean speech and noise-only frames. However, employing clean speech is not very interesting from the practical application point of view and in (Yoma & Villar, 2002) a different approach was followed by observing the error rate for a range of values of c_m. Solving (2), $\overline{s_m^2}$ can be written as:

$$\overline{s_m^2}(\phi, \overline{n_m^2}, \overline{x_m^2}) = 2 \cdot A_m^2 \cdot \cos^2(\phi) + B_m - 2 \cdot A_m \cdot \cos(\phi) \cdot \sqrt{A_m^2 \cdot \cos^2(\phi) + B_m} \tag{3}$$

where $A_m = \sqrt{\overline{n_m^2} \cdot c_m}$ and $B_m = \overline{x_m^2} - \overline{n_m^2}$. Notice that $\overline{n_m^2}$ can be replaced with an estimate of the noise energy made in non-speech intervals, $E\left[\overline{n_m^2}\right]$, $\overline{x_m^2}$ is the observed noisy signal energy and ϕ can be considered as a random variable. If $f_\phi(\phi)$, the probability density function of ϕ, is considered as being uniformly distributed between $-\pi$ and π, it can be shown that:

$$E\left[\log(\overline{s_m^2}(\phi))\right] = \int_{-\pi}^{\pi} \log(\overline{s_m^2}(\phi)) \cdot f_\phi(\phi) \cdot d(\phi) \cong \log(E[B_m]) \tag{4}$$

where $E[B_m] = \overline{x_m^2} - E\left[\overline{n_m^2}\right]$. To simplify the notation, $\overline{n_m^2}$ and $\overline{x_m^2}$ are withdrawn as arguments of the function $\overline{s_m^2}(\,)$ defined in (3). It is important to emphasize that $\overline{x_m^2} - E\left[\overline{n_m^2}\right]$ can be seen as the spectral subtraction (SS) estimation of the clean signal.

In (Yoma et al., 1998-A; 1998-B) the uncertainty in noise canceling was modeled as being the variance:

$$Var\left[\log(\overline{s_m^2}(\phi))\right] = E\left[\log^2(\overline{s_m^2}(\phi))\right] - E^2\left[\log(\overline{s_m^2}(\phi))\right] \tag{5}$$

where $E\left[\log^2(\overline{s_m^2}(\phi))\right]$ was computed by means of numerical integration.

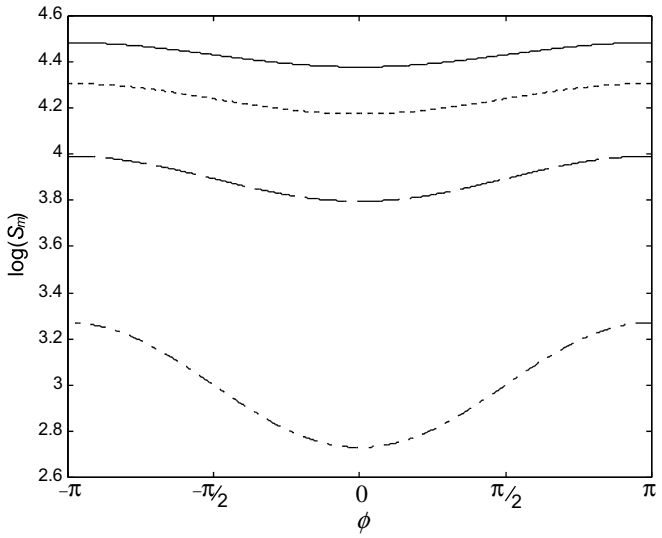


Figure 1. The log energy at filter m, $\log(S_m) = \log(\overline{s_m^2}(\phi))$, vs. ϕ, where $\overline{s_m^2}(\phi)$ is defined according to (5), for $\overline{x_m^2}/\overline{n_m^2}$ equal to 28 (——), 18 (- - - -), 8 (— — ·) and 2 (— · —). $\overline{n_m^2}$ was made equal to 1000 and c_m to 0.1.

2.1 Approximated expressions for the additive noise model

Figure 1 shows the function $\log\left(\overline{s_m^2}(\phi)\right)$, when $\overline{s_m^2}(\phi)$ is given by (3), for several values of the ratio $\overline{x_m^2}/\overline{n_m^2}$. As suggested in Fig.1, and easily verified in (3), the function $\log\left(\overline{s_m^2}(\phi)\right)$ is even and its minimum and maximum values are, respectively, $\log\left(\overline{s_m^2}(0)\right)$ and $\log\left(\overline{s_m^2}(\pi)\right)$ or $\log\left(\overline{s_m^2}(-\pi)\right)$. Employing $\log(1+x) \cong x$ for $x << 1$ and considering $B_m >> A_m^2$, which is easily satisfied at moderate SNR (greater or equal than 6dB), it is possible to show that (see appendix):

$$\log\left(\overline{s_m^2}(\phi)\right) \cong -\frac{2\cdot A_m}{\sqrt{B_m}}\cdot\cos(\phi) + E\left[\log(\overline{s_m^2}(\phi))\right] \;\cong\; -\frac{2\cdot A_m}{\sqrt{B_m}}\cdot\cos(\phi) + \log\left(E[B_m]\right) \tag{6}$$

Using (6), it can be shown that the uncertainty variance defined in (5) can be estimated with:

$$Var\left[\log\left(\overline{s_m^2}(\phi)\right)\right] \cong \frac{2E\left[A_m^2\right]}{E[B_m]} \tag{7}$$

where $E\left[A_m^2\right] = c_m \cdot E\left[\overline{n_m^2}\right]$ and $E[B_m]$ is defined above. Due to the fact that (6) and (7) are derived considering that $B_m >> A_m^2$, this condition imposes a domain where these expressions can be used. Assuming that B needs to be greater or equal than $10 \cdot A_m^2$, to satisfy the condition above, means that (7) is valid when $\overline{x_m^2} - E\left[\overline{n_m^2}\right] \geq 10 \cdot c_m \cdot E\left[\overline{n_m^2}\right]$.

When $\overline{x_m^2} - E\left[\overline{n_m^2}\right] < 10 \cdot c_m \cdot E\left[\overline{n_m^2}\right]$ a linear extrapolation could be used and (7) is modified to:

$$Var\left[\log\left(\overline{s_m^2}(\phi)\right)\right] = \begin{cases} \dfrac{2 \cdot c_m \cdot E\left[\overline{n_m^2}\right]}{\overline{x_m^2} - E\left[\overline{n_m^2}\right]} & \text{if } \overline{x_m^2} - E\left[\overline{n_m^2}\right] \geq 10 \cdot c_m \cdot E\left[\overline{n_m^2}\right] \\ -\dfrac{\overline{x_m^2} - E\left[\overline{n_m^2}\right]}{50 \cdot c_m \cdot E\left[\overline{n_m^2}\right]} + 0.4 & \text{if } \overline{x_m^2} - E\left[\overline{n_m^2}\right] < 10 \cdot c_m \cdot E\left[\overline{n_m^2}\right] \end{cases} . \quad (8)$$

2.2 Spectral subtraction

As mentioned above, (4) could be considered as a definition for SS (spectral subtraction). However, (4) presents the same problems at low SNR when the additive noise model loses accuracy and $E[B_m] = \overline{x_m^2} - E\left[\overline{n_m^2}\right]$ can be negative, which in turn is incompatible with the log operator. In (Yoma & Villar, 2003) the clean signal was estimated using the SS defined as:

$$SSE_m = \max\left\{\overline{x_m^2} - E\left[\overline{n_m^2}\right] \; ; \; \beta \cdot \overline{x_m^2}\right\} \quad (9)$$

which corresponds to a simplified version of an SS defined in (Vaseghi & Milner, 1997). SSE_m denotes the estimation of the clean signal energy by means of SS.

In order to improve the applicability at low segmental SNR of the additive noise model discussed here, some modifications would be necessary: first, the domain of ϕ requires to be modified, affecting the integral in (4), to satisfy the condition $\overline{s_m^2}(\phi) \geq 0$; second, the noise energy $\overline{n_m^2}$ should also be treated as a random variable at low SNR, but the estimation of its distribution may require long non-speech intervals, which imposes restrictions on the dynamics of the corrupting additive process; third, a more accurate model should also take into consideration an a priori distribution of the clean speech energy. Consequently, employing the SS defined as in (9) is an interesting compromise between the applicability of the approach proposed here and the theoretical model for the addition of noise discussed in section 2. The SS as in (9) reduces the distortion at low SNR by setting a lower threshold proportional to the noisy signal energy.

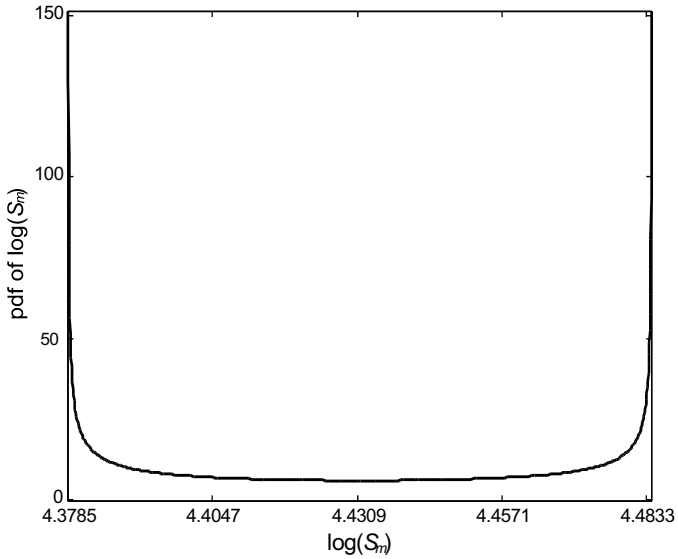


Figure 2. Probability density function of $\log(S_m) = \log(\overline{s_m^2}(\phi))$ assuming that ϕ is a random variable uniformly distributed between $-\pi$ and π. $\overline{x_m^2}/\overline{n_m^2}$ was made equal to 28, $\overline{n_m^2}$ to 1000 and c_m to 0.1. The p.d.f. curve of $\log(S_m)$ was estimated using the following theorem (Papoulis, 1991): to find $f_y(y)$ for a specific y, the equation $y = g(x)$ is solved; if its real roots are denoted by x_n, then $f_y(y) = f_x(x_1)/|g'(x_1)| + \cdots + f_x(x_n)/|g'(x_n)|$ where $g'(x)$ is the derivative of $g(x)$. In this case $y = \log(\overline{s_m^2}(\phi))$ and $x = \phi$.

2.3 Uncertainty variance in the cepstral domain

Most speech recognizers and speaker verification systems compute cepstral coefficients from the filter log energies. The static cepstral coefficient C_n is defined as:

$$C_n = \sum_{m=1}^{M} \log\left(\overline{s_m^2}(\phi)\right) \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right) \tag{10}$$

where M is the number of DFT filters. Observing that (10) is a sum and assuming that $\log\left(\overline{s_m^2}(\phi)\right)$ with $1 \le m \le M$ are independent random variables, C_n tends to a random variable with Gaussian distribution according to the Central Limit Theorem (Papoulis, 1991). The independence hypothesis is strong but substantially simplifies the mapping between the log and cepstral domain for the uncertainty variance. Consequently, the variance of C_n is given by (Yoma et al., 1998-A; Yoma & Villar, 2002):

$$Var[C_n] = \sum_{m=1}^{M} Var\left[\log\left(\overline{s_m^2}(\phi)\right)\right]\cdot\cos^2\left(\frac{\pi\cdot n}{M}\cdot(m-0.5)\right). \quad (11)$$

In order to counteract the limitation discussed in section 2.2, $E\left[\log\left(\overline{s_m^2}(\phi)\right)\right]$ was replaced with $\log(SSE_m)$, where SSE_m is defined according to (9), to estimate $E[C_n]$:

$$E[C_n] = \sum_{m=1}^{M} \log(SSE_m)\cdot\cos\left(\frac{\pi\cdot n}{M}\cdot(m-0.5)\right). \quad (12)$$

The probability density functions (p.d.f.) of $\log\left(\overline{s_m^2}(\phi)\right)$ and C_n are shown in Figs.2 and 3. As can be seen in Fig.3, approximating the distribution of C_n with a Gaussian seems a reasonable approach.

Considering the variables $\log\left(\overline{s_m^2}(\phi)\right)$ as being independent should be interpreted as a hypothesis that is inaccurate for contiguous filters but more realistic when the separation between filters increases. This assumption is able to simplify the formulation of the approach proposed here and to lead to significant improvements in the system performance as shown later. Assuming $\log\left(\overline{s_m^2}(\phi)\right)$ is correlated requires a more complex analysis to estimate the uncertainty variance in the cepstral domain and the distribution of the cepstral coefficients of the hidden clean signal. This analysis, which would incorporate further knowledge about the speech signal in the spectral domain but also would make the estimation of the expected value of the output probability in section 3 more difficult, is not addressed in (Yoma & Villar, 2002) although could still lead to some improvements when compared with the current model.

In speech recognition and speaker verification systems delta cepstral coefficients are used in combination with the static parameters. The delta cepstral coefficient in frame t, $\delta C_{t,n}$ is defined as:

$$\delta C_{t,n} = \frac{C_{t+1,n} - C_{t-1,n}}{2}. \quad (13)$$

where $C_{t+1,n}$ and $C_{t-1,n}$ are the static cepstral features in frames $t+1$ and $t-1$. If the frames are supposed uncorrelated, the same assumption made by HMM, the uncertainty mean and variance of $\delta C_{t,n}$ are, respectively, given by:

$$E\left[\delta C_{t,n}\right] = \frac{E\left[C_{t+1,n}\right] - E\left[C_{t-1,n}\right]}{2}. \quad (14)$$

$$Var\left[\delta C_{t,n}\right] = \frac{Var\left[C_{t+1,n}\right] + Var\left[C_{t-1,n}\right]}{4}. \quad (15)$$

Concluding, the cepstral coefficients could be treated as random variables with normal distribution whose mean and variance are given by (12) (11) and (14) (15). As a result, the HMM output probability needs to be modified to represent the fact that the spectral features should not be considered as being constants in noisy speech.

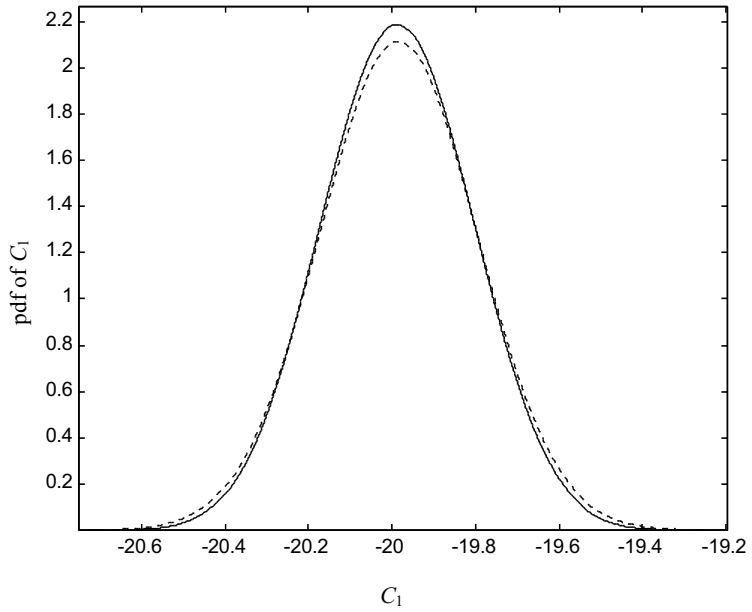


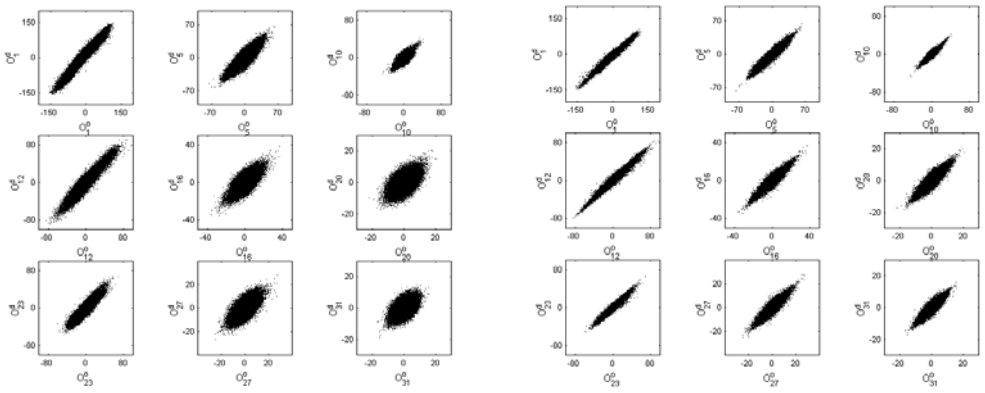
Figure 3. Probability density function of the static cepstral coefficient C_1 computed with 20 log energies $\log(\overline{s_m^2}(\phi))$. As a consequence, this density function corresponds to the convolution (———) of 20 p.d.f.´s similar to the one shown in Fig. 2. The theoretic Normal p.d.f. with the same mean and variance is represented with (– – –).

3. Modelling low-bit rate coding-decoding distortion

As discussed in (Yoma et al., 2006), 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 (ITU-T, 1996) 13 kbps GSM (ETSI, 1992), 5.3 kbps G723.1 (ITU-T, 1996-B), 4.8 kbps FS-1016 (Campbell et al, 1991) and 32 kbps ADPCM (ITU-T, 1990). After that, the original and coded-decoded speech signals, which were sampled at a rate of 8000 samples/second, were divided in 25ms frames with 12.5ms 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. 4-9.

It is worth mentioning that the estimation and compensation of the coding-decoding distortion proposed in (Yoma et al., 2006) was tested with SI continuous speech recognition experiments using LATINO-40 database (LDC, 1995). 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). 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. Trigram language model was employed during recognition.

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 respect to the diagonal axis in the 8 kbps CS-CELP (Fig. 4a) and in the 32 kbps ADPCM (Fig. 4b). This suggests that the coding-decoding distortion, defined as $D_n = O_n^o - O_n^d$, 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_n^o in those cases. However, the same behavior is not observed in the 13 kbps GSM coder (Fig. 5) where the pairs (O_n^o, O_n^d) seems to be symmetrically distributed around a center near (0, 0).



(a) 8 kbps CS-CELP (b) 32 kbps ADPCM

Figure 4. Cepstral coefficients from uncoded (O^o) vs. 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.

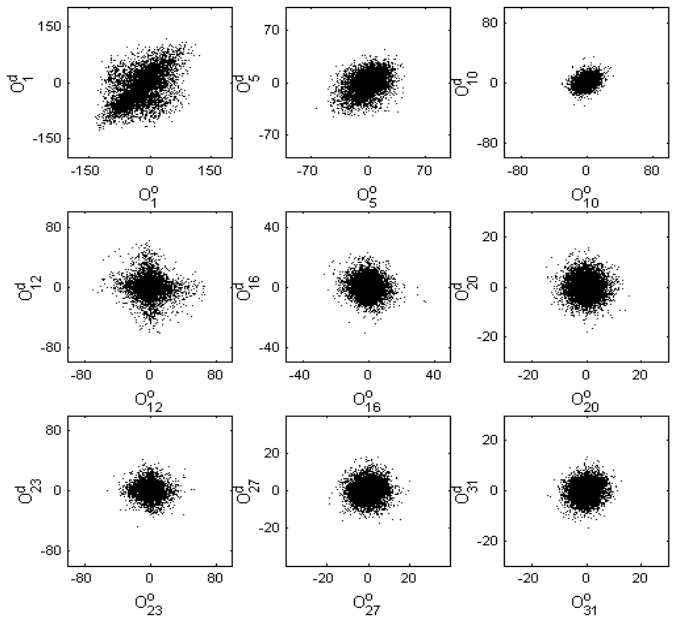


Figure 5. Cepstral coefficients from uncoded (O^o) vs. coded-decoded (O^d) speech signals. The coder is the 13 kbps GSM from the ETSI GSM-06.10 Full Rate Speech Transcoding. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral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

The histograms presented in Fig. 6 (8 kbps CS-CELP) and Fig. 7 (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 (Yoma et al., 2006). The expected value, normalized with respect to the range of the observed O^o_n, of the coding-decoding distortion vs. O^o_n is shown in Fig. 8. Notice that the dependence of the expected value on O^o_n 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^o_n itself and displays an odd symmetry. It is interesting to emphasize that the fuzzy circular-like (O^o_n,O^d_n) patterns observed with the 13 kbps GSM (Fig. 5) 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 vs. O^o_n is shown in Fig. 9. According to Fig. 9, the assumption related to the independence of the variance with respect O^o_n does not seem to be unrealistic. Moreover, this assumption is strengthen by the fact that the distribution of O^o_n tends to be concentrated around $O^o_n = 0$.

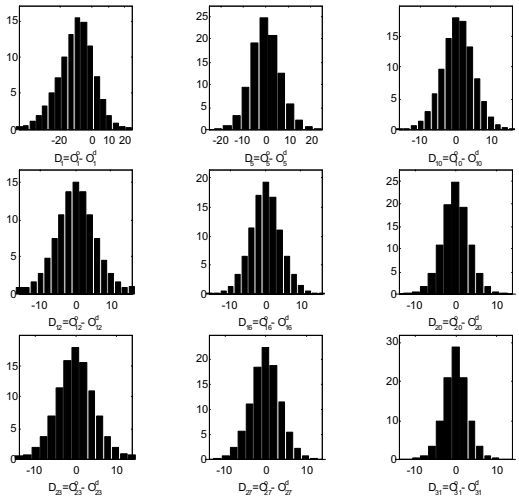


Figure 6. Distribution of coding distortion ($O^o - O^d$) with signals processed by 8 kbps CS-CELP from the ITU-T standard G.729. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated with the same data employed in Fig. 4.

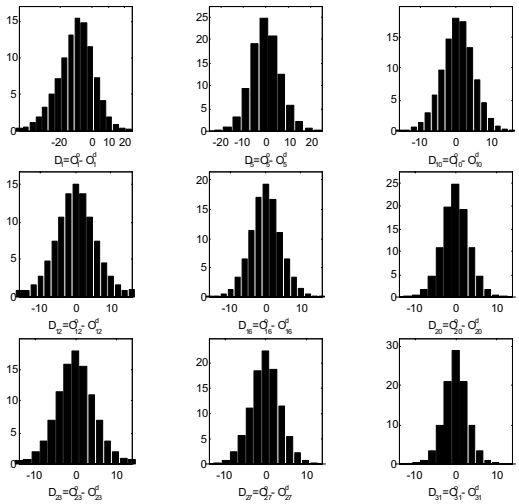


Figure 7. Distribution of coding distortion ($O^o - O^d$) with signals processed by 5.3 kbps G723-1 from the ITU-T standard G.723.1. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated with the same data employed in Fig. 4.

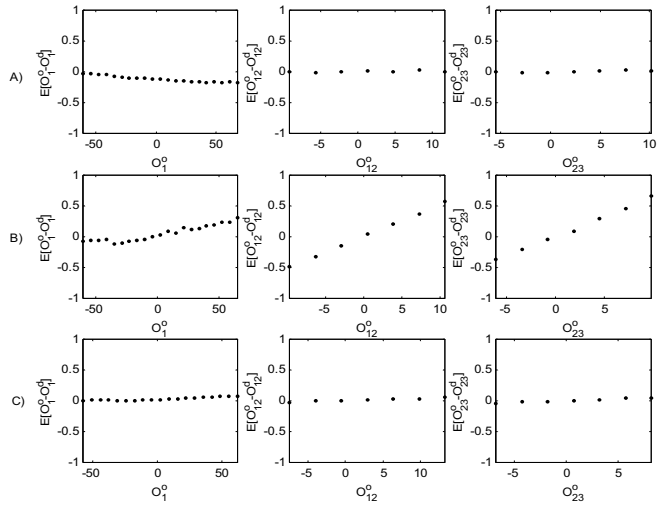


Figure 8. Expected value of the coding-decoding error, $E[O_n^o - O_n^d] = m_n^d$, vs. O^o. The expected value is normalized with respect to the range of observed O^o. The following coders are analyzed: A) 8 kbps CS-CELP; B) 13 kbps GSM; and, C) 32 kbps ADPCM . The cepstral coefficients correspond to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

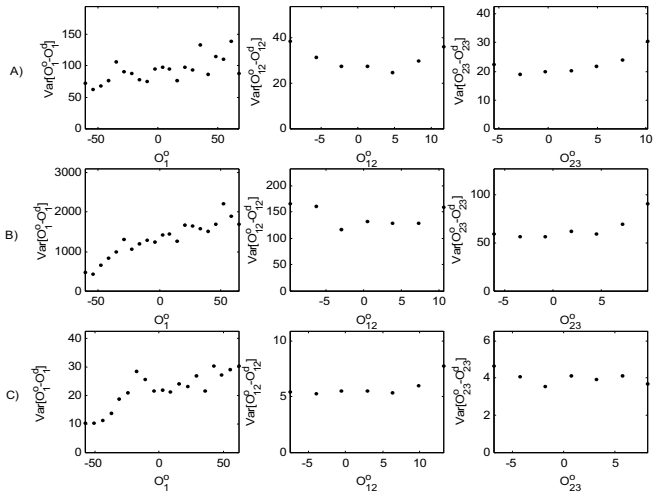


Figure 9. Variance of the coding-decoding error, $Var[O_n^o - O_n^d] = v_n^d$, vs. O^o. The following coders are analyzed: A) 8 kbps CS-CELP; B) 13 kbps GSM; and, C) 32 kbps ADPCM . The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23).

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, $O^o_{t,n}$, could be given by (Yoma et al., 2006):

$$O^o_{t,n} = O^d_{t,n} + D_n \tag{16}$$

where $O^d_{t,n}$ 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. $f_{D_n}(D_n) = N(m^d_n, v^d_n)$ that does not depend on the value of the cepstral coefficient n, and therefore the phonetic class; $N(m^d_n, v^d_n)$ is a Gaussian distribution with mean m^d_n and variance v^d_n. 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 (Yoma et al., 2006).
In a real situation, $O^d_{t,n}$ is the observed cepstral parameter and $O^o_{t,n}$ is the hidden information of the original speech signal. From (16), the expected value of $O^o_{t,n}$ is given by:

$$E\left[O^o_{t,n}\right] = O^d_{t,n} + m^d_n \tag{17}$$

Concluding, according to the model discussed in this section, the distortion caused by the coding-decoding scheme is represented by the mean vector $M^d = \left[m^d_1, m^d_2, m^d_3, ... m^d_n, ..., m^d_N\right]$ and the variance vector $V^d = \left[v^d_1, v^d_2, v^d_3, ... v^d_n, ..., v^d_N\right]$. Moreover, this distortion could be considered independent of the phonetic class and is consistent with the analysis presented in (Huerta, 2000).

4. Estimation of coding-decoding distortion

In this section the coding-decoding distortion as modeled in section 3 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 above. In (Yoma et al., 2006) 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 p.d.f. 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 $\mu^o_j = \left[\mu^o_{j,1}, \mu^o_{j,2}, ..., \mu^o_{j,N}\right]$ and variance $(\sigma^o_j)^2 = \left[(\sigma^o_{j,1})^2, (\sigma^o_{j,2})^2, ..., (\sigma^o_{j,N})^2\right]$ are computed, and the distribution of frames in the cells is supposed to be Gaussian:

$$f(O^o_t / \phi^o_j) = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\Sigma^o_j\right|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\left(O^o_t - \mu^o_j\right)^t \left(\Sigma^o_j\right)^{-1} \left(O^o_t - \mu^o_j\right)} \tag{18}$$

where N is the number of cepstral coefficients and also the dimension of the code-book; Σ_j^o is the N-by-N covariance matrix that is supposed diagonal; and, $\phi_j^o = \left(\mu_j^o, \Sigma_j^o\right)$. 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:

$$f(O_t^o / \Phi^o) = \sum_{j=1}^{J} f(O_t^o \mid \phi_j^o) \cdot \Pr(cw_j) \tag{19}$$

where $\Phi^o = \left\{\phi_j^o \;\middle|\; 1 \le j \le J\right\}$ denotes all the means and variances of the code-book. Equation (19) 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 $\Phi^d = \left\{\phi_j^d \middle| 1 \le j \le J\right\}$, where $\phi_j^d = \left(\mu_j^d, \Sigma_j^d\right)$ and,

$$\mu_j^d = \mu_j^o - M^d \tag{20}$$

$$(\sigma_{j,n}^d)^2 = (\sigma_{j,n}^o)^2 + v_n^d \tag{21}$$

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 (Yoma et al., 2006) 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 (Huang et al., 1990; Moon, 1996) was employed. Given an adaptation utterance O^d distorted by a coding-decoding scheme and composed of T frames,

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

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

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

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:

$$Q(\Phi, \hat{\Phi}) = E\left[\log\left(f(O^d, Y^d / \hat{\Phi})\right) \middle|\ O^d, \Phi\right] \tag{22}$$

where $\hat{\Phi} = \left\{\hat{\phi}_j \middle| 1 \le j \le J\right\}$, where $\hat{\phi}_j = \left(\mu_j^d, \Sigma_j^d\right)$ denotes the parameters that are estimated in an iteration by maximizing $Q(\Phi, \hat{\Phi})$. It can be shown that (22) can be decomposed in two terms:

$$A = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr(cw_j \mid O_t^d, \hat{\Phi}) \cdot \log\left(\hat{\Pr}(cw_j)\right) \tag{23}$$

and

$$B = \sum_{t=1}^{T}\sum_{j=1}^{J} \Pr(cw_j \mid O_t^d, \Phi_j) \cdot \log\left(f(O_t^d \mid cw_j, \hat{\Phi}_j)\right) \tag{24}$$

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

$$\hat{\Pr}\left(cw_j\right) = \frac{1}{T}\sum_{t=1}^{T} \Pr\left(cw_j \mid O_t^d, \phi_j\right) \tag{25}$$

The distortion parameters defined in (16) 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 $\Phi = \left\{\phi_j \middle| 1 \le j \le J\right\}$ and $\phi_j = \left(\mu_j, \Sigma_j\right)$.
2. Compute $\Pr(cw_j \middle| O_t^d, \phi_j)$

$$\Pr(cw_j \middle| O_t^d, \phi_j) = \frac{f(O_t^d \mid \phi_j) \cdot \Pr(cw_j)}{\sum_{k=1}^{J} f(O_t^d \mid \phi_k) \cdot \Pr(cw_k)} \tag{26}$$

3. Estimate $\hat{\Pr}(cw_j)$ with (25)
4. Estimate $\Delta\mu_n$ with

$$\Delta\mu_n = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\hat{\Pr}(cw_j \mid O_t^d, \phi_j) \cdot \frac{\left(O_{t,n}^d - \mu_{j,n}\right)}{\sigma_{j,n}^2}\right)}{\sum_{t=1}^{T}\sum_{j=1}^{J}\left(\frac{\hat{\Pr}(cw_j \mid O_t^d, \phi_j)}{\sigma_{j,n}^2}\right)} \tag{27}$$

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

$$\hat{\mu}_{j,n} = \mu_{j,n} + \Delta\mu_n \tag{28}$$

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

$$\hat{\sigma}_{j,n}^2 = \frac{\sum_{t=1}^{T}\hat{\Pr}(cw_j \mid O_t^d, \phi_j) \cdot \left(O_{t,n}^d - \hat{\mu}_{j,n}\right)^2}{\sum_{t=1}^{T}\hat{\Pr}(cw_j \mid O_t^d, \phi_j)} \tag{29}$$

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

$$f(O^d / \hat{\Phi}) = \sum_{t=1}^{T} \sum_{j=1}^{J} f(O_t^d \mid \hat{\phi}_j) \cdot \hat{\Pr}(cw_j) \tag{30}$$

8. Update parameters:

$$\Phi = \hat{\Phi}$$

$$\Pr(cw_j) = \hat{\Pr}(cw_j)$$

9. If convergence was reached, stop iteration; otherwise, go to step 2.
10. Estimate M^d and V^d:

$$m_n^d = -(\mu_{j,n} - \mu_{j,n}^o) \tag{31}$$

for any $1 < j < J$, and

$$v_n^d = \frac{\sum_{j=1}^{J} \left[\sigma_{j,n}^2 - (\sigma_{j,n}^o)^2 \right] \cdot \Pr(\mathrm{cw_j})}{\sum_{j=1}^{J} \Pr(\mathrm{cw_j})} \tag{32}$$

where $1 < n < N$. If $v_n^d < 0$, v_n^d is made equal to 0.

It is worth observing that (27) was derived with $\frac{\partial B}{\partial(\Delta\mu_n)} = 0$, where B is defined in (24), $\hat{\mu}_{j,n} = \mu_{j,n} + \Delta\mu_n$ corresponds to the re-estimated code-word mean in an iteration. Expression (29) was derived by $\frac{\partial B}{\partial \hat{\sigma}_{j,n}^2} = 0$. Moreover, expressions (31) and (32) assume that the coding-distorting is independent of the code-word or class, and (32) 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 (Gauvain & Lee, 1994) 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 in (Yoma et al., 2006) 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 (Gales, 1998), the proposed computation of the coding-decoding distortion requires fewer parameters to estimate, although it should still lead to high improvements in word accuracy as a speaker adaptation method. Finally, the method discussed in this section to estimate the coding-decoding distortion is similar to

the techniques employed in (Acero and Stern, 1990; Moreno et al., 1995; Raj et al., 1996) to compensate additive/convolutional noise and estimate the unobserved clean signal. In those papers the p.d.f. 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, (Yoma et. al, 2006) 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 (Acero & Stern, 1990; Moreno et al., 1995; Raj et al., 1996). In contrast, M^d and V^d are considered independent of the code-word in (Yoma et. al, 2006). 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 (27) to estimate M^d is the same expression employed to estimate convolutional distortion (Acero & Stern, 1990) if additive noise is not present (Yoma, 1998-B), the methods in (Acero & Stern, 1990; Moreno et al., 1995; Raj et al., 1996) 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 (Yoma & Villar, 2002), which leads to a variance compensation that clearly depends on the phonetic class and code-word.

5. The expected value of the observation probability: The Stochastic Weighted Viterbi algorithm

In the ordinary HMM topology the output probability of observing the frame O_t at state s, $b_s(O_t)$, is computed, either in the training or in the testing algorithms, considering O_t as being a vector of constants. As can be seen in (Yoma & Villar; 2002; Yoma et al., 2006) the observation vector is composed of static, delta and delta-delta cepstral coefficients, and according to sections 2 and 3 these parameters should be considered as being random variables with normal distributions when the speech signal is corrupted by additive noise and coding-decoding distortion. Therefore, to counteract this incompatibility (Yoma & Villar; 2002) proposes to replace, in the Viterbi algorithm, $b_s(O_t)$ with $E[b_s(O_t)]$ that denotes the expected value of the output probability. This new output probability, which takes into consideration the additive noise model, can be compared an empiric weighting function previously proposed in (Yoma et al., 1998-B).

5.1 An empiric weighting function

The uncertainty in noise canceling variance was estimated in each one of the DFT mel filters and employed to compute a coefficient $w(t)$ to weight the information provided by the frame t (Yoma et al., 1998-B). This weighting coefficient was included in the Viterbi algorithm by means of raising the output probability of observing the frame O_t at state s, $b_s(O_t)$, to the power of $w(t)$. The weighting parameter was equal to 0 for noise-only signal and equal to 1 for clean speech. As a consequence, if $w(t)$=0, $[b_s(O_t)]^{w(t)} = 1$ that means that the frame does not give any reliable information. This weighted Viterbi algorithm was able to show reductions in the error as high as 80 or 90% in isolated word speech recognition experiments. However, the approach presented some drawbacks: first, the function to