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.

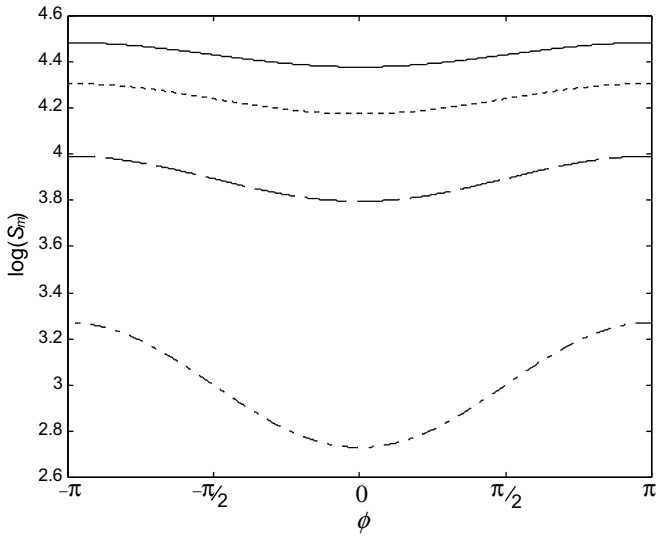


Figure 1. The log energy at filter m, $\log(S_m) = \log(\overline{s_m^2}(\phi))$, vs. ϕ, where $\overline{s_m^2}(\phi)$ is defined according to (5), for $\overline{x_m^2}/\overline{n_m^2}$ equal to 28 (——), 18 (- - - -), 8 (— — ·) and 2 (— · —). $\overline{n_m^2}$ was made equal to 1000 and c_m to 0.1.

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.

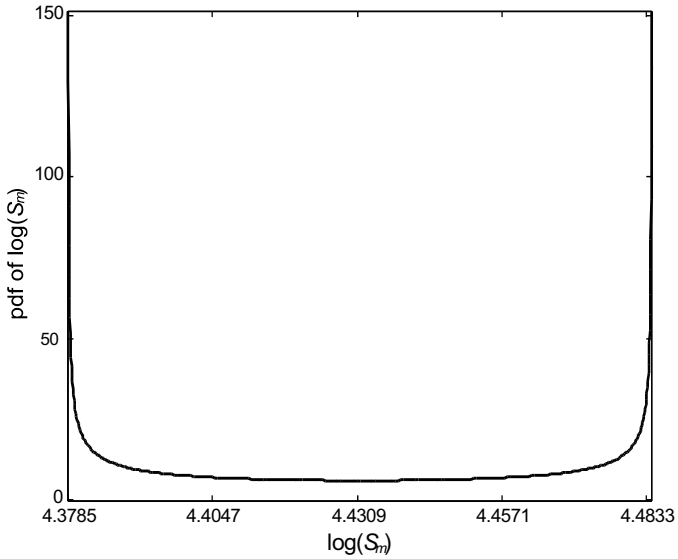


Figure 2. Probability density function of $\log(S_m) = \log(\overline{s_m^2}(\phi))$ assuming that ϕ is a random variable uniformly distributed between $-\pi$ and π. $\overline{x_m^2}/\overline{n_m^2}$ was made equal to 28, $\overline{n_m^2}$ to 1000 and c_m to 0.1. The p.d.f. curve of $\log(S_m)$ was 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.

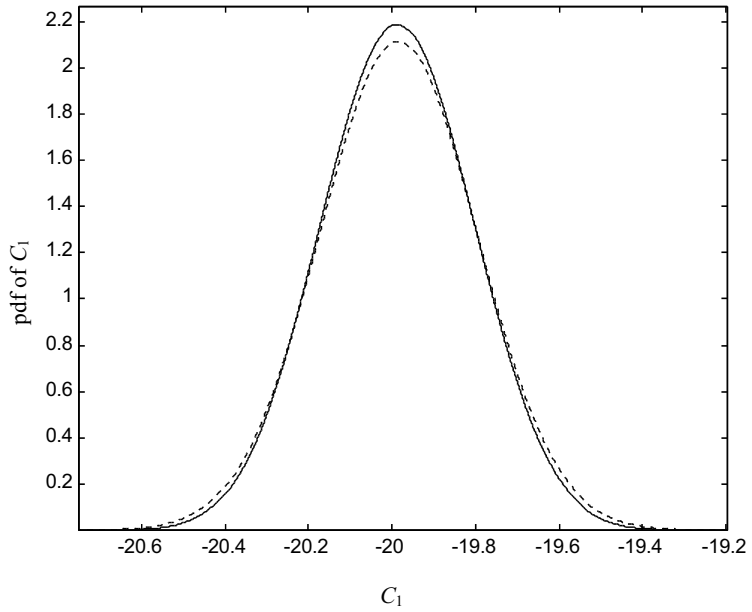


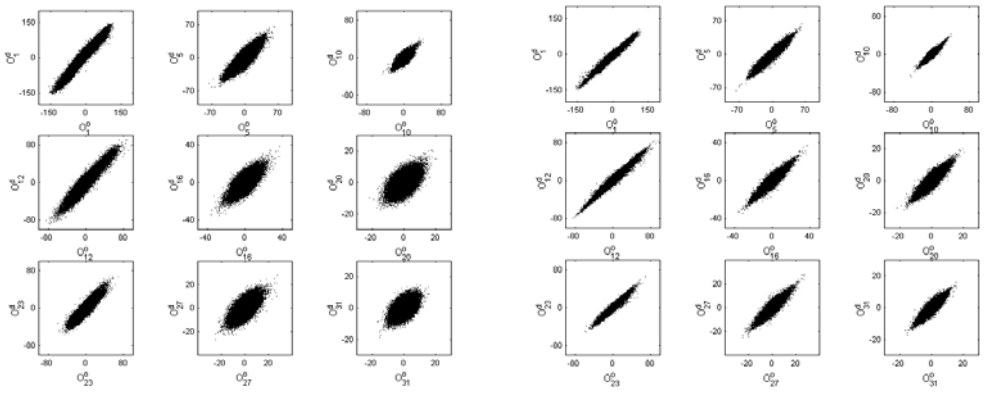
Figure 3. Probability density function of the static cepstral coefficient C_1 computed with 20 log energies $\log(\overline{s_m^2}(\phi))$. As a consequence, this density function corresponds to the convolution (———) of 20 p.d.f.´s similar to the one shown in Fig. 2. The theoretic Normal p.d.f. with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

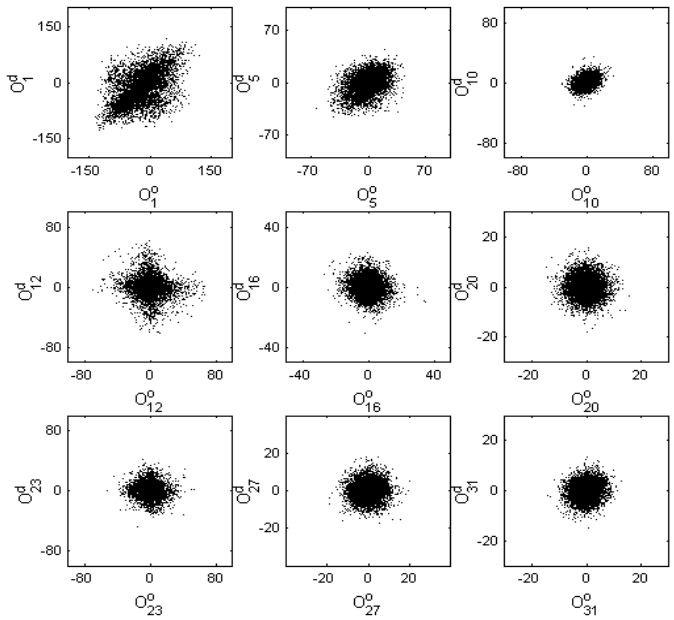


Figure 5. Cepstral coefficients from uncoded (O^o) vs. coded-decoded (O^d) speech signals. The coder is the 13 kbps GSM from the ETSI GSM-06.10 Full Rate Speech Transcoding. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral 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 $O^o_n = 0$.

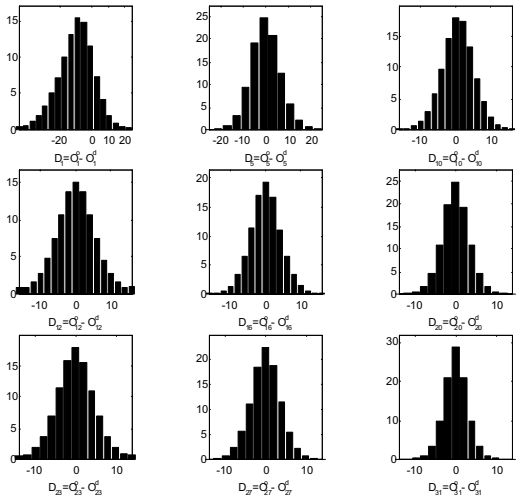


Figure 6. Distribution of coding distortion ($O^o - O^d$) with signals processed by 8 kbps CS-CELP from the ITU-T standard G.729. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated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

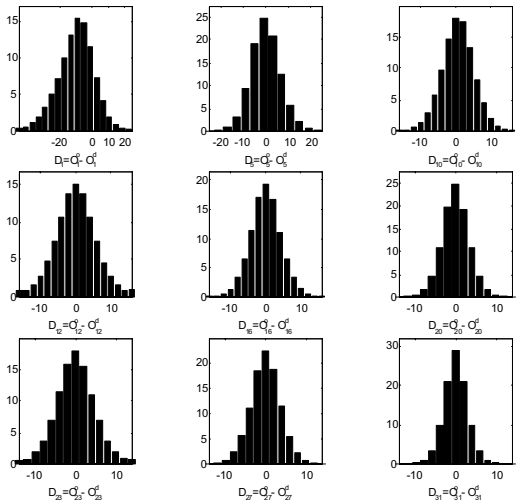


Figure 7. Distribution of coding distortion ($O^o - O^d$) with signals processed by 5.3 kbps G723-1 from the ITU-T standard G.723.1. The parameters employed in the figures correspond to static (1, 5, 10), delta (12, 16, 20) and delta-delta (23, 27, 31) cepstral coefficients. The histograms were generated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

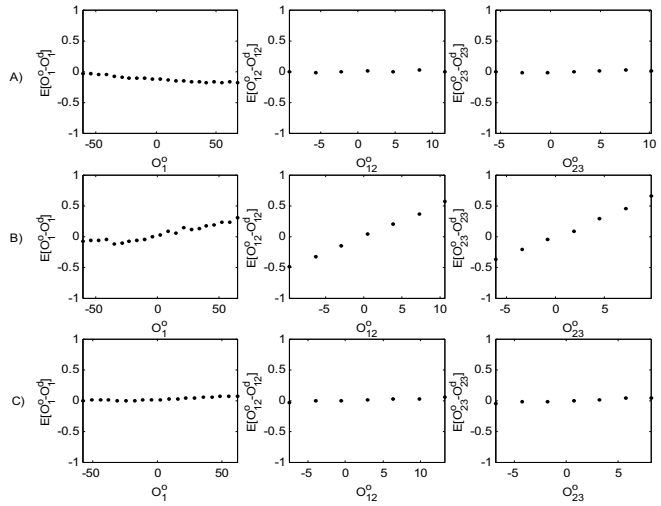


Figure 8. Expected value of the coding-decoding error, $E[O_n^o - O_n^d] = m_n^d$, vs. O^o. The expected value is normalized with respect to the range of observed O^o. The following coders are analyzed: A) 8 kbps CS-CELP; B) 13 kbps GSM; and, C) 32 kbps ADPCM . The cepstral coefficients correspond to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

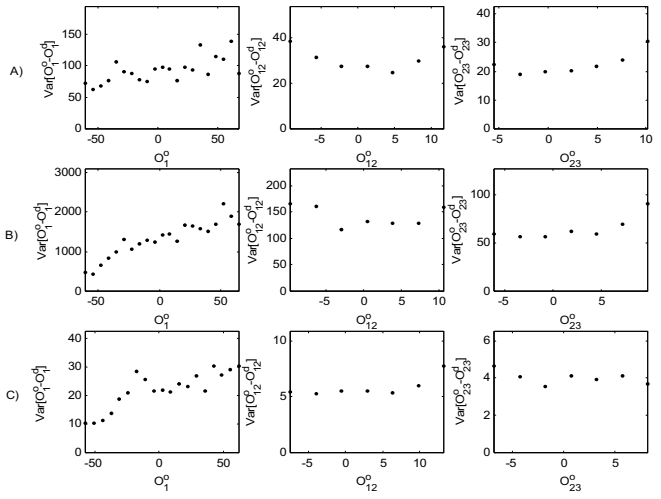


Figure 9. Variance of the coding-decoding error, $Var[O_n^o - O_n^d] = v_n^d$, vs. O^o. The following coders are analyzed: A) 8 kbps CS-CELP; B) 13 kbps GSM; and, C) 32 kbps ADPCM . The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23).

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

estimate the weighting coefficient from the uncertainty variance was empiric, although coherent; second, the variance in (5) was estimated using numeric approximations which resulted in a high computational load. It is worth highlighting that the same weighting $[b_s(O_t)]^{w(t)}$ has been used later by other authors. For instance, in (Bernard & Alwan, 2002; Tan, Dalsgaard, & Lindberg, 2005) this weighting function was used to address the problem of speech recognition in packet based and wireless communication. Notice that a lost packet would corresponds to reliability in signal estimation equal to zero.

5.2 The expected value of the output probability

In most HMM systems the output probability is modeled with a mixture of Gaussians with diagonal covariance matrices (Huang et al., 1990):

$$b_s(O_t) = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} (2\cdot\pi)^{-0.5} \cdot \left(Var_{s,g,n}\right)^{-0.5} \cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}-E_{s,g,n}\right)^2}{Var_{s,g,n}}} \tag{33}$$

where *s, g, n* are the indices for the states, the Gaussian components and the coefficients, respectively; p_g is a weighting parameter; $O_t = \left[O_{t,1},\ O_{t,2},\ ...,\ O_{t,N}\right]$ is the observation vector composed of *N* coefficients (static, delta and delta-delta cepstral parameters); and, $E_{s,g,n}$ and $Var_{s,g,n}$ are the HMM mean and variance, respectively. Assuming that the coefficients $O_{t,n}$ are uncorrelated, which in turn results in the diagonal covariance matrices, the expected value of $b_s(O_t)$ is given by:

$$E\left[b_s(O_t)\right] = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} E\left[\frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}} \cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}-E_{s,g,n}\right)^2}{Var_{s,g,n}}}\right] \tag{34}$$

where

$$E\left[\frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}} \cdot e^{\frac{1}{2}\cdot\frac{\left(O_{t,n}-E_{s,g,n}\right)^2}{Var_{s,g,n}}}\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}} \cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}-E_{s,g,n}\right)^2}{Var_{s,g,n}}} \cdot G\left(O_{t,n}; E(O_{t,n}); Var(O_{t,n})\right) \cdot dO_{t,n} \tag{35}$$

and, according to sections 2 and 3, $G\left(O_{t,n}; E(O_{t,n}); Var(O_{t,n})\right)$ is the Gaussian distribution of $O_{t,n}$. When the speech signal is corrupted with addittive noise, the mean, $E(O_{t,n})$, and variance, $Var(O_{t,n})$, are estimated with (11) (12) and (14) (15) for the static and delta cepstral coefficients, respectively. The delta-delta cepstral parameters can be computed using the same strategy employed in (14) (15). When the speech signal is affected by coding-decoding distortion, $E(O_t) = M^d$ and $Var(O_t) = V^d$, as discussed in section 3. As a consequence, it is possible to show that:

$$E\left[\frac{1}{\sqrt{2\cdot\pi\cdot Var_{s,g,n}}}\cdot e^{-\frac{1}{2}\cdot\frac{\left(O_{t,n}-E_{s,g,n}\right)^2}{Var_{s,g,n}}}\right]=\frac{1}{\sqrt{2\cdot\pi\cdot Vtot_{s,g,n,t}}}\cdot e^{-\frac{1}{2}\cdot\frac{\left(E\left[O_{t,n}\right]-E_{s,g,n}\right)^2}{Vtot_{s,g,n,t}}} \tag{36}$$

where $Vtot_{s,g,n,t}=Var_{s,g,n}+Var(O_{t,n})$. Therefore, (34) can be written as:

$$E[b_s(O_t)]=\sum_{g=1}^{G}p_g\cdot\prod_{n=1}^{N}\frac{1}{\sqrt{2\cdot\pi\cdot Vtot_{s,g,n,t}}}\cdot e^{-\frac{1}{2}\cdot\frac{\left(E\left[O_{t,n}\right]-E_{s,g,n}\right)^2}{Vtot_{s,g,n,t}}} \tag{37}$$

This is an elegant and generic result, and deserves some comments. Firstly, the expression (37) means that the expected value of the output probability is also represented by a sum of Gaussian functions. Secondly, if $Var(O_{t,n})\to 0$ (i.e. high SNR) $O_{t,n}$ can be considered as a constant and (37) is reduced to the ordinary output probability because $E\left[O_{t,n}\right]=O_{t,n}$. Finally, if $Var(O_{t,n})$ is high (i.e. low SNR) the expected value given by (37) tends to zero independently of $E\left[O_{t,n}\right]$, and of the HMM parameters $E_{s,g,n}$ and $Var_{s,g,n}$, which means that the information provided by a noisy observation vector is not useful and has a low weight in the final decision procedure of accepting or rejecting a speaker. The weighting mechanism could be defined by the fact that the original output probability is mapped to the same value (1 in the empirical weighting function, and 0 in (37)) when the segmental SNR is very low. As a consequence, the expression (37) is consistent with the weighting function mentioned in section 6.1 and can define a stochastic version of the weighted Viterbi algorithm, which in turn was proposed to take into consideration the segmental SNR.

5.3 SWV applied to speaker verification with additive noise

As shown in (Yoma & Villar; 2002), experiments with speech signal corrupted by car noise show that the expected value of the output probability using the additive noise model combined with SS led to reductions of 10%, 34%, 35% and 31% in the EER_{SD} at SNR=18dB, 12dB, 6dB and 0dB, respectively, when compared with the ordinary Viterbi algorithm also with SS. In the same conditions, the reductions in the EER_{SI} were 26%, 41%, 43% and 30% at, respectively, SNR=18dB, 12dB, 6dB and 0dB as shown in Table 1. Although an optimum might be considered around c_m=*0.25*, according to Figs. 10 and 11 the EER_{SD} and the EER_{SI} did not present a high variation with c_m, which confirms the stability of the approach proposed. Preliminary experiments showed that the lower the reduction due to spectral subtraction, the higher the improvement due to the weighted Viterbi algorithm. The effectiveness of spectral subtraction is closely related to how low SNR frames are processed.
According to the experiments presented in (Yoma & Villar; 2002) and Table 1 the weighted Viterbi algorithm defined by the expected observation probability in (37) can improve the accuracy of the speaker verification system even if SS is not employed. For instance, the average reduction in EER_{SD} and EER_{SI} without SS is 11%. As can be seen in (Yoma & Villar; 2002) shown in Table 2 and in Fig. 12, the expected value of the output probability using the additive noise model substantially reduced the variability of $TEER_{SD}$ and $TEER_{SI}$ with and without SS. According to Table 2, the differences $TEER_{SD}$(18dB) - $TEER_{SD}$ (0dB) and $TEER_{SI}$(18dB) - $TEER_{SI}$(0dB) with SS are, respectively, 53% and 55% lower with the weighted

Viterbi algorithm than with the ordinary one. This must be due to the fact that, when the segmental SNR decreases, $Var(O_{t,n})$ increases and the output probability according to (37) tends to 0 for both the client and global HMM in the normalized log likelihood (log $L(O)$) (Furui, 1997)::

$$\log L(\mathrm{O}) = \log P(\mathrm{O} \mid \lambda_i) - \log P(\mathrm{O} \mid \lambda_g) \tag{38}$$

where $P(\mathrm{O} \mid \lambda_i)$ is the likelihood related to the speaker *i;* and $P(\mathrm{O} \mid \lambda_g)$ is the likelihood related to the global HMMs.

The results presented in (Yoma & Villar; 2002) with speech noise basically confirmed the tests with car noise. The expected observation probability in (37) led to average reductions in EER_{SD} and in EER_{SI} equal to 23% and 30%, respectively, with SS. Significance analysis with the McNamar´s testing (Gillik & Cox, 1989) shows that this improvement due to the expected value of the output probability using the additive noise model combined with SS, when compared with the ordinary Viterbi algorithm also with SS, are significant ($p<0.1$ at SNR=18dB and $p<0.001$ at SNR=12, 6 and 0dB). Also, the differences $TEER_{SD}$(18dB) - $TEER_{SD}$ (0dB) and $TEER_{SI}$(18dB) - $TEER_{SI}$(0dB) were dramatically improved by the weighted Viterbi algorithm in combination with the additive noise model.

Finally, it is worth mentioning that the performance of SS is highly dependent on the parameters related to the thresholds (Berouti et al., 1979; Vaseghi & Milner, 1997) that are defined to make the technique work properly. In the case of the SS as defined in (9), parameter β, which defines the lower bound for the estimated signal energy, was not optimized for each SNR although its optimum values is case dependent. For instance, Table 3 shows that the expected observation probability led to a reduction of 26% in the EER_{SI} at SNR=18dB although SS alone did not give any improvement. This result suggests that the weighted Viterbi algorithm also improves the robustness of SS by means of giving a lower weight to those frames with low segmental SNR, where in turn SS is not reliable.

SNR	18dB	12dB	6dB	0dB
Vit-Nss	2.50	5.11	14.70	32.94
Vit-SS	2.59	4.71	11.14	26.73
SWVit-NSS	2.26	4.40	12.22	31.35
SWVit-SS	1.92	2.79	6.40	18.85

Table 1. EER_{SI} (speaker-independent Equal Error Rate) % with speech corrupted by additive noise (car noise). The correction coefficient c_m was made equal to 0.25.

	Vit-NSS	Vit-SS	SWVit-NSS	SWVit-SS
$TEER_{SD}$	1.72	1.93	1.01	0.90
$TEER_{SI}$	1.62	1.88	0.93	0.84

Table 2. Difference in the threshold of equal error rate at 18dB and 0dB, TEER(18dB)-TEER(0dB), with speech corrupted by additive noise (car noise). The correction coefficient c_m was made equal to 0.25.

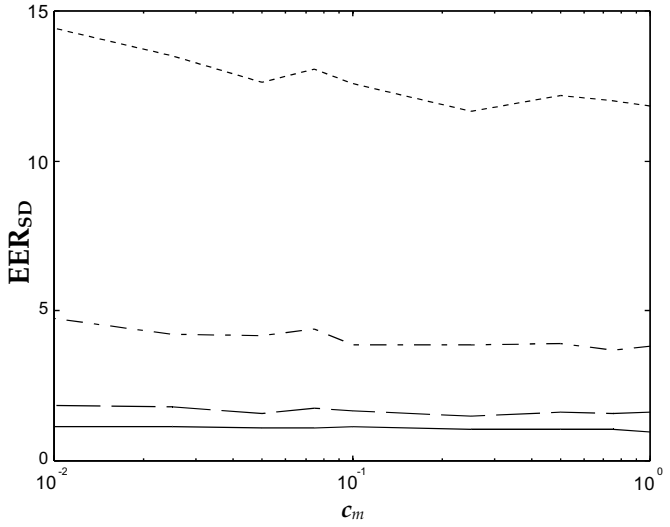


Figure 10. EER_{SD} vs. c_m with speech corrupted by additive noise (car noise): 18dB (——), 12dB (— —), 6dB (— · —) and 0dB (- - -).

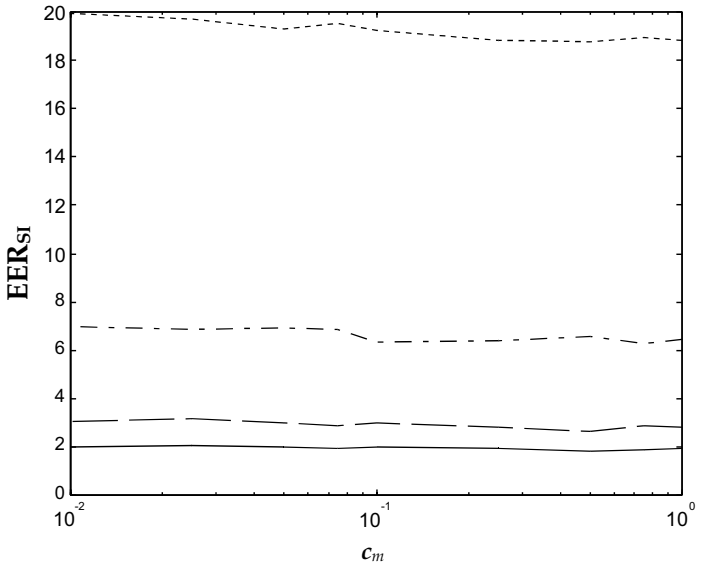


Figure 11. EER_{SI} vs. c_m with speech corrupted by additive noise (car noise): 18dB (——), 12dB (— —), 6dB (— · —) and 0dB (- - -).

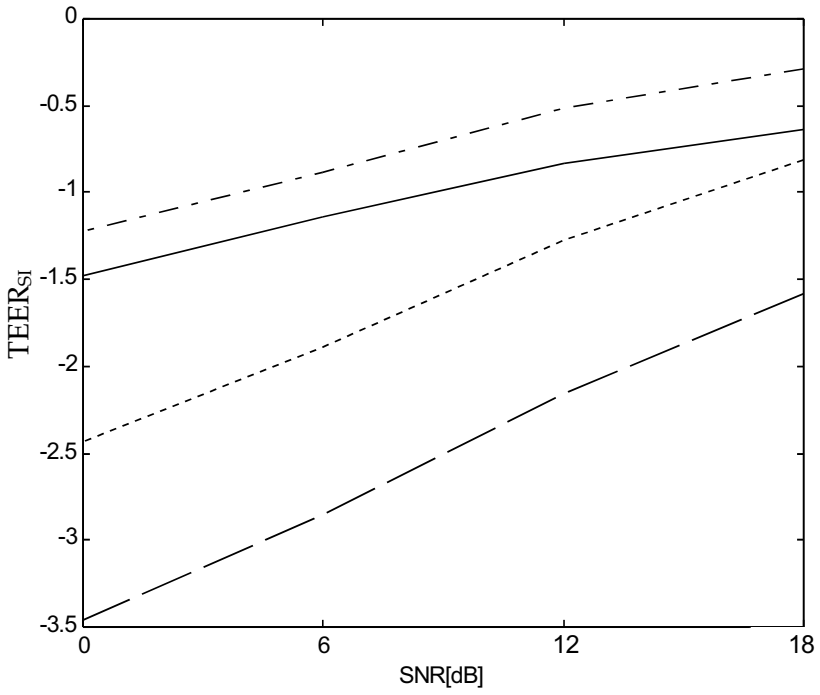


Figure 12. Speaker-independent threshold of equal error rate ($TEER_{SI}$) vs. SNR with speech corrupted by additive noise (car noise): *WVit-SS* (——); *Vit-SS* (— —); *WVit-NSS* (— · —) and *Vit-NSS* (- - -). The correction coefficient c_m was made equal to 0.25.

5.4 SWV applied to low-bit rate coding-decoding distortion compensation

The code-book to model the non-distorted speech process was composed of 256 code-words and was generated with the uncoded training utterances. The techniques 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 explained in section 4. Observe that *Baseline* indicates that no HMM compensation was applied. The baseline system with non-distorted speech and without any compensation gave a WER equal to 5.9%.

According to the results presented in (Yoma et. al., 2006) and shown in Table 3, 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 3, 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.

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 3. WER (%) with signal processed with the following coders: 32 kbps ADPCM, 13 kbps GSM, 8kbps 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.

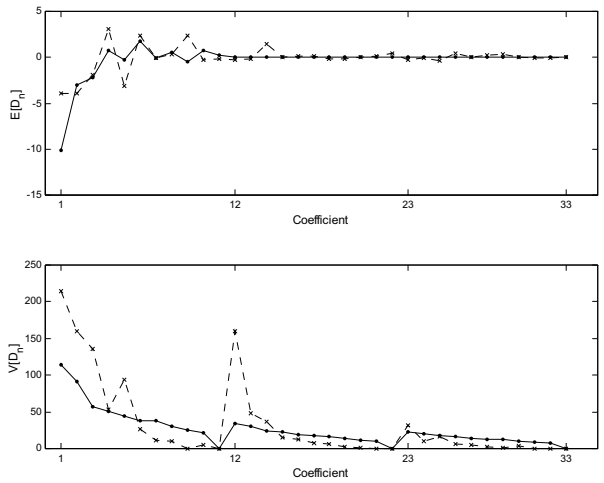


Figure 13. M^d (top) and V^d (bottom) estimated with the EM based algorithm (--×--) and computed with the training database by directly aligning uncoded and coded-decoded speech samples (——). The signals were processed by the 8 kbps CS-CELP from the ITU-T standard G.729.

According to Fig. 13, 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. 13) could be due to fact that the coding-decoding distortion depends on the speaker. As can be seen in Table 3, 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 approaches described in sections 3 and 4 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, the results presented in (Yoma et. al., 2006), show 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 (Zhao, 1994), where additive bias compensation in the cepstral domain for speaker adaptation was studied. Also according to Table 3, 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 was composed of utterances from just 36 speakers. Consequently, the fact that the EM compensation method also introduces a speaker adaptation effect would be consistent with the size of the database. Most of 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 was shown in (Yoma et. al., 2006), 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 *HMM-Comp-EM* technique; and second, the improvement in word accuracy given by the method presented in (Yoma et. al., 2006) is not due to under trained conditions.

The EM adaptation method is unsupervised and requires only one adaptation utterance. In (Yoma et. al., 2006), RATZ (Moreno et. al., 1995), without variance compensation and supervised ML estimation (Afify et. al., 1998), based on forced Viterbi alignment was compared with *HMM-Comp-EM* algorithm. According to (Moreno et. al., 1995), 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. Notice that blind RATZ is an unsupervised method. Word accuracy given by RATZ strongly depends on the number of adapting utterances employed to compute $O^{o}_{t,n}$. 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 gave a WER even higher than the one achieved with the baseline system. It is worth highlighting that *HMM-Comp-EM* provides higher recognition accuracy even when the whole testing data was employed by RATZ. Supervised ML, *Superv-ML,* estimation evaluated in (Yoma et al., 2006) is similar to the one presented in (Afify et. al., 1998) except for the fact that the Forward-Backward procedure was replaced with the Viterbi algorithm. The improvement in WER given by *Superv-ML* also depends on the number of adapting utterances. The stochastic model employed by the proposed EM unsupervised algorithm is more robust than the one provided by the *Superv-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*. When the number of adapting utterances is equal to 500, *Superv-ML* could give improvements in WER 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.

5.5. SWV to address the problem of joint compensation of additive noise and low-bit rate coding-decoding distortion

As can be seen in Fig. 14 (Yoma et. al., 2003), the problem of additive noise and low-bit rate coding-decoding distortion corresponds to a clean signal *s(t)* firstly corrupted by an additive noise in the temporal domain, *x(t)*, and then coded and decoded, $x^D(t)$. The observation parameter vectors of the signals *s(t)*, *x(t)* and $x^D(t)$. are $O_t^{S,U}$, $O_t^{X,U}$ and $O_t^{S,D}$, respectively. *S* and *X* denote the clean and noisy signal, respectively; *U* and *D* correspond to the signals before (uncoded) and after (distorted) the coding-decoding process.

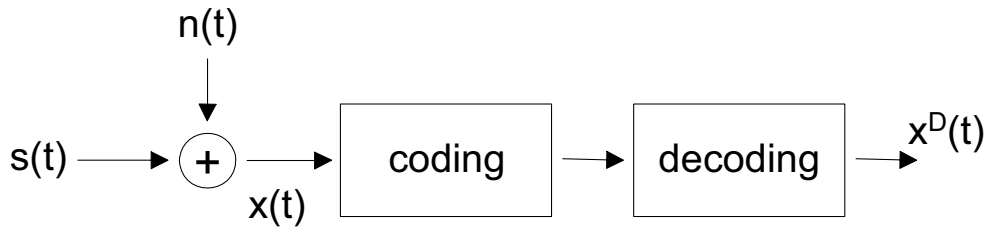


Figure 14. Additive noise and coding distortion.

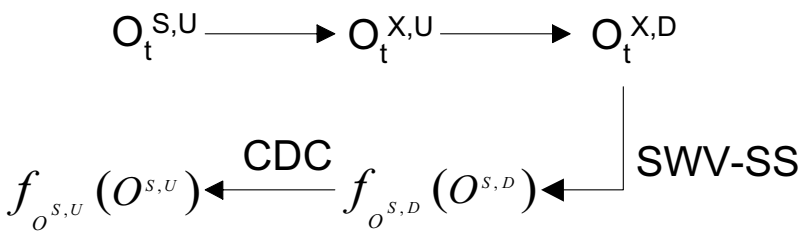


Figure 15. Join compensation of additive noise and coding distortion: $f_{O^{S,D}}(O^{S,D})$ denotes the p.d.f. of the distorted by coding clean signal; $f_{O^{S,U}}(O^{S,U})$ corresponds to the p.d.f. of the uncoded clean signal.

As is shown in Fig. 15, the method proposed in (Yoma et. al., 2003) firstly compensates the presence of additive noise by applying SS and estimating the uncertainty variance in noise canceling as in section 2 using $x^D(t)$. As a result, the p.d.f. of the distorted by coding clean speech, $f_{O^{S,D}}(O^{S,D})$, is generated. Then, as discussed in section 3, $f_{O^{S,U}}(O^{S,U})$ is estimated by adding M^d and V^d to the mean and variance, respectively, of $f_{O^{S,D}}(O^{S,D})$. Finally, by taking the expected value of the output p.d.f., the compensation of the additive noise and of the coding distortion are incorporated in the Viterbi decoding as in (37).

As can be seen in Tables 4 and 5 (Yoma et al., 2003), the additive noise and the coder dramatically degraded the WAC at SNR equal to 18dB and 12dB. SWV and SS substantially reduced the WER, but the highest improvement was achieved when coding-decoding compensation was also applied. Reductions as high as 50% or 60% in WER were observed at 18dB and 12dB. Nevertheless, the degradation of the system at 12dB is still too severe. According to Tables 4 and 5, the additive noise has probably a more significant effect on rising the WER than the coding-decoding distortion. As a result, improving the accuracy of the additive noise model (Yoma et. al., 1998-B) at low SNR should certainly increase the effectiveness of the approach proposed here.

SNR	18dB	12dB
Baseline	27.4	38.5
SWV-SS	11.9	18.3
SWV-SS-CDC	10.2	16.9

Table 4. WER (%) with signal corrupted with additive noise (car noise) and coded by 8kbps CS-CELP.

SNR	18dB	12dB
Baseline	26.2	37.9
SWV-SS	11.7	17.5
SWV-SS-CDC	10.0	15.3

Table 5. WER (%) with signal corrupted with additive noise (speech noise) and coded by 8kbps CS-CELP.

6. Language model accuracy and uncertainty in noise canceling in SWV

No significant improvements were observed when the SWV algorithm in combination with the additive noise model proposed in (Yoma et al, 1998-B) and SS was applied to the connected digit task. This result must be due to fact that the SWV algorithm makes the HMM observation p.d.f. lose discrimination ability at noisy frames. This hypothesis means that the Viterbi decoding should be guided by the information from higher layers, such as language modeling, in those intervals with low SNR. In contrast, the connected digit task employs a flat language model. In (Yoma et al, 2003-B), the SWV algorithm was applied to a continuous speech, medium vocabulary, speaker independent (SI) task opening a new paradigm in speech recognition where the noise canceling could interact with the information from higher layers in the same way the human perceptions works. Bigram and trigram language models were tested and, in combination with spectral subtraction, the SWV algorithm coul lead to reductions as high as 20% or 45% in word error rate (WER) using a rough estimation of the additive noise made in a short non-speech interval. Also, the results presented in (Yoma et al, 2003-B) suggest that the higher the language model accuracy, the higher the improvement due to SWV. Consequently, the problem of noise robustness in speech recognition should be classified in two different contexts: firstly, at the acoustic-phonetic level only, as in small vocabulary tasks with flat language model; and, by integrating noise canceling with the information from higher layers.

7. Conclusions

The Stochastic Weighted Viterbi algorithm offers a unified framework to reduce the effect of additive/convolutional noise and low-bit rate coding-decoding distortion. SWV started a new paradigm in speech processing by considering the original speech signal information as a stochastic variable. Consequently, the ordinary HMM observation probability needs to be replaced with its expected value. SWV is interesting from the theoretic and applied points of view: first, it is based on stochastic models of additive noise and low-bit rate coding-decoding distortion; and second, it assumes reasonable hypotheses such as a rough estimation of additive noise and a low number of adaptation utterances. It is worth emphasizing that SWV allows the interaction between the higher layers of language modeling (semantic, syntactic, etc…) and acoustic models in ASR just like in human perception: the higher layer of the linguistic information should have a higher weight in

those frames with low SNR or low reliability. Finally, the concepts of uncertainty in noise canceling and weighted recognition algorithms, which were firstly proposed by the first author of this chapter, have also widely been employed elsewhere in the fields of ASR and SV in later publications.

8. Acknowledgement

This research described here was funded by Conicyt - Chile under grants Fondecyt Nº 1030956, Fondecyt Nº 1000934, and Fondef Nº D02I-1089.

9. References

Acero, A. & Stern, R. (1990). Environmental robustness in automatic speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´90*,pp. 849-852.

Acero, A., Deng, L. & Droppo, J. G. (2006-A). Method of iterative noise estimation in a recursive framework. *United States Patent 7139703*.

Acero, A.; Deng, L. & Droppo, J. G. (2006-B). Non-linear observation model for removing noise from corrupted signals. *United States Patent 7047047*.

Afify, M.; Gong, Y. & Haton, J. (1998). A General Joint Additive and Convolutive Bias Compensation Approach Applied to Noise Lombard Speech Recognition *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 6, pp. 524-538.

Arrowood, J.A. & Clements, M.A. (2004). Extended cluster information vector quantization (ECIVQ) for robust classification, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2004*, pp. I889-I892.

Barger, P. and Sridharan, S. (1997). Robust speaker identification using multi-microphone systems. *Proceedings of TENCON '97*, pp. 261 –264.

Bernard, A. & Alwan, A. (2002). Low-bitrate distributed speech recognition for packet-based and wireless communication. *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 8, pp. 570-579.

Berouti, M.; Schwartz, R. & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'79*. pp.208-211.

Breton, P. A. (2005). Method and device for voice recognition in environments with fluctuating noise levels. *United States Patent 6859773*.

Campbell, J. P.; Tremain, T. E. & Welch, V. C. (1991). The federal standard 1016 4800 bps CELP voice coder. *Digital Signal Processing*, Vol. 1, No. 3, pp. 145--155,.

Chan, S.M. & Siu, M.H. (2004). Discrimination power weighted subword-based speaker verification. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2004*, pp. I45-I4.

Cho, H.Y.; Kim, L.Y. & Oh, Y.H. (2002). Segmental reliability weighting for robust recognition of partly corrupted speech. *IEE Electronics Letters*, Vol. 38, No. 12, pp. 611-612.

Delaney, B. (2005). Increased robustness against bit errors for distributed speech recognition in wireless environments. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2005,* pp. I313-I316.

Deng, L.; Droppo, J. & Acero, A. (2005). Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, pp. 412-421.

Drygajlo, A. & El-Maliki, M. (1998). Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´98*, pp. 121 –124.

Erzin, E.; Yemez, Y. & Tekalp, A.M. (2005). Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia*, Vol. 7, No. 5, pp. 840-852.

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

Furui, S. (1982). Cepstral analysis technique for automatic speaker verication. *IEEE Transactions on Speech and Audio Processing*, Vol. 29, No.2, pp.254-272.

Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters*, Vol. 18, pp. 859-872.

Gales M.J.F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, Vol. 12, no. 2, pp. 75-98.

Gales, M.J.F. & Young, S.J. (1993). HMM recognition in noise using parallel model combination. *Proceedings of. Eurospeech'93*, pp. 837-840.

Gales, M.J.F. (1997). "Nice" model-based compensation schemes for robust speech recognition. *Proceedings of. Esca-Nato Workshop on robust speech recognition for unknown channels.* pp. 55-64.

Gauvain J. & Lee, C-H (1994). Maximum a posteriori estimation for multivariate Gaussian Mixture Observation of Markov Chains, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298.

Gillik, L. & Cox, S.J. (1989). Some statistical issues in the comparison of speech recognition algorithms. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´98*, pp.532-535.

Gomez, A. M.; Peinado, A. M. & Sanchez, V. (2006). Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels. *IEEE Transactions on Multimedia*, Vol. 8, No. 6, pp. 1228-1238.

Hardt, D. & Fellbaum, K. (1997). Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´97.*, pp. 867 –870.

Hermansky, H. et al. (1991). Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). *Procedings of Eurospeech' 91*, pp.1367-1370.

Huang, X.D.; Ariki, Y. & Jack, M. (1990). Hidden Markov Models for speech recognition. Edinburgh University Press.

Huerta , J.M. (2000). Speech Recognition in Mobile Environments. *Ph.D thesi*s, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

Hung, J.; Shen, J. & Lee, L. (1998). Improved robustness for speech recognition under noisy conditions using correlated Parallel Model Combination. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'98*, Vol. 1, pp. 549-552 .

ITU-T (1990). Recommendation G.726, 40-,32-,24-, and 16-Kb/s adaptive differential pulse code modulation.

ITU-T (1996). Recommendation G.729-Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-CELP).

ITU-T (1996-B). Recommendation G.723.1 Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbps, Marzo 1996.

Keung, C. L.; Au, O. C.; Yim, C. H. & Fung, C. C. (2000). Probabilistic Compensation of Unreliable Feature Components for Robust Speech Recognition, *Proceedings of International Conference of Spoken Language Processing ICSLP 2000*, pp. 1085-1087.

Kitaoka, N. & Nakagawa, S. (2002). Evaluation of spectral subtraction with smoothing of time direction on the aurora 2 task. *Proceedings of International Conference on Spoken Language Processing ICSLP 2002*, pp. 1085-1087.

LDC (1995). Latino database provided by Linguistic Data Consortium (LDC), University of Pennsylvania: http://www.ldc.upenn.edu/Catalog/LDC95S28.html.

Li, S. C. (2003). Applying eigenvoice model adaptation for user verbal information verification. *M.Sc. Thesis in Computer Science and Information Engineering*, National Chen Kung University, Tainan, Taiwan, R.O.C.

Liao, H. & Gales, M.J.F. (2005). Joint Uncertainty Decoding for Noise Robust Speech Recognition. *Proceedings of Interspeech 2005*, pp. 3129-3132.

Moon, T.K. (1996). The Expectation-Maximization Algorithm. *IEEE Signal Processing Magazine*, Vol. 13, No.6, pp.47-60.

Moreno P. J., Raj B., Govea E. and Stern R. M. (1995). Multivariate Gaussian Based Cepstral Normalization for Robust Speech Recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'95*, pp. 137-140.

Ortega-Garcia, J. & Gonzalez-Rodriguez, J. (1997). Providing single and multi-channel acoustical robustness to speaker identification systems. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´97*, pp. 1107-1110.

Papoulis, A. (1991). Probability, random variables, and stochastic processes. McGraw-Hill International Editions.

Pfitzinger, H.R. (2000). Removing Hum from Spoken Language Recources. *Proceedings of International Conference of Spoken Language Processing ICSLP 2000*, pp. 618-621.

Pitsikalis, V.; Katsamanis, A.; Papandreou, G. & Maragos. P. (2006). Adaptive Multimodal Fusion by Uncertainty Compensation. *Proceedings of International Conference of Spoken Language Processing ICSLP 2006,* pp. 2458-2461.

Raj, B.; Gouvea, E. B.; Moreno, P. J. & Stern, R. M. (1996). Cepstral compensation by polynomial approximation for environment-independent speech recognition. *Proceedings of International Conference of Spoken Language Processing ICSLP'96*, Vol 4, pp. 2340-2343.

Reynolds, D.A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No.4, pp. 639-643.

Rose, R.C. ; Hofstetter, E.M. & Reynolds D.A. (1994). Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No.2 , pp. 245-257.

Tan, Z. H.; Dalsgaard, P. & Lindberg, B. (2005). Automatic speech recognition over error-prone wireless networks, *Speech Communication*, Vol. 47, No. 1-2, pp. 220-242.

van Vuuren, S. (1996). Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. *Proceedings of International Conference of Spoken Language Processing ICSLP'*96, pp. 1788 –1791.

Vaseghi, S.V. & Milner, B.P. (1997). Noise compensation methods for Hidden Markov Model speech recognition in adverse environments. *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 1, pp. 11-21.

Vildjiounaite, E.; Makela, S.M.; Lindholm, M.; Riihimaki, R.; Kyllonen, V.; Mantyjarvi, J. & Ailisto, H. H. (2006). Unobtrusive multimodal biometrics for ensuring privacy and information security with personal devices. *Lecture Notes in Computer Science*, No. 3968, pp. 187-201.

Wu, C. H. & Chen, Y. J. (2001). Multi-keyword spotting of telephone speech using a fuzzy search algorithm and keyword-driven two-level CBSM, *Speech Communication*. Vol. 33, No. 3, pp. 197-212.

Yoma, N.B.; McInnes, F. & Jack, M. (1995). Improved Algorithms for Speech Recognition in Noise using Lateral Inhibition and SNR Weighting. *Proceedings of Eurospeech'95*, pp.461-464

Yoma, N.B.; McInnes, F. & Jack, M. (1996-A). Lateral inhibition net and weighted matching algorithm for speech recognition in noise. *IEE Proceedings of Vision, Image and Signal Processing*, Vol. 143, No. 5, pp. 324-330.

Yoma, N.B.; McInnes, F. & Jack, M. (1996-B). Use of a Reliability coefficient in noise canceling by Neural Net and Weighted Matching Algorithms. *Proceedings of International Conference of Spoken Language Processing ICSLP'96,* pp. 2297-2300.

Yoma, N.B.; McInnes, F. & Jack, J. (1997-A). Weighted Matching Algorithms and Reliability in Noise Canceling by Spectral Subtraction. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'97*, Vol.2, pp. 1171-1174.

Yoma, N.B.; McInnes, F. & Jack, J. (1997-B). Spectral Subtraction and Mean Normalization in the context of Weighted Matching Algorithms. *Proceedings of Eurospeech'97*, pp. 1411-1414.

Yoma, N.B.; McInnes, F. & Jack, J. (1998-A). Weighted Viterbi Algorithm and State Duration Modeling for Speech Recognition in Noise. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'98*, pp. 709-712.

Yoma, N.B.; McInnes, F. & Jack, J. (1998-B). Improving Performance of Spectral Subtraction in speech recognition using a model for additive noise. *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 6, pp. 579-582.

Yoma, N.B. (1998-C). Speech recognition in noise using weighted matching algorithms. *Ph.D. Thesis*, University of Edinburgh, UK.

Yoma, N. B.; Ling, L. L.& Dotto, S. (1999). Robust connected word speech recognition using weighted Viterbi algorithm and context-dependent temporal constraints. *Proceedings of Eurospeech'99*, pp. 2869-2872.

Yoma, N. B. & Villar, M. (2001). Additive and convolutional noise canceling in speaker verification using a stochastic weighted Viterbi algorithm". *Proceedings of Eurospeech 2001,* pp. 2845-2848.

Yoma, N. B. & Villar, M. (2002). Speaker Verification in noise using a stochastic version of the weighted Viterbi algorithm". *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No 3, pp. 158-166.

Yoma, N. B.; Silva, J.; Busso, C. & Brito, I. (2003-A). On compensating additive noise and CS-CELP distortion in speech recognition using the stochastic weighted Viterbi algorithm. *IEE Electronics Letters*, Vol 39, No. 4, pp. 409-411.

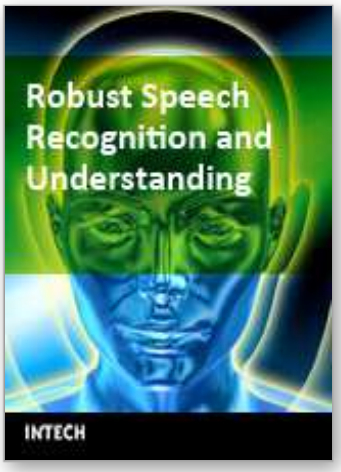
Yoma, N. B.; Brito, I. & Silva, J. (2003-B). Language Model Accuracy and Uncertainty in Noise Canceling in the Stochastic Weighted Viterbi Algorithm. *In Proceedings of Eurospeech 2003,* pp. 2193-2196.

Yoma, N. B.; Brito, I. & Molina, C. (2004). The stochastic weighted Viterbi algorithm: a framework to compensate additive noise and low bit rate coding distortion. *Proceedings of International Conference of Spoken Language Processing ICSLP 2004,* pp. 2821-2824.

Yoma, N. B.; Molina, C.; Silva, J. & Busso, C. (2005). Modelling, estimating and compensating low-bit rate coding distortion in speech recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 14, No.1, pp. 246-255.

Yoma, N. B.; Molina, C. (2006). Feature-dependent compensation of coders in speech recognition. *Signal Processing (Elsevier)*, Vol. 86, No. 1, pp. 38-49.

Zhao, Y. (1994). An acoustic-phonetic based speaker adaptation technique for improving speaker independent continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, pp. 380-394.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821