

THE STOCHASTIC WEIGHTED VITERBI ALGORITHM: A FRAME WORK TO COMPENSATE ADDITIVE NOISE AND LOW – BIT RATE CODING DISTORTION

Néstor Becerra Yoma, Iván Brito, Carlos Molina
Electrical Engineering Department, University of Chile
Av. Tupper 2007, P.O.Box 412-3, Santiago, CHILE.
nbecerra@ing.uchile.cl

ABSTRACT

A solution to the problem of speech recognition with signals corrupted by additive noise and distorted by low-bit rate coders is presented in this paper. The additive noise and the coding distortion are cancelled according to the following scheme: firstly, the pdf of the clean coded-decoded speech is estimated with an additive noise model; second, the pdf of the clean uncoded signal is also estimated with a coding distortion model; and finally, the HMM is compensated by using the expected value of the observation pdf in the context of the stochastic weighted Viterbi (SWV) algorithm. The approach leads to reductions as high as 50% or 60% in word error rate.

1. INTRODUCTION

Improving robustness to noise is one of the most important problems that needs to be solved in order to make speech recognition successful in real applications. Moreover, the evolution and popularity of mobile and TCP/IP networks have created the problem of improving the recognition accuracy for speech distorted by low-bit rate coders. The distortion of coding schemes in speech recognizers cannot be solved by applying conventional noise canceling techniques [3]. Some of the techniques that have been proposed to cancel or compensate additive or/and convolutional noise are: spectral subtraction, SS [5]; Rasta [2]; Parallel Model Combination, PMC [1]; and, Cepstral Mean Normalization.

A stochastic version of the weighted Viterbi (SWV, Stochastic Weighted Viterbi) algorithm was proposed and successfully applied to compensate additive noise in text-dependent speaker verification [6] and continuous speech recognition [7]. In those papers the observation parameters were considered as random variables with normal distributions. As a result, the HMM observation pdf was replaced with its expected value. Moreover, in [7] it was shown that noise canceling could interact with the information from higher layers emulating the human perception in some extend. The SWV algorithm was combined with bigram and trigram language models, which in turn led the Viterbi decoding in those intervals where the information provided by noisy frames is not reliable.

In [8], the cepstral coefficients from uncoded and coded-decoded speech signals were linearly aligned to estimate a model for the distortion introduced by coding schemes. As a result, this distortion was modeled with a Gaussian distribution whose mean and variance did not depend on the phonetic class. Then, a HMM compensation method was proposed by considering the original and unseen uncoded cepstral parameters as random variables and by estimating the expected value of the output probability density function as in [6] and [7]. In this paper, the additive noise and the coding distortion are cancelled according to the following scheme: 1) the pdf of the clean coded-decoded speech is estimated; then, 2) the pdf of the clean uncoded signal is also estimated; and, 3) the HMM is compensated by using the expected value of observation pdf. The approach leads to reductions as high as 50% or 60% in WER.

2. THE STOCHASTIC WEIGHTED VITERBI (SWV) ALGORITHM

In the ordinary HMM topology the output pdf of observing the frame O_t at state s , $b_s(O_t)$, is computed considering O_t as being a vector of constants. In this paper the observation vector is composed of static, delta and delta-delta cepstral coefficients, and according to [6] these parameters should be considered as being random variables with normal distributions. Therefore, to counteract this incompatibility, $b_s(O_t)$ is replaced with $E[b_s(O_t)]$ in the Viterbi algorithm, where $E[b_s(O_t)]$ denotes the expected value of the output pdf.

2.1 Uncertainty variance in the cepstral domain

According to [6], the expected value and the uncertainty in noise canceling variance of the static cepstral coefficient C_n are:

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

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

where SSE_m is the SS estimation of the clean signal energy and is defined according to [5]:

$$SSE_m = \max \left\{ \overline{x_m^2} - \alpha \cdot E \left[\overline{n_m^2} \right]; \beta \cdot \overline{x_m^2} \right\} \quad (3)$$

where β is a constant, $\alpha = \alpha_0 - \mu \cdot SNR$, $\mu = \frac{\alpha_0 - 1}{18}$

and $1.0 \leq \alpha \leq \alpha_0$. In (2), $Var \left[\log \left(\overline{s_m^2(\phi)} \right) \right]$ denotes

the uncertainty in noise canceling variance in the log-filter domain and is estimated as proposed in [6]. In this paper the delta cepstral coefficient in frame t , $\delta C_{t,n}$ is defined as:

$$\delta C_{t,n} = \frac{C_{t+1,n} - C_{t-1,n}}{2} \quad (4)$$

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 :

$$E[\delta C_{t,n}] = \frac{E[C_{t+1,n}] - E[C_{t-1,n}]}{2} \quad (5)$$

$$Var[\delta C_{t,n}] = \frac{Var[C_{t+1,n}] + Var[C_{t-1,n}]}{4} \quad (6)$$

The same approach can be applied to delta-delta cepstral coefficient in frame t , $\delta^2 C_{t,n}$, that is defined as:

$$\delta^2 C_{t,n} = \frac{\delta C_{t+1,n} - \delta C_{t-1,n}}{2} \quad (7)$$

and,

$$E[\delta^2 C_{t,n}] = \frac{E[\delta C_{t+1,n}] - E[\delta C_{t-1,n}]}{2} \quad (8)$$

$$Var[\delta^2 C_{t,n}] = \frac{Var[\delta C_{t+1,n}] + Var[\delta C_{t-1,n}]}{4} \quad (9)$$

2.2. Modified output pdf

Consider that the HMM output pdf, $b_s(O_t)$, is modeled with a mixture of Gaussians with diagonal covariance matrices. The expected value of $b_s(O_t)$, $E[b_s(O_t)]$, is given by:

$$E[b_s(O_t)] = \sum_{g=1}^G p_g \cdot \prod_{n=1}^N \frac{1}{\sqrt{2 \cdot \pi \cdot V_{tot,s,g,n,t}}} \cdot e^{-\frac{1}{2} \frac{(E[O_{t,n}] - E_{s,g,n})^2}{V_{tot,s,g,n,t}}} \quad (10)$$

where s, g, n are the indices for the states, the Gaussian components and the coefficients, respectively; p_g is a weighting parameter; $O_t = [O_{t,1}, O_{t,2}, \dots, O_{t,N}]$ is the

observation vector; $E_{s,g,n}$ and $Var_{s,g,n}$ are the HMM mean and variance, respectively; the mean, $E(O_{t,n})$, and variance, $Var(O_{t,n})$, are estimated with (1) (2), (5) (6), and (8) (9) for the static, delta and delta-delta cepstral coefficients, respectively; and

$$V_{tot,s,g,n,t} = Var_{s,g,n} + Var(O_{t,n}) \quad (11)$$

In order to adapt the decreasing rate of the output pdf discrimination ability to the task and to the language model in the Viterbi decoding, expression (11) was modified to [7]:

$$V_{tot,s,g,n,t} = Var_{s,g,n} + ALMI \cdot Var(O_{t,n}) \quad (12)$$

where $ALMI$ (Acoustic and Language Model Integration) is a constant that needs to be tuned empirically.

3. COMPENSATION OF THE LOW-BIT RATE CODING DISTORTION

In [8] the coding-decoding distortion, D_n , was modeled in the cepstral domain as an additive random variable with Gaussian distribution $f_{D_n}(D_n) = N(m_n^d, v_n^d)$ where n, m_n^d and v_n^d denote the cepstral coefficient, the mean and the variance, respectively. Moreover, D_n could be considered as independent of the phonetic identity (Fig 1-2). Therefore, the cepstral coefficient n in frame t of the original signal, $O_{t,n}^o$, is given by:

$$O_{t,n}^o = O_{t,n}^d + D_n \quad (13)$$

where $O_{t,n}^d$ is the cepstral coefficient corresponding to the decoded-coded speech signal. In a real application $O_{t,n}^d$ is the observed cepstral parameter. From (13), the expected value of $O_{t,n}^o$ is given by:

$$E[O_{t,n}^o] = O_{t,n}^d + m_n^d \quad (14)$$

Concluding, the distortion caused by the coding-decoding scheme is represented by the mean vector $M^d = [m_1^d, m_2^d, m_3^d, \dots, m_n^d, \dots, m_N^d]$ and the variance vector $V^d = [v_1^d, v_2^d, v_3^d, \dots, v_n^d, \dots, v_N^d]$. Moreover, this distortion would not depend on the phonetic class, which is consistent with the analysis presented in [3].

The HMM compensation method proposed in [8] considers the original and unseen uncoded cepstral parameter $O_{t,n}^o$ as a random variable. So the output probability $b_s(O_t)$ should be replaced with its expected value $E[b_s(O_t)]$ as in [6]:

$$E[b_{s(O_t)}] = \sum_{g=1}^G p_g \prod_{n=1}^N \frac{1}{\sqrt{2\pi V_{tot_{s,g,n}}^d}} e^{-\frac{1}{2} \frac{(E[O_{t,n}^o] - E_{s,g,n})^2}{V_{tot_{s,g,n}}^d}} \quad (15)$$

where $E[O_{t,n}^o]$ is given by (14) and

$$V_{tot_{s,g,n}}^d = Var_{s,g,n} + v_n^d \quad (16)$$

where s, g, n are the indices for the states, the Gaussian components and the coefficients, respectively; p_g is a weighting parameter. The compensation according to (14) and (16) with M^d and V^d almost completely cancelled the coding-decoding distortion caused by CS-CELP and ADPCM vocoders in [6].

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

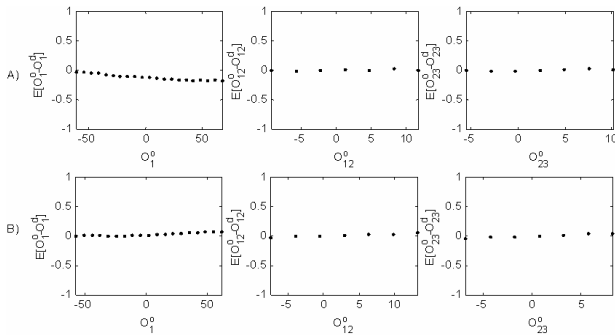
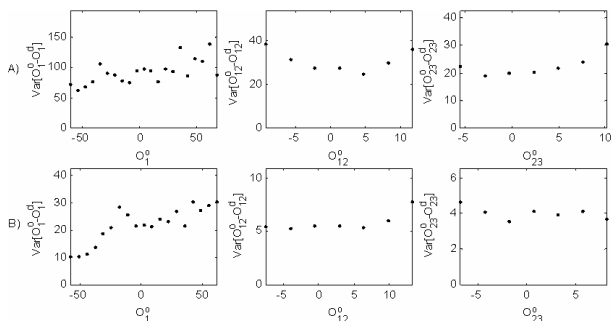


Figure 2: Variance of the coding-decoding error $Var[D_n]$ vs. O^o . The following coders are analyzed: A) 8kbps CS-CELP; and, B) 32kbps ADPCM. The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23).



4. JOINT COMPENSATION OF ADDITIVE NOISE AND CODING DISTORTION

As can be seen in Fig.3, the problem presented here 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 $O_t^{X,D}$, 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. As is shown in Fig. 4, the method proposed in this paper firstly compensates the presence of additive noise by applying SS and estimating the uncertainty variance in noise canceling as in section 2.1 using $x^D(t)$. As a result, the fdp of the distorted by coding clean speech, $f_{O_{s,D}}(O^{S,D})$, is generated. Then $f_{O_{s,U}}(O^{S,U})$ 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 fdp, the compensation of the additive noise and of the coding distortion are incorporated in the Viterbi decoding as in (10) and (15).

Figure 3: Additive noise and coding distortion.

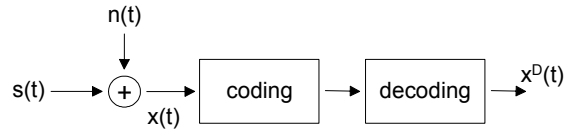


Figure 4: Joint compensation of additive noise and coding distortion: $f_{O_{s,D}}(O^{S,D})$ denotes the fdp of the distorted by coding clean signal; $f_{O_{s,U}}(O^{S,U})$ corresponds to the fdp of the uncoded clean signal.

$$O_t^{S,U} \longrightarrow O_t^{X,U} \longrightarrow O_t^{X,D} \\ f_{O_{s,U}}(O^{S,U}) \xleftarrow{\text{CDC}} f_{O_{s,D}}(O^{S,D}) \xleftarrow{\text{SWV-SS}}$$

5. EXPERIMENTS

The compensation methods proposed in this paper were tested with SI continuous speech recognition experiments using the LATINO-40 database [4]. This database is composed of speech from 40 Latin American native speakers, each reading 125 sentences

from newspapers in Spanish. The training utterances were 4500 uncoded clean sentences provided by 36 speakers and context – dependent phoneme HMMs were employed. The vocabulary 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. Speech signals were sampled at a rate of 8000 samples/second and 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, SS was applied 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.

The noise was estimated using only 10 non-speech frames before the beginning of the utterance, and α_0 and β in (3) were made equal to 4 and 0.5, respectively. The 500 testing clean signals were used to create the noisy utterances by adding car and speech noise from the Noisex database at 2 global-SNR levels: +18dB and +12dB. Then, the noisy signals were coded and decoded using the 8 kbps CS-CELP (ITU-T, 1996). The coding-decoding distortion parameters M^d and V^d were estimated by directly aligning uncoded and coded-decoded training utterances.

The techniques are indicated as follow: *Baseline* without any HMM compensation; *SWV-SS*, the *SWV* algorithm combined with *SS*; and, *SWV-SS-CDC* the *SWV* algorithm combined with both *SS* and coding distortion compensation (*CDC*). The results are shown in Tables 1 and 2. The baseline system with clean signal gave a WER equal to 5.9%.

6. DISCUSSION AND CONCLUSIONS

As can be seen in Tables 1 and 2, 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 CDC 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 1 and 2, 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 [6] at low SNR should certainly increase the effectiveness of the approach proposed here.

Table 1: WER (%) with signal corrupted with additive noise (car noise) and coded by 8kbps CS-CELP.

SNR	18dB	12dB
<i>Baseline</i>	27.4	38.5
<i>SWV-SS</i>	11.9	18.3
<i>SWV-SS-CDC</i>	10.2	16.9

Table 2: WER (%) with signal corrupted with additive noise (speech noise) and coded by 8kbps CS-CELP.

SNR	18dB	12dB
<i>Baseline</i>	26.2	37.9
<i>SWV-SS</i>	11.7	17.5
<i>SWV-SS-CDC</i>	10.0	15.3

7. ACKNOWLEDGEMENTS

This research was funded by Conicyt, Chile.

8. REFERENCES

- [1] Gales, M.J.F. and Young, S.J. (1993). *HMM recognition in noise using parallel model combination*. Proc. Eurospeech93, pp. 837-840, 1993.
- [2] Hermansky, H. et al. (1991). *Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)*. Proc. Eurospeech 91, pp.1367-1370, 1991.
- [3] Huerta, J.M. (2000). *Speech Recognition in Mobile Environments*. Ph.D thesis, Dept.of Elec. and Comp. Engineering, Carnegie Mellon University, April, 2000.
- [4] LDC (1995). Latino-40 database provided by *Linguistic Data Consortium* (LDC), University of Pennsylvania: <http://www ldc.upenn.edu/>
- [5] Vaseghi, S.V. and Milner, B.P. (1997). *Noise compensation methods for Hidden Markov Model speech recognition in adverse environments*. IEEE Trans. on Speech and Audio Processing, 5 (1): 11-21, 1997.
- [6] Yoma, N.B., Villar, M. (2002). *Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm*. IEEE Trans. on Speech and Audio Processing, Vol. 10, No.3, pp. 158-166, 2002.
- [7] Yoma, N.B., Brito, I., Silva, J. *Uncertainty in noise canceling and language model perplexity in the Stochastic Weighted Viterbi*. Proc. Eurospeech 2003.
- [8] Yoma, N.B., Silva, J., Brito, I, Busso, C. (2002). *Modelling, estimating and compensating low-bit rate coding distortion in speech recognition*. Submitted to IEEE Trans. on SAP, September, 2002.