

FEATURE-DEPENDENT COMPENSATION IN SPEECH RECOGNITION

Iván Brito, Néstor Becerra Yoma, Carlos Molina

Electrical Engineering Department, University of Chile
 Av. Tupper 2007, P.O.Box 412-3, Santiago, CHILE.
nbecerra@ing.uchile.cl

ABSTRACT

Several mismatch conditions can be modeled as an additive bias. This bias is considered independent of the observation vectors, although this approximation is not always accurate. In this paper the dependence of the bias on the observation vectors is taken into consideration in the context of compensating the GSM coding distortion in speech recognition. However, the results presented here can easily be generalized to deal with other types of mismatch. The coding-decoding distortion is modeled here as feature-dependent. This model is employed to propose an Expectation-Maximization (EM) estimation algorithm of the coding-decoding distortion that is able to cancel the effect of GSM coder with as few as one adapting utterance. Finally, the feature-dependent adaptation can give word error rate (WER) 26% lower than the feature-independent model.

1. INTRODUCTION

Adaptation to Lombard effect [1], speaker [2] and noisy environments [3] are achieved by modeling and estimating an additive bias in the output probability density function. The estimation algorithms consider the bias as independent of the observation vectors. This assumption is probably accurate in some cases, but this model can certainly be improved.

The impressive growth of mobile networks around the world has created the problem of improving the recognition accuracy for speech distorted by low-bit rate coders. The GSM standard is certainly the most popular and this paper focuses on estimating and compensating the distortion of GSM coder in speech recognition. This distortion cannot be solved by applying conventional noise canceling techniques [6]: spectral subtraction [8]; Rasta [5]; PMC [4]; and, Cepstral Mean Normalization. Empirical observations suggested that the coding-decoding distortion in cepstral coefficient n in frame t could be modelled as [10]:

$$O_{t,n}^o = O_{t,n}^d + D_n \quad (1)$$

where $O_{t,n}^o$ and $O_{t,n}^d$ are the cepstral coefficients corresponding to the original and coded-decoded, speech signal, respectively; D_n is the distortion caused by the coding-decoding process with p.d.f. $f_{D_n}(D_n) = N(m_n^d, v_n^d)$, which does not depend on the value of the cepstral coefficient and is modelled as a Gaussian distribution with mean $m_n^d = E[D_n] = E[O_{t,n}^o - O_{t,n}^d]$ and variance $v_n^d = Var[D_n]$. According to [10], the HMM compensation is achieved by replacing the output probability by its expected value in the Viterbi algorithm [9], which in turn leads to replacing the observed $O_{t,n}^d$ and variance with:

$$E[O_{t,n}^o] = O_{t,n}^d + E[D_n] \quad (2)$$

$$V_{tot_{h,s,g,n}}^d = Var_{h,s,g,n} + v_n^d \quad (3)$$

where $E[O_{t,n}^o]$ is the expected value of the unseen cepstral coefficient $O_{t,n}^o$, considered as a random variable, in the original speech signal according to (1); $Var_{h,s,g,n}^d$ and $Var_{h,s,g,n}$ are, respectively, the compensated and original variances in HMM h , state s , Gaussian component g and coefficient n . Note that $m_n^d = E[D_n]$ and v_n^d could be considered as approximately constant in some cases (Figs. 1-2) [10].

2. FEATURE-DEPENDENT CODING-DECODING DISTORTION IN GSM CODEC

The model described by (1), (2) and (3) describes well the distortion caused by the G.729 CS-CELP coder. However, according to Fig. 1, the mean distortion $E[D_n] = E[O_{t,n}^o - O_{t,n}^d]$ in the GSM coder clearly depends on the value of the cepstral coefficient:

$$E[D_n] = E[O_{t,n}^o - O_{t,n}^d] = f(O_{t,n}^o) \quad (4)$$

According to Fig 1, this dependence could be modeled as:

$$f(O_{t,n}^o) = B_n \cdot O_{t,n}^o + A_n \quad (5)$$

where B_n and A_n are constants. In contrast to [10], the expected value of the GSM coded-decoded distortion depends on the uncoded speech feature. This problem could be counteracted by applying the expected value operator again to (4):

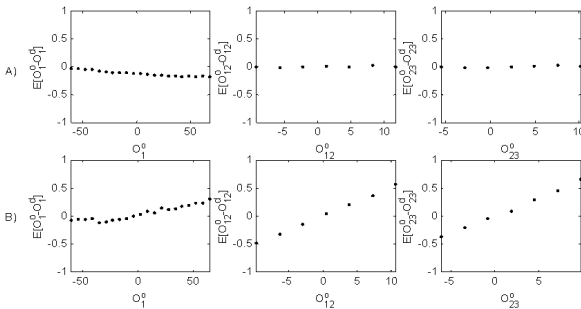
$$E[D_n] = E[O_{t,n}^o - O_{t,n}^d] = E[f(O_{t,n}^o)] \quad (6)$$

Considering (5) and (6), $E[O_{t,n}^o]$ can be written as:

$$E[O_{t,n}^o] = \frac{O_{t,n}^d + A_n}{1 - B_n} \quad (7)$$

Although not shown here, the coding-decoding distortion variance, v_n^d , also depends on the value of the uncoded speech feature. Nevertheless, the introduction of this dependence does not result in analytical solutions in the EM estimation algorithm employed here. Consequently, v_n^d was supposed to be a constant, and the HMM is compensated with (7), instead of (2), and with (3).

Figure 1: Expected value of the coding-decoding error $E[D_n]$ 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) 13kbps GSM. The cepstral coefficients correspond to a static (1), a delta (12) and a delta-delta (23).



3. ESTIMATION OF GSM CODING-DECODING DISTORTION

Estimating the coding-decoding distortion is equivalent to find the vectors A_n and B_n . In this paper these parameters are estimated with the EM algorithm using a code-book, where every code-word corresponds to a

multivariate Gaussian, built with uncoded speech signals. Inside each code-word cw_j the mean

$$\mu_j^o = [\mu_{j,1}^o, \mu_{j,2}^o, \dots, \mu_{j,n}^o, \dots, \mu_{j,N}^o] \quad \text{and} \quad \text{variance}$$

$$(\sigma_j^o)^2 = [(\sigma_{j,1}^o)^2, (\sigma_{j,2}^o)^2, \dots, (\sigma_{j,n}^o)^2, \dots, (\sigma_{j,N}^o)^2] \quad \text{are}$$

computed, where N is the number of cepstral coefficients and the dimension of the code-book. If there are J code-words, the p.d.f. associated to the frame $O_t^o = [O_{t,1}^o, O_{t,2}^o, \dots, O_{t,n}^o, \dots, O_{t,N}^o]$ given the uncoded speech signal model is:

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

where $\Phi^o = \{\phi_j^o \mid 1 \leq j \leq J\}$ and $\phi_j^o = (\mu_j^o, \Sigma_j^o)$; $Pr(cw_j)$

is the *a priori* probability of codeword j ; and, Σ_j^o

is the N -by- N covariance matrix that is supposed diagonal.

If A_n and B_n are considered independent of the code-word or class, it is possible to show that the coded-decoded speech signal is represented by the model whose parameters are denoted by

$\Phi^d = \{\phi_{j,t}^d \mid 1 \leq j \leq J\}$, where $\phi_{j,t}^d = (\mu_{j,t}^d, \Sigma_j^d)$ and,

$$\mu_{j,n,t}^d = \mu_j^o - E[D_n] = \mu_{j,n,t}^o - f(O_{t,n}^o) \quad (9)$$

$$(\sigma_{j,n}^d)^2 = (\sigma_{j,n}^o)^2 + v_n^d \quad (10)$$

Again, applying the expected value to (9) as in (6), and

replacing $E[O_{t,n}^o]$ with (7), $\mu_{j,n,t}^d$ can be written as:

$$\mu_{j,n,t}^d = \mu_{j,n,t}^o - Z_n \cdot (A_n + O_{t,n}^d) - A_n \quad (11)$$

where $Z_n = \frac{B_n}{1 + B_n}$. In this paper A_n , B_n and v_n^d are

estimated with the maximum likelihood criterion using adaptation utterances. The maximization of the likelihood does not lead to analytical solutions, so the EM algorithm was employed. Given an adaptation

utterance O^d distorted by a coding-decoding scheme and composed of T

frames, $O^d = [O_1^d, O_2^d, \dots, O_t^d, \dots, O_T^d]$, O^d is also called

observable data. The unobserved data is $Y^d = [y_1^d, y_2^d, \dots, y_t^d, \dots, y_T^d]$ where y_t^d is the hidden

number that refers to the code-word or density of the observed frame O_t^d . The EM algorithm defines the

function $Q(\Phi, \hat{\Phi})$:

$$Q(\Phi, \hat{\Phi}) = E \left[\log \left(f(O^d, Y^d / \hat{\Phi}) \right) \middle| O^d, \Phi \right] \quad (12)$$

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

1. Start with $\Phi = \Phi^o$, where

$$\Phi = \{\phi_{j,t} \mid 1 \leq j \leq J, 1 \leq t \leq T\} \quad \text{and} \\ \phi_{j,t} = (\mu_{j,t}^o, \Sigma_j^o).$$

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

$$Pr(cw_j \mid O_t^d, \phi_j) = \frac{f(O_t^d / \phi_{j,t}) \cdot Pr(cw_j)}{\sum_{k=1}^J f(O_t^d / \phi_{k,t}) \cdot Pr(cw_k)} \quad (13)$$

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

$$\hat{Pr}(cw_j) = \frac{1}{T} \sum_{t=1}^T Pr(cw_j \mid O_t^d, \phi_{j,t}) \quad (14)$$

4. Estimate \hat{A}_n with

$$\hat{A}_n = \frac{\sum_{t=1}^T \sum_{j=1}^J \left(\hat{Pr}(cw_j \mid O_t^d, \phi_{j,t}) \cdot \frac{(\mu_{j,n}^o - O_{t,n}^d - Z_n \cdot O_{t,n}^d)}{\sigma_{j,n}^2} \right)}{(Z_n + 1) \cdot \sum_{t=1}^T \sum_{j=1}^J \left(\frac{\hat{Pr}(cw_j \mid O_t^d, \phi_{j,t})}{\sigma_{j,n}^2} \right)} \quad (15)$$

5. Estimate \hat{Z}_n

$$\hat{Z}_n = \frac{\sum_{t=1}^T \sum_{j=1}^J \left(\hat{Pr}(cw_j \mid O_t^d, \phi_{j,t}) \cdot \frac{(\mu_{j,n}^o - A_n - O_{t,n}^d) \cdot (A_n + O_{t,n}^d)}{\sigma_{j,n}^2} \right)}{\sum_{t=1}^T \sum_{j=1}^J \left(\frac{\hat{Pr}(cw_j \mid O_t^d, \phi_{j,t}) \cdot (A_n + O_{t,n}^d)^2}{\sigma_{j,n}^2} \right)} \quad (16)$$

6. Update $\hat{\mu}_{j,n,t}^d$ $1 < j < J$, $1 < n < N$, and $1 < t < T$ with (11).

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

$$\hat{\sigma}_{j,n}^2 = \frac{\sum_{t=1}^T \hat{Pr}(cw_j \mid O_t^d, \phi_{j,t}) \cdot (O_{t,n}^d - \hat{\mu}_{j,n,t}^d)^2}{\sum_{t=1}^T \hat{Pr}(cw_j \mid O_t^d, \phi_{j,t})} \quad (17)$$

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

$$f(O^d / \hat{\Phi}) = \sum_{t=1}^T \sum_{j=1}^J f(O_t^d / \hat{\phi}_{j,t}) \cdot \hat{Pr}(cw_j) \quad (18)$$

9. Update parameters: $Pr(cw_j) = \hat{Pr}(cw_j)$;

$$A_n = \hat{A}_n; B_n = \hat{B}_n; \sigma_{j,n}^2 = \hat{\sigma}_{j,n}^2.$$

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

11. Estimate v_n^d :

$$v_n^d = \frac{\sum_{j=1}^J [\sigma_{j,n}^2 - (\sigma_{j,n}^o)^2] \cdot Pr(cw_j)}{\sum_{j=1}^J Pr(cw_j)} \quad (19)$$

for any $1 < j < J$ where $1 < n < N$.

Note that the maximization of $Q(\Phi, \hat{\Phi})$ does not lead to an analytical solution for v_n^d , which in turn was estimated with (19). Finally, if $Z_n = 0$, the distortion model is similar to the one employed in [10].

4. EXPERIMENTS

The compensation method was tested with speaker-independent continuous speech recognition experiments using LATINO-40 database [7]. This database is composed of 40 Latin American native speakers, each reading 125 sentences from newspapers in Spanish. The training utterances were 4500 uncoded clean sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary has almost 6000 words. The testing database was composed of 500 utterances provided by 4 testing speakers (two females and two males). Each context-dependent phoneme was modelled with a 3-state left-to-right topology, with 8 multivariate Gaussian densities per state and diagonal covariance matrices. Trigram language modelling was employed. The frame energy plus ten Mel Frequency Cepstral Coefficients (MFCC), and their first and second time derivatives were computed. The 500 testing uncoded

signals were coded and decoded with the 8kHz G.729 CS-CELP and the 13kHz GSM coders to create the corrupted database. The techniques that were employed are indicated as follows: *Baseline*, without HMM compensation; *Feature-Independent-Compensation*, where the EM algorithm presented here estimated only A_n and v_n^d by making $B_n = Z_n = 0$; and, *Feature-Dependent-Compensation*, where the EM algorithm computed A_n , B_n and v_n^d . The EM algorithm estimated the coding-decoding distortion sentence-by-sentence. The code-book was composed of 256 code-words and was generated with the uncoded training signals. The results are shown in Table 1. The baseline system with uncoded speech gave a WER equal to 5.9%.

5. CONCLUSIONS AND DISCUSSIONS

As can be seen in Table 1, the feature-dependent compensation method proposed here completely cancelled the coding-decoding distortion and gave a WER 26% lower than the one achieved with the feature-independent model in experiments with GSM coder. This result confirms the hypothesis suggested by Fig. 1. However, the feature-dependent and independent estimation techniques gave similar WER with the G.729 CS-CELP. This result is also consistent with Fig. 1 where the G.729 CS-CELP average coding-decoding distortion almost does not depend on the value of the feature. Note that the compensation methods gave a WER even lower than the baseline system with uncoded speech. This is probably due to the fact that the compensation also provides an adaptation to testing condition beyond the type of codification (e.g. speaker adaptation). It is worth emphasizing that the EM estimation needs only one adapting utterance and the approach described here is certainly suitable for telephone dialogue systems. Finally, reducing the computational load of the estimation algorithm, the feature-dependent computation of the coding-distortion variance, and the applicability of the method to speaker adaptation are proposed as future work.

Table 1: WER (%) with signals processed with 8 kbps G.729 CS-CELP and 13 kbps GSM coders.

CODER	Baseline	FI Comp.	FD Comp.
CS-CELP	6.2	2.7	2.4
GSM	7.5	3.4	2.5

6. ACKNOWLEDGEMENTS

This research was funded by Conicyt, Chile.

7. REFERENCES

- [1] Afify M., Gong Y. and Haton J. *A General Joint Additive and Convolutional Bias Compensation Approach Applied to Noise Lombard Speech Recognition*. IEEE Transaction Speech and Audio Processing. Vol. 6, No. 6, pp. 524-538, November 1998.
- [2] Gauvain J., Lee, C-H. *Maximum a posteriori estimation for multivariate Gaussian Mixture Observation Chains*, IEEE SAP April 1994.
- [3] Raj, Gouvea E. B., Moreno P. J. and Stern R. M. *Cepstral compensation by polynomial approximation for environment-independent speech recognition*. Proc. of Int. Conf. Spoken Language Processing, Philadelphia, PA, pp. 2340-2343, Oct. 1996.
- [4] Gales, M.J.F. and Young, S.J. *HMM recognition in noise using parallel model combination*. Proc. Eurospeech93, pp. 837-840, 1993.
- [5] Hermansky, H. et al. *Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)*. Proc. Eurospeech 91, pp.1367-1370, 1991.
- [6] Huerta, J.M. *Speech Recognition in Mobile Environments*. Ph.D thesis, Dept. of Elec. and Comp. Engineering, Carnegie Mellon University, April, 2000.
- [7] LDC. Latino-40 database provided by *Linguistic Data Consortium* (LDC), University of Pennsylvania.
- [8] Vaseghi, S.V. and Milner, B.P. *Noise compensation methods for Hidden Markov Model speech recognition in adverse environments*. IEEE Trans. on Speech and Audio Processing, 5 (1): 11-21, 1997.
- [9] Yoma, N.B., Villar, M. *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.
- [10] Yoma, N. B., Silva, J., Brito, I. and Busso, C. *Modeling, estimating and compensating low-bit rate coding distortion in speech recognition*. Submitted to IEEE Transactions on Speech and Audio Proc. 2002.