A UNIFIED MAXIMUM LIKELIHOOD APPROACH TO ACOUSTIC MISMATCH COMPENSATION : APPLICATION TO NOISY LOMBARD SPEECH RECOGNITION

Mohamed Afify[1] *Yifan Gong*[2] *Jean-Paul Haton*[1]

[1]CRIN/CNRS-INRIA-Lorraine,B.P. 239 54506 Vandeouvre,Nancy,France

[2] Speech Research,Personal Systems Laboratory,Texas Instruments, P.O.BOX 655303 MS 8374, Dallas TX 75265, U.S.A.

ABSTRACT

In the context of continuous density hidden Markov model (CDHMM) we present a unified maximum likelihood (ML) approach to acoustic mismatch compensation. This is achieved by introducing additive Gaussian biases at the state level in both the mel cepstral and linear spectral domains. Flexible modelling of different mismatch effects can be obtained through appropriate bias tying. A Maximum likelihood approach for joint estimation of both mel cepstral and linear spectral biases from the observed mismatched speech given only one set of clean speech models is presented, where the obtained bias estimates are used for the compensation of clean speech models during decoding. The proposed approach is applied to the recognition of noisy Lombard speech, and significant improvement in the word recognition rate is achieved.

1. INTRODUCTION

It is well known that acoustic mismatch can cause severe degradation to the performance of current speech recognition systems. Also recent research results indicate that most sources of acoustic mismatch occuring in practice can be modelled as an additive bias in either the mel cepstral or linear spectral domains at different levels (e.g. phoneme, broad phonetic class, all speech,....)[1].In the linear spectral domain, stationary additive white noise is an additive bias that is common to all speech, whereas stationary additive coloured noise has additive bias components at different levels depending on its frequency response. Speaker difference can be viewed as an additive bias in the cepstral domain having both a global component and phone dependent components [5]. A communication channel or a transducer is also an additive bias in the cepstral domain which is common to all speech events. Finally it was shown that stressed speech and Lombard effect can be modelled by using an additive bias at the word level [6], or the broad phonetic level [7].

To provide a general model of acoustic mismatch in the framework of the continuous density hidden Markov model (CDHMM) we propose to introduce additive Gaussian biases at the state level in both the linear spectral and mel cepstral domains, where different mismatch effects can be modelled through appropriate bias tying. A Maximum likelihood approach to joint estimation of both mel cepstral and linear spectral biases from the observed mismatched speech given only one set of clean speech models defined in the mel cepstral domain is presented. These bias estimates are used for the compensation of clean speech models during decoding. Based on this formulation specific bias models in the mel cepstral and linear spectral domains are derived, among which a new polynomial trend cepstral bias model proved effective for Lombard speech compensation.

The proposed approach is applied to the recognition of noisy Lombard speech, where (based on previous research evidence)it is assumed that the mismatch has both mel cepstral and linear spectral components. Using the joint bias compensation algorithm significant improvement in word recognition accuracy is obtained, when using a limited amount of training data in the mismatch environment (no separate stress database is required).

The paper is organized as follows. Section 2 gives a general formulation of the problem. Bias parameter estimation and model compensation are considered in Sections 3 and 4 respectively. Section 5 shows the implementation of the algorithm to noisy Lombard speech. Experimental results are given in Section 6 followed by conclusion in Section 7.

2. PROBLEM FORMULATION

In a CDHMM framework, we assume that for model state i, the observed mismatched speech (Y) is obtained through the corruption of clean speech (X) by *state dependent, statistically independent, additive, Gaussian* biases in both the *mel cepstral* (B_i^c) and the *linear spectral* (B_i^l) domains. In the mel cepstrum domain this can be expressed mathematically as:

$$Y = C\log(\exp(C^{-1}(X + B_i^c)) + \exp(C^{-1}B_i^l)) \qquad (1)$$

Where C is the DCT transformation matrix, and exp() and log() means component-wise exponential and logarithm functions respectively.

The ML bias model parameter estimates given the observed mismatched speech Y and the clean speech models λ_x can be written as:

$$(\lambda_{b_c}^*, \lambda_{b_l}^*) = \underset{(\lambda_{b_c}, \lambda_{b_l})}{\operatorname{argmax}} P(Y|\lambda_x, \lambda_{b_c}, \lambda_{b_l}) \qquad (2)$$

Due to the nonlinear nature of the global optimization in (2) we propose to successively improve the bias estimates through the iterative application of the local maximizations in (3) and (4).

$$\lambda^*_{b_c} = \underset{\lambda_{b_c}}{\operatorname{argmax}} P(Y|\lambda_x, \lambda_{b_c}, \lambda^*_{b_l}) \quad (3)$$

$$\lambda^*_{b_l} = \underset{\lambda_{b_l}}{\operatorname{argmax}} P(Y|\lambda_x, \lambda_{b_l}, \lambda^*_{b_c}) \quad (4)$$

Parallel model combination (PMC) [2] provides a flexible framework for model combination and tranformation. A generalized view of PMC consists of applying the following model combinations and transformations:

- A transformation $\mathcal{T}(\lambda)$ from the mel cepstral to the linear spectral domains.
- Additive model combination $\mathcal{C}(\lambda_1, \lambda_2)$(which can be applied in either the mel cepstral or linear spectral domains).
- And by making the assumption that the sum of two lognormal variables is also lognormal, an inverse transformation $\mathcal{T}^{-1}(\lambda)$ from the linear spectral to the mel cepstral domains.

Details of the transformations can be found in e.g. [2],[9].

Using the above model combinations and transformations (3) and (4) can be rewritten as:

$$\lambda^*_{b_c} = \underset{\lambda_{b_c}}{\operatorname{argmax}} P(Y|\lambda_{xl}, \lambda_{b_c}) \quad (5)$$

$$\lambda^{l*}_{b_l} = \underset{\lambda^l_{b_l}}{\operatorname{argmax}} P(Y^l|\lambda^l_{xc}, \lambda^l_{b_l}) \quad (6)$$

where λ_{xl} , λ^l_{xc} ,and $\lambda^l_{b_l}$ are given by:

$$\lambda_{xl} \stackrel{\text{def}}{=} \mathcal{T}^{-1}(\mathcal{C}(\mathcal{T}(\lambda_x), \mathcal{T}(\lambda^*_{b_l}))) \quad (7)$$

$$\lambda^l_{xc} \stackrel{\text{def}}{=} \mathcal{T}(\mathcal{C}(\lambda_x, \lambda^*_{b_c})) \quad (8)$$

$$\lambda^l_{b_l} \stackrel{\text{def}}{=} \mathcal{T}(\lambda_{b_l}) \quad (9)$$

and Y^l is the mismatched speech transformed to the linear spectral domain.

The basic idea of the proposed algorithm is to use only one set of clean speech models λ_x defined in the mel cepstral domain to successively obtain bias parameter estimates through the application of (5) and (6)in conjunction with the model combinations and transformations (7)-(9). These bias estimates are used for clean speech model compensation during decoding.

3. PARAMETER ESTIMATION

In this section we present a maximum likelihood approach to bias parameter estimation. A unified view of both (5) and (6) can be written as:

$$\lambda^*_b = \underset{\lambda_b}{\operatorname{argmax}} P(Y|\lambda_x, \lambda_b) \quad (10)$$

The problem in (10) is that of ML estimation of additive bias parameters $\lambda_b \stackrel{\text{def}}{=} (\mu_{b,i}, \Sigma_{b,i})$, where statistics are Gaussian for (5) and lognormal (due to mel cepstrum to linear spectral transformation) for (6). An EM based solution for this parameter estimation problem in the Gaussian case can be found in the literature(e.g. [3],[8]), the mean and covariance estimates for one obseravtion sequence can be written as:

$$\mu_{b,i} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)E[b_t|y_t,k,i]}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)} \quad (11)$$

$$\Sigma_{b,i} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)E[b_t b_t^T|y_t,k,i]}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)} - \mu_{b,i}\mu_{b,i}^T \quad (12)$$

where T is the observation sequence length, M is the number of mixture components per state, and $\gamma_t(i,k) = P(s_t = i, m_t = k|Y)$ which can be calculated using the forward backward algorithm. The expected values in equations (11) and (12) are calculated as:

$$E[b_t|y_t,k,i] = \Sigma_{b,i}(\Sigma_{x,i,k} + \Sigma_{b,i})^{-1}(y_t - \mu_{x,i,k}) + \Sigma_{x,i,k}(\Sigma_{x,i,k} + \Sigma_{b,i})^{-1}\mu_{b,i} \quad (13)$$

$$E[b_t b_t^T|y_t,k,i] = E[b_t|y_t,k,i]E[b_t^T|y_t,i,k] + \Sigma_{b,i}\Sigma_{x,i,k}(\Sigma_{x,i,k} + \Sigma_{b,i})^{-1} \quad (14)$$

Thus, bias training consists of the iterative application of the E-step (equations (13)-(14)), and the M-step (equations (11)-(12)) starting from some initial value of the bias, where each iteration ensures the increase of the observed data likelihood. The above EM algorithm can be extended to the lognormal case by considering the following :

- It is regarded as a first and second moment matching technique rather than an exact ML algorithm.
- No closed form expressions exist for the required expectations, and these expectations can be obtained by using numerical integration.

The output of this training procedure will be the optimal bias parameter estimates in the mel cepstral ($\lambda^*_{b_c}$) and the linear spectral ($\lambda^{l*}_{b_l}$) domains.

4. MODEL COMPENSATION

After obtaining the optimal bias parameter estimates they are used in compensating the clean speech models. The model compensation process can be summarized as follows:

- Combine λ_x and $\lambda^*_{b_c}$ in the cepstral domain, which can be represented as $\lambda_{xc} = \mathcal{C}(\lambda_x, \lambda^*_{b_c})$.
- Combine the resulting model and $\lambda^{l*}_{b_l}$ in the linear spectral domain, which can be represented as $\lambda^l_{xcl} = \mathcal{C}(\mathcal{T}(\lambda_{xc}), \lambda^{l*}_{b_l})$.
- Transform the combined model to the cepstral domain, which can be expressed as $\lambda_{xcl} = \mathcal{T}^{-1}(\lambda^l_{xcl})$. This model is used for mismatched speech decoding.

5. IMPLEMENTATION ISSUES

The implementation focuses on noisy Lombard speech recognition. Some implementation issues of the algorithm in both the mel cepstral and linear spectral domains are given below.

5.1. Cepstral domain

In the cepstral domain two bias models are considered. An *Independent bias model* where bias components are assumed statistically independent, and a *Polynomial trend bias model* where bias components are assumed to follow a polynomial function along the cepstral dimension. The use of the polynomial trend model results in better utilization of bias training data, and is also a tractable approximation to the empirical finding in [6] that stressed speech bias follows an exponential trend.

In the case of independent bias model a scalar version of the estimation algorithm presented above can be separately applied to each cepstral component. The estimation equations remain the same as (11) and (12), while the expected values calculation is simplified to:

$$E[b_t|y_t,k,i] = \frac{\sigma^2_{b,i}}{\sigma^2_{x,i,k}+\sigma^2_{b,i}}(y_t-\mu_{x,i,k}) + \frac{\sigma^2_{x,i,k}}{\sigma^2_{x,i,k}+\sigma^2_{b,i}}\mu_{b,i} \tag{15}$$

$$E[b_t^2|y_t,k,i] = E^2[b_t|y_t,k,i] + \frac{\sigma^2_{b,i}\sigma^2_{x,i,k}}{\sigma^2_{x,i,k}+\sigma^2_{b,i}} \tag{16}$$

In the case of polynomial trend bias we consider the bias b_t generated according to the equation:

$$b_t = Za + e_t \tag{17}$$

where b_t is a $P \times 1$ bias vector,
e_t is a zero mean $P \times 1$ error vector, assumed Gaussian,
a is a $Q \times 1$ coefficient vector,
and Z is a $P \times Q$ matrix of powers, where $Z_{p,q} = p^q$.

It can be shown that (see [4]) the EM parameter estimates corresponding to those in Section 3 can be obtained as:

$$a_{b,i} = \frac{(Z^TZ)^{-1}Z^T\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)E[b_t|y_t,i,k]}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)} \tag{18}$$

$$\Sigma_{b,i} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)E[b_tb_t^T|y_t,i,k]}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(i,k)} - (Za_{b,i})(Za_{b,i})^T \tag{19}$$

Furthermore if the error covariance matrix $\Sigma_{b,i}$ is assumed to be diagonal, the expected values in the (18)and(19) can be calculated from (15) and (16). In all experiments with the polynomial trend model the covariance is assumed to be diagonal and a third order function is used based on preliminary experimentation.

Appropriate bias tying is a key issue in the successful application of the proposed algorithm. In this work three cepstral bias patterns are empirically evaluated, these patterns consist of:

- 8 bias vectors/word denoted $B8$.
- 4 bias vectors/word denoted $B4$.
- 1 bias vector/word denoted $B1$.

For all the three patterns the bias variance is tied for all states of all words.

5.2. Linear spectral domain

In implementing the linear spectral bias estimation algorithm (6) we face the following difficulties:

- Expected values have no closed form solution.
- Full covariance matrices are used in the linear spectral domain,making the approach computationally expensive.
- A suitable gain mismatch compensation technique should be developed.

Hence, in this work linear spectral bias is estimated from the speech pauses and used in (5). However this simplified form of the algorithm jointly compensates for both additive and convolutive biases. And as we are treating stationary additive white noise we use a single bias pattern tied to all states of all words.

5.3. Outline of the training algorithm

In this subsecton we give an outline of the implemented training algorithm.

1. Estimate linear spectral bias from speech pauses.
2. Form the linearly compensated model λ_{xl} as in (7).
3. For each training sample using λ_{xl}
   - Perform forward backward algorithm and calculate $\gamma_t(i,k)$.
   - Calculate expected values as in (15) and (16).
   - Update counters.
4. Estimate cepstral bias parameters (equations (11)-(12)) for indpendent bias model, or (equations (18)-(19)) for polynomial trend model.
5. If convergence is not met goto step 3.

6. EXPERIMENTAL RESULTS

The speech database consists of 21 confusable words, uttered by 24 speakers (12 male/ 12 female), each word is uttered two times by each speaker. The same speech corpus is available under the following conditions, representing the mismatched environment: Lombard (no noise added), 15 db additive white noise and Lombard, 5 db additive white noise and Lombard. Speaker independent clean speech recognition rate using an eight state left to right continuous density HMM with 2 mixture components per state and defined in a 12 dimension MFCC space is 61%. Bias training data consists of one repitition of each word from a male and female speaker. Results of applying joint bias compensation with independent cepstral bias model are shown in Table 1. Note that for Lombard speech without noise only cepstral compensation is applied. Table 2 is the same as Table 1, but for the polynomial trend cepstral bias model. Table 3 shows the results of applying cepstral compensation (independent model)only to the noisy Lombard conditions.

In all experiments a significant improvement in recognition accuracy is obtained using a limited amount of training data in the mismatch environment, which indicates the efficiency of the proposed algorithm and its generalization capability. For cepstral compensation i.e. only Lombard

effect both the independent and polynomial trend models perform very similarly, with the polynomial model using a smaller number of parameters, which shows the validity of the polynomial assumption. For joint cepstral and spectral bias the proposed approach significantly outperform both cepstral compensation and PMC (see, for example, 5dB noisy Lombard, 4 bias/word, 32.4%) than either independent cepstral bias model (19.6%) or PMC (19.0%). However, in this case the independent bias model outperforms the polynomial trend model, this can be attributed to the fact linear spectral compensated noisy Lombard speech has a complex pattern in the cepstral which doesn't fit the polynomial assumption. It is also interesting to note the effect of bias tying, in experiments with Lombard speech the pattern $B1$ gave best performance, while for noisy Lombard speech the patterns $B8$ and $B4$ gave superior results, which can be attributed to the higher variability of the noisy Lombard speech.

	B8	B4	B1	No comp.	PMC
Lombard	46.7	46.5	48.3	37.5	N.A.
15dB NL	36.8	38.9	36.8	9.6	23.8
5dB NL	33.1	32.4	30.1	6.6	19.0

Table 1. Recognition rates (%) obtained when applying the joint bias compensation algorithm with independent cepstral bias model for different mismatch conditions. Results using clean speech models and PMC are also shown.

	B8	B4	B1	No comp.	PMC
Lombard	46.9	46.7	47.6	37.5	N.A.
15dB NL	33.9	33.2	33.4	9.6	23.8
5dB NL	30.8	29.5	28.9	6.6	19.0

Table 2. Recognition rates (%) obtained when applying the joint bias compensation algorithm with third order polynomial trend cepstral bias model for different mismatch conditions. Results using clean speech models and PMC are also shown.

	B8	B4	B1
15 dB NL	30.7	25.8	23.9
5 dB NL	20.3	19.6	17.4

Table 3. Recognition rates (%) for cepstral bias compensation of noisy Lombard speech with independent cepstral bias model.

7. CONCLUSION

In this paper we present a unified approach to joint compensation of additive and convolutive biases in the CDHMM framework. This is achieved by introducing state level additive biases in the mel cepstral and linear spectral domains. ML bias parameter estimation and model compensation are discussed, and a new polynomial trend bias model is proposed. The approach is applied to noisy Lombard speech recognition, and it significantly outperform both cepstral compensation and linear spectral compensation on a confusable word recognition task.In this work we have used a supervised version of the compensation algorithm and treated only the mel cepstral coefficients. Extensions to unsupervised adaptation, and compensation of the difference coefficients in the proposed framework are given in [4].

REFERENCES

[1] Y.Gong,"Speech recognition in noisy environments: A survey,"Speech communication, vol. 16, pp 261-291, June 1995.

[2] M.Gales, S.Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, Vol.12, No. 3, pp.231-239.

[3] A.Sankar, C.H.Lee, "Robust speech recognition based on stochastic matching," Proc. ICASSP-95, vol. 1, pp.121-124, 1995.

[4] M.Afify, Y.Gong, and J.P.Haton, "A general joint additive and convolutive bias compensation approach applied to noisy Lombard speech recognition," Submitted to IEEE Trans. Speech and Audio processing, May 1996.

[5] Y.Zhao,"An acoustic-phonetic based speaker adaptation technique for improving speaker independent continuous speech recognition ," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 3, pp.380-394, July 1994.

[6] Y.Chen, "Cepstral domain talker stress compensation for robust speech recognition," IEEE Trans. ASSP, vol. 36, pp.433-439, Apr. 1988.

[7] J.H.L.Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation for speech recognition in noise and Lombard effect," IEEE Trans. Speech and Audio Processing, Vol. 2, No.4 , pp.598-614, Oct. 1994.

[8] R.Rose, E.Hofstetter, D.Reynolds, "Integrated models of signal and background with application to speaker identification in noise, " IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp.245-257, Apr. 1994.

[9] M.Gales, S.Young, "A fast and flexible implementation of parallel model combination," Proc. ICASSP-95, vol. 1, pp.133-136, 1995.