CHANNEL AND NOISE ADAPTATION VIA HMM MIXTURE MEAN TRANSFORM AND STOCHASTIC MATCHING

Shuen Kong Wong and Bertram Shi

Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clearwater Bay, Hong Kong
eewsk95@engsvr.ust.hk eebert@ee.ust.hk

ABSTRACT

We present a non-linear model transformation for adapting Gaussian Mixture HMMs using both static and dynamic MFCC observation vectors to additive noise and constant system tilt. This transformation depends upon a few compensation coefficients which can be estimated from channel distorted speech via Maximum-Likelihood stochastic matching. Experimental results validate the effectiveness of the adaptation. We also provide an adaptation strategy which can result in improved performance at reduced computational cost compared with a straightforward implementation of stochastic matching.

1. INTRODUCTION

Gales and Young [4] proposed parallel model combination (PMC) for robust recognition of speech corrupted by additive and convolutional noise. PMC combines HMM models for clean speech with a model for noise to create HMM models for noisy speech. The means and variances of static MFCC features are modified according to a nonlinear function of the means and variances of clean speech and noise. One of the strengths of this method is that all of the speech models can be modified using a single transformation which depends only upon a few parameters. A similar transform was studied by Vaseghi and Miller [9].

Our previous work [11] showed that a transformation similar to those used above can be used to adapt not only the means of static MFCC coefficients, but also delta and acceleration coefficients, to the presence of additive noise. In brief, clean Mel-scaled Log-spectral Filterbank Coefficient (MLFC) means $\{m_1, m_2, \ldots, m_R\}$ are obtained by applying the inverse DCT to the clean Mel-scaled Frequency Cepstral Coefficient (MFCC) means of the HMM output probability distributions. Noise transformed MLFC means $\{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_R\}$ for the static and dynamic coefficients are obtained via the following transformations.

$$\mathbf{E}\left[\hat{m}_i\right] = K_i \ln(P_o^i + P_n^i) \tag{1}$$

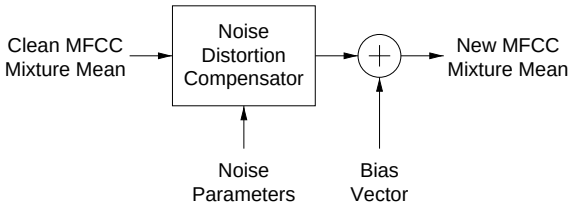

Figure 1: Gaussian mixture mean transformation for ML stochastic matching optimization

$$\mathbf{E}\left[\Delta^k \hat{m}_i\right] \approx \left(\frac{P_o^i}{P_o^i + P_n^i}\right)^k \cdot \mathbf{E}\left[\Delta^k m_i\right] \tag{2}$$

where the K_i's are pre-determined constants determined by the sub-band filter shape, the $\vec{P}_o = \{P_o^1, P_o^2, \ldots, P_o^R\}$ are determined from the equation $\mathbf{E}\left[m_i\right] = K_i \ln(P_o^i)$ and $\vec{P}_n = \{P_n^1, P_n^2, \ldots, P_n^R\}$ are estimates of the noise power within each sub-band filter. The noise transformed MLFC means are then converted to noise transformed MFCC means via the DCT.

At a signal to noise ratio of 10dB, transforming the static coefficients alone increased recognition accuracy to 81% as compared with the baseline performance of 8.9%. Transforming the dynamic coefficients as well resulted in a further 8% increase. Additionally, we showed that the noise parameters could not only be estimated from pure noise samples, but also from speech embedded in noise using maximum likelihood stochastic matching[8].

In this work, we extend this transformation to compensate for constant convolutional distortion by adding a bias $\vec{b} = \{b_1, b_2, \ldots, b_q\}$ to the MFCC means. See Figure 1. The bias parameters can also be adapted using ML stochastic matching. Experimental results demonstrate that combining the noise and bias adaptation can successfully compensate for both noise and channel distortions. In addition, we show that separating the estimation of the noise and bias parameters results in improved performance with less computational cost. However, the order in which the estimations

are performed is critical when we only have a good initial value of either $\vec{b}$ or $\vec{P}_n$. Finally, we demonstrate that the addition of the bias and ML stochastic matching can compensate for the approximations used in deriving the noise transformation. In particular, we show that using stochastic matching to adapt both channel and noise parameters is superior to adapting the noise parameters alone, even if the convolutional bias is known exactly.

2. MIXTURE MEAN OPTIMIZATION VIA ML STOCHASTIC MATCHING

Adaptation of the bias and noise parameters $\{\vec{P}_n, \vec{b}\}$ via stochastic matching is accomplished through the auxiliary function $Q(\cdot)$ and the EM algorithm. Due to the non-linearity of the transformation, it is difficult to find a closed form solution for the maximization step. Therefore, we solve the problem iteratively by gradient ascent. At each iteration, for a system using static and dynamic MFCC coefficients,

$$P_n^{i'}(k) = P_n^{i'}(k-1) + \epsilon(k)\frac{\partial Q(P_n^{i'}, \vec{b}'|\vec{P}_n, \vec{b})}{\partial P_n^{i'}}$$

$$b_i'(k) = b_i'(k-1) + \epsilon(k)\frac{\partial Q(b_i', \vec{P}_n'|\vec{P}_n, \vec{b})}{\partial b_i'}$$

where the gradient components are given by

$$\frac{\partial Q(P_n^{i'}, \vec{b}'|\vec{P}_n, \vec{b})}{\partial P_n^{i'}} = \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M} p(t,n,m|W,\vec{P}_n,\vec{b},\Lambda_X) \cdot \left\{ \sum_{d=1}^{q} \frac{y_{t,d} - \sum_{k=1}^{R} c_{kd}\mathbf{E}\left[\hat{m}_k'\right] - b_d}{\xi_{mn,d}} \cdot \frac{c_{id}K_i}{P_o^i + P_n^{i'}} - \sum_{d=1}^{q} \frac{\Delta y_{t,d} - \sum_{k=1}^{R} c_{kd}\mathbf{E}\left[\Delta\hat{m}_k'\right] - b_d}{\Delta\xi_{nm,d}} \cdot \frac{c_{id}\mathbf{E}\left[\Delta\hat{m}_i'\right]}{P_o^i + P_n^{i'}} \right\}$$

and,

$$\frac{\partial Q(b_i', \vec{P}_n'|\vec{P}_n, \vec{b})}{\partial b_i'} = \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M} p(t,n,m|W,\vec{P}_n,\vec{b},\Lambda_X) \cdot \sum_{d=1}^{q} \frac{y_{t,d} - \sum_{k=1}^{R} c_{kd}\mathbf{E}\left[\hat{m}_k'\right] - b_d}{\xi_{mn,d}}$$

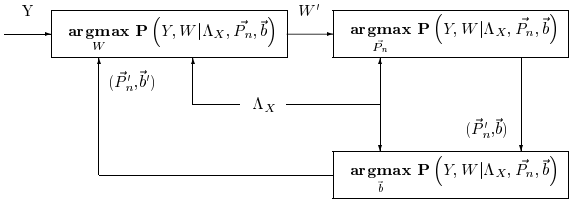

Figure 2: Scheme 1 - for good initial bias guess

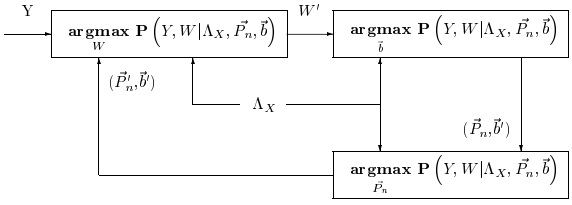

Figure 3: Scheme 2 - for good initial noise guess

The coefficients c_{lk} are given by

$$c_{lk} = \begin{cases} \sqrt{\frac{1}{R}}\cos\frac{\pi(2l-1)(k-1)}{2R}, & 1 \le l \le R,\ k = 1 \\ \sqrt{\frac{2}{R}}\cos\frac{\pi(2l-1)(k-1)}{2R}, & 1 \le l \le R,\ 2 \le k \le R \end{cases}$$

The coefficients $\xi_{mn,d}$ and $\Delta\xi_{mn,d}$ are the variances of d-th static and dynamic MFCC component of mixture m in state n. The probability $p(t,n,m|W,\vec{P}_n,\vec{b},\Lambda_X)$ is the joint likelihood of distorted observation $\vec{Y}$, mixture m in state n with known word sequence W, parameters $\{\vec{P}_n,\vec{b}\}$ and clean speech model Λ_X at time t.

Although it is difficult to find a closed form solution which maximizes the auxiliary function over both $\vec{P}_n$ and $\vec{b}$, a closed form solution for the maximization over $\vec{b}$ for constant $\vec{P}_n$ does exist[8]. Thus, the total computation effort might be reduced by optimizing over $\vec{b}$ and $\vec{P}_n$ separately . Two possible combinations for separating the two optimization steps are depicted in Figures 2 and 3. Scheme 1 shown in Figure 2 optimizes the noise parameters first, while Scheme 2 shown in Figure 3 optimizes the bias parameters first. We expect Scheme 1 to perform better if we have a good initial guesses of the system tilt and Scheme 2 to perform better when we have a good initial guess of the noise parameters.

3. REMARKS ON NOISE PARAMETERS OPTIMIZATION

To avoid negative values for the noise power estimates, we introduce the continuous limiter function $P_n^i = e^{p_n^i}$. Instead of optimizing the auxiliary function over $\vec{P}_n$, we optimize over $\vec{p}_n$. However, this introduces a term $\frac{e^{p_n^i}}{P_o^i + e^{p_n^i}}$ in the calculation of $\frac{\partial Q(\cdot)}{\partial p_n^i}$. This term may cause pre-mature

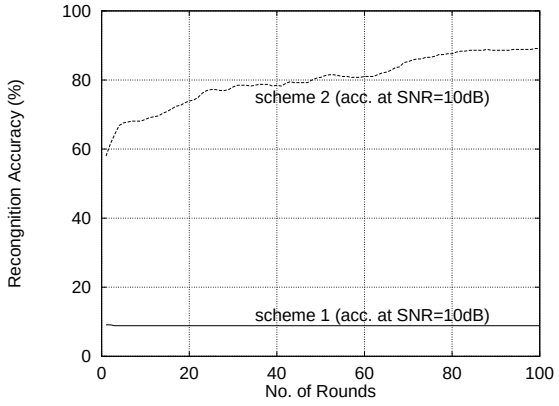

Figure 4: Recognition accuracy of scheme 1 and 2 for the first 100 rounds at SNR=10dB in supervised mode

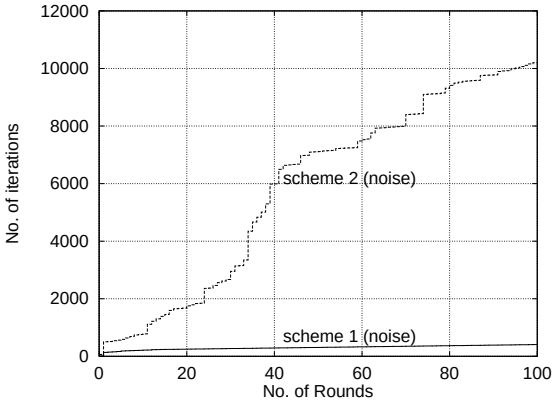

Figure 5: Total number of noise coefficients iterations of scheme 1 and 2 for the first 100 rounds at SNR=10dB in supervised mode

termination of the gradient ascent algorithm if any p_n^i is under-estimated by a large extent, since the gradient becomes almost zero. We refer to this situation as noise parameter dead-lock. We use the following measures during maximization step in EM algorithm to prevent parameter dead-lock:

1. Use initial noise parameters which are over-estimated.

2. Limit the norm of the parameter update to prevent parameter dead-lock due to overshoot. In the tests reported below, we have fixed the ceiling value to $|\epsilon(k) \cdot \nabla Q| < 5$.

4. EXPERIMENTAL RESULTS

4.1. Base-line system

Our baseline system is a speaker dependent connected digit recognizer trained on the *TIMIT Connected Digits Corpus*. The vocabulary consists of 10 digits ('zero' through 'nine') plus 'oh' and silence. Each digit is modeled by a 9 state left-right HMM with 5 mixtures per state. For every frame, a 26-dimensional feature vector is extracted based on C_0 to C_{12} (and delta coefficients) of a 22-nd order MFCC extractor. The analysis frames were 25ms wide with 15ms overlap. The testing set is generated by corrupting speech samples with discrete white Gaussian noise at Signal to Noise Ratio (SNR) of 10dB. Channel distortion is simulated by suppressing 10-th to 22-nd MLFC's by a factor of 100, corresponding to a low pass filter with cut-off frequency 1000Hz. A distorted utterance containing the digit string "8379261" is used for our training token.

4.2. Experiments

In our experiments, we compare the performance of three adaptation algorithms. The first one is a standard gradient ascent algorithm where both the bias and noise parameters are updated simultaneously using a fixed step size ($\epsilon(k) = 0.1$ for all k). The remaining two use schemes 1 and 2 where the bias and noise parameters are optimized separately, but in different order. A contractive variable step size searching algorithm is used in the noise optimization block to ensure that the auxiliary function increases at every iteration. In all cases, the parameters are optimized in supervised mode, where the word sequence is assumed to be known. We choose as initial parameters $\vec{b} = \vec{0}$ and $p_n^i = 10$ for all i. Our previous experiments indicate $p_n^i = 10$ is a good guess for white noise at SNR=10dB.

For schemes 1 and 2, define one "round" to be one trip through the blocks for updating the noise and bias coefficients. In this case, one round may correspond to many noise parameter updates. After each round, recognition accuracy is evaluated over the testing set.

4.3. Results

After separating the optimizations over $\vec{b}$ and $\vec{P}_n$, the number of bias coefficient iterations is negligible compared with that of noise coefficients. The recognition accuracy of schemes 1 and 2 and the corresponding total number of coefficient iterations are shown in Figure 4 and 5.

It is clear that the transformation can compensate for the additive noise and constant channel distortion successfully. Figure 4 shows that the recognition accuracy reaches 89.11% after the 100-th round using Scheme 2. In comparison, recognition accuracy of the baseline system (without adaptation) is 8.9%.

Separating optimization of $\vec{b}$ and $\vec{P}_n$ gives better performance for the same computational effort. The standard gradient ascent approach with fixed step size converges after 4012 iterations at an accuracy of 56.71%. By the 35th round, Scheme 2 has performed a similar number of iterations, but the accuracy reaches 77.64%. Even with half the number of iterations, the accuracy still reaches 76.74%. Most of the gains in the adaptation algorithm are observed in the first several rounds.

The initial value plays an important role in the performance of EM-based optimization process. The ordering of the separate optimization blocks is very critical if only one of initial guesses of $\vec{b}$ and $\vec{P}_n$ is good. This point is illustrated by the performance difference between Scheme 1 and 2. After passing through the first noise parameter optimization block in Scheme 1, the resulting transformation actually decreases performance due to the poor initial guess for the bias vector. We have observed that over half the noise parameters are dead-locked after the first round.

Even if we know the exact value of system bias, better performance is observed if we optimize the bias and noise coefficients as a whole. By fixing $\vec{b}$ to be the simulated constant channel distortion and optimizing over $\{\vec{P}_n\}$, the system recognition accuracy reached 85.57%. However, the maximum recognition accuracy obtained by scheme 2, where both noise and bias coefficients are optimized, is 89.11%. Thus, the constant bias adaptation not only enables the system to compensate for channel distortion, but also to compensate for approximations made in deriving the noise transformation.

5. CONCLUSION

We have presented a mixture mean transformation which can compensate for additive noise and constant channel distortion via ML stochastic matching. For best performance, several rounds of optimization should be done over the noise and bias parameters separately. The ordering of noise and bias optimization blocks is critical if only one of noise and bias coefficients ($\vec{P}_n$ and $\vec{b}$) has a good initial guess. Finally, we found that the constant bias vector can compensate not only for channel distortion, but also for approximations and over-simplifications used in deriving the noise transformation.

6. REFERENCES

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. "A maximum technique occurring in the statistical analysis of probabilistic functions of Markov chains". *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[2] J. R. J. Deller, J. G. Proakis, and J. H. L. Hansen. *"Discrete-time processing of speech signals"*. New York: Macmillan, 1993.

[3] A. P. Dempster, N. M. Larid, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of Royal Statistical Society*, 39(1):1–38, 1977.

[4] M. F. Gales and S. J. Young. "Robust speech recognition in additive and convolutional noise using parellel model cobination". *Computer, Speech and Language*, 9:289–307, 1995.

[5] B. A. Hanson and T. H. Applebaum. "Robust speaker-independent work recognition using static, dynamic and acceleration features: Experiments with lombard and noisy speech". *ICASSP*, pages 857–860, 1990.

[6] B. H. Juang. "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains". *AT&T Technical Journal*, 64(6):1235–1249, July-August 1985.

[7] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. "Integrated models of signal and background with application to speaker identification in noise". *IEEE Transactions of Speech and Audio Processing*, 2(2):245–257, April 1994.

[8] A. Sankar and C. H. Lee. "A maximum-likelihood approach to stochastic matching for robust speech recognition". *IEEE Transactions of Speech and Audio Processing*, 4(3):190–202, May 1996.

[9] S. V. Vaseghi and B. P. Milner. "Noise-adaptive hidden Markov models based on Wiener filters". *EUROSPEECH*, pages 1023–1026, 1993.

[10] S. V. Vaseghi and B. P. Milner. "Noisy speech recognition based on HMMs Wiener filters and re-evaluation of most likely candidates". *ICASSP*, II:103–106, 1993.

[11] S. K. Wong and B. Shi. "A non-linear model transformation for ML stochastic matching in additive noise". In *IEEE Proc. MMSP-98*, Los Angeles, CA, 1998.