MMSE-BASED STEREO FEATURE STOCHASTIC MAPPING FOR NOISE ROBUST SPEECH RECOGNITION

Xiaodong Cui[1]*, Mohamed Afify*[2] *and Yuqing Gao*[1]

IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY, 10598, USA[1]
ITIDA, Ministry of Communications and Information Technology, Cairo, Egypt[2]
Emails: cuix@us.ibm.com, mohamed_afify2001@yahoo.com, yuqing@us.ibm.com

ABSTRACT

A stochastic mapping approach under the MMSE criterion based on stereo features is investigated in this paper for noise robust speech recognition. By learning the mapping from a joint GMM distribution of clean and noisy features, the MMSE estimate of the clean feature is shown to be a piece-wise linear transformation of the noisy feature. The mathematical relationship between the proposed MMSE mapping and other piece-wise linear estimates for noise robustness (i.e. MAP mapping and SPLICE) is also analyzed and discussed. Experimental results show that the proposed MMSE-based stochastic mapping yields superior performance over the MAP mapping on DARPA Transtac large vocabulary spontaneous speech test sets when using clean and multi-style acoustic models.

Index Terms— speech recognition, noise robust, MMSE, stereo feature, stochastic mapping.

1. INTRODUCTION

Noise robustness is crucial when a speech recognition system is deployed in real-world applications. In recent years, the IBM multilingual real-time automatic speech-to-speech translation system [1][2][3] has been targeting its deployment in military conditions through the DARPA Transtac project. The automatic speech recognition component in the speech-to-speech translation system is demanded to be robust to the environment, especially with military noise, to carry the conversation through. Therefore, accomplishing noise robust speech recognition is a fundamental and important issue to the success of the translation system in this scenario.

Stereo data are widely used in achieving noise robustness in speech recognition [4][5][6][7][8][9]. The approaches of using stereo data are able to learn the statistical relationship between clean and noisy speech signals directly from the data for denoising, requiring no model between clean and noisy speech signals. The earliest research was initiated in [4] where the SNR-Dependent Cepstral Normalization (SDCN) and Codeword-Dependent Cepstral Normalization (CDCN) were proposed for noise robust speech recognition based on stereo data that were recorded simultaneously from two channels. In [6] and [7], the Stereo-based Piecewise Linear Compensation for Environments (SPLICE) algorithm was investigated and obtained impressive performance on the Aurora 2 database. Recently, an iterative MAP-based stochastic mapping approach utilizing stereo data was studied in [9] where a GMM distribution is assumed for the joint stereo features and the estimation of the clean feature from the noisy feature was carried out iteratively by the EM algorithm in the GMM framework.

In this paper, we follow the same assumption as [9] in modeling the joint distribution of the stereo features as a GMM. From the information-theoretic perspective, this joint distribution contains all the information between the clean and noisy features. Given the estimated GMM and noisy feature, an MMSE estimate of the clean feature is derived which can be shown as a piece-wise linear function. Since a number of GMM based estimations can result in piecewise linear estimates, it is insightful to investigate the mathematical relationships among the estimates derived from distinctive optimization criteria. In this paper, we analyze the connection of the piecewise linear functions among the proposed MMSE mapping, MAP mapping and SPLICE estimates in [9] and [6], respectively. The MMSE based estimation offers a computational advantage over its MAP counterpart because it requires no EM iterations and no matrix inversion during run-time. In addition, it offers improved performance in our experimental evaluation. We also show that one iteration of the MAP estimate approximately equals to (differ by a posterior probability) the MMSE under a special tying of the parameters. This can partially explain the improved performance of the MMSE estimate in the experiments.

The remainder of the paper is organized as follows. In Section 2, we give the mathematical derivation of the MMSE-based stochastic mapping in the GMM framework of the joint stereo features. In Section 3, we discuss the theoretical relationship among the MMSE, MAP and SPLICE estimates. Experimental results are presented in Section 4 and summary and conclusions are provided in Section 5.

2. MATHEMATICAL FORMULATION

Denote a set of stereo feature as $\{(x_i, y_i)\}$, where x is the clean speech feature vector and y is the corresponding noisy speech feature vector. In the most general case, y_i can be L_n noisy vectors used to predict L_c clean vectors in x_i. Define $z \equiv (x, y)$ as the concatenation of the two channels. A GMM of the joint distribution is shown in Eq.1 and trained from stereo features

$$p(z) = \sum_{k=1}^{K} c_k \mathcal{N}(z; \mu_{z,k}, \Sigma_{zz,k}) \tag{1}$$

where K is the number of mixture components, c_k, $\mu_{z,k}$, and $\Sigma_{zz,k}$, are the mixture weight, mean, and covariance of each component, respectively. Both the mean and covariance can be partitioned as

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \tag{2}$$

$$\Sigma_{zz,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix} \tag{3}$$

where subscripts x and y indicate the clean and noisy speech features respectively.

Given the observed noisy speech feature y, the MMSE estimate of clean speech x is given by

$$\hat{x} = E[x|y] \tag{4}$$

Since $p(x, y)$ is a GMM, Eq.4 can be further written as

$$\begin{aligned} \hat{x} &= \int_x p(x|y)x dx \\ &= \sum_k p(k|y) \int_x p(x|k, y)x dx \\ &= \sum_k p(k|y)E[x|k, y] \end{aligned} \tag{5}$$

The posterior probability term $p(k|y)$ in Eq.5 can be computed as

$$\begin{aligned} p(k|y) &= \frac{p(k, y)}{p(y)} \\ &= \frac{p(y|k)p(k)}{\sum_k p(y|k)p(k)} \end{aligned} \tag{6}$$

The expectation term $E[x|k, y]$ in Eq.5 can be computed as

$$\begin{aligned} E[x|k, y] &= \mu_{x|y,k} \\ &= \mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y - \mu_{y,k}) \end{aligned} \tag{7}$$

From Eq.5 and Eq.7, it is obvious that the MMSE estimate of x is a piece-wise linear function of the noisy feature y, as we can re-write Eq.5 in the following form

$$\hat{x} = \sum_k p(k|y)(A_k y + b_k) \tag{8}$$

where

$$A_k = \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \tag{9}$$

$$b_k = \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \tag{10}$$

In other words, this MMSE-based stochastic mapping is a weighted summation of linear functions contributed by each Gaussian component k from the joint GMM distribution $p(x, y)$. The weight is the posterior probability $p(k|y)$ and the linear function comes naturally as a result of joint Gaussian distribution of each component.

3. COMPARATIVE DISCUSSION

Similar to the proposed MMSE estimate, there are a few other GMM-based techniques for noise robust speech recognition which lead to piece-wise linear functions too, e.g. the SPLICE estimate in [6] and the MAP estimate in [9]. It would be interesting to investigate the relationship between them.

3.1. MMSE vs. SPLICE

In [6], the estimate of clean feature $\hat{x}$ is obtained as

$$\hat{x} = \sum_k p(k|y)(y + r_k) \tag{11}$$

where the bias term r_k of each component is trained upon stereo data (x_n, y_n)

$$r_k = \frac{\sum_n p(k|y_n)(x_n - y_n)}{\sum_n p(k|y_n)} \tag{12}$$

compared to Eqs. 8, 9 and 10, we have several observations. First, SPLICE builds GMM on noisy features while in this paper GMM is built on the joint clean and noisy features (Eq.1). Consequently, the posterior probability $p(k|y)$ in Eq.11 is computed from the noisy feature distribution while $p(k|y)$ in Eq.8 is computed from the joint distribution. Second, SPLICE assumes the transformation matrix A_k is an identity matrix, which is a special case of the MMSE when $\Sigma_{xy,k} = \Sigma_{yy,k}$. If a perfect correlation is assumed between the clean feature x_n and noisy feature y_n, then $p(k|x_n)$ and $p(k|y_n)$ are approximately identical from the joint GMM distribution $p(x, y)$. In this case, Eq.12 can be written as

$$\begin{aligned} r_k &= \frac{\sum_n p(k|y_n)(x_n - y_n)}{\sum_n p(k|y_n)} \\ &= \frac{\sum_n p(k|y_n)x_n - \sum_n p(k|y_n)y_n}{\sum_n p(k|y_n)} \\ &= \frac{\sum_n p(k|y_n)x_n}{\sum_n p(k|y_n)} - \frac{\sum_n p(k|y_n)y_n}{\sum_n p(k|y_n)} \\ &\doteq \frac{\sum_n p(k|x_n)x_n}{\sum_n p(k|x_n)} - \frac{\sum_n p(k|y_n)y_n}{\sum_n p(k|y_n)} \\ &= \mu_{x,k} - \mu_{y,k} \end{aligned} \tag{13}$$

These are A_k and b_k in the MMSE estimate in Eqs. 9 and 10 when $\Sigma_{xy,k} = \Sigma_{yy,k}$.

3.2. MMSE vs. MAP

In [9], a stochastic mapping is estimated under the MAP criterion

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, p(x|y) \tag{14}$$

which results in an iterative piece-wise linear estimate of the clean feature

$$\hat{x}^{(l)} = \sum_k p(k|\hat{x}^{(l-1)}, y)(A_k y + b_k) \tag{15}$$

where $\hat{x}^{(l-1)}$ is the estimate of x from previous iteration and

$$\begin{aligned} A_k &= \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma_{x|y,k}^{-1}\right)^{-1} \cdot \\ &\quad \Sigma_{x|y,k}^{-1}\Sigma_{xy,k}\Sigma_{yy,k}^{-1} \end{aligned} \tag{16}$$

$$\begin{aligned} b_k &= \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma_{x|y,k}^{-1}\right)^{-1} \cdot \\ &\quad \Sigma_{x|y,k}^{-1}\left(\mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k}\right) \end{aligned} \tag{17}$$

By comparing the MAP piece-wise linear estimate in Eqs. 15, 16 and 17 with that of the MMSE estimate in Eqs. 8, 9 and 10, one can easily observe the difference between the two estimates. First, the posterior probability in the MAP estimate is $p(k|\hat{x}^{(l-1)}, y)$, which is computed against the joint Gaussian distribution $p(x, y)$; the posterior probability in the MMSE estimate is $p(k|y)$, which is computed against the marginal noisy distribution $p(y)$ from the joint distribution $p(x, y)$. Second, there is an extra term

$$\left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma^{-1}_{x|y,k}\right)^{-1} \quad (18)$$

in the MAP estimate which is the inversion of the weighted summation of conditional covariance matrices from each individual Gaussian component. This term is the only matrix inversion needed at run-time, otherwise can be computed and saved off-line. It is worth noting that the MMSE estimate has the advantage over the MAP estimate from the computational perspective since it needs no EM iterations and requires no run-time matrix inversion in Eq.18. If we assume the conditional covariance matrix $\Sigma_{x|y,k}$ in Eq.18 is constant across k, i.e. all Gaussians in the GMM share the same conditional covariance matrix $\Sigma_{x|y}$, the Eq.18 turns to

$$\begin{aligned} & \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma^{-1}_{x|y,k}\right)^{-1} \\ = & \left(\Sigma^{-1}_{x|y}\sum_k p(k|\hat{x}^{(l-1)}, y)\right)^{-1} \\ = & \left(\Sigma^{-1}_{x|y}\cdot 1\right)^{-1} = \Sigma_{x|y} \end{aligned} \quad (19)$$

Accordingly, Eqs.16 and 17 can be written as

$$A_k = \Sigma_{x|y}\Sigma^{-1}_{x|y}\Sigma_{xy,k}\Sigma^{-1}_{yy,k} = \Sigma_{xy,k}\Sigma^{-1}_{yy,k} \quad (20)$$

$$\begin{aligned} b_k &= \Sigma_{x|y}\Sigma^{-1}_{x|y}\left(\mu_{x,k} - \Sigma_{xy,k}\Sigma^{-1}_{yy,k}\mu_{y,k}\right) \\ &= \mu_{x,k} - \Sigma_{xy,k}\Sigma^{-1}_{yy,k}\mu_{y,k} \end{aligned} \quad (21)$$

which are literally Eqs. 9 and 10 of the MMSE estimate, respectively. The assumption of $\Sigma_{x|y,k}$ being constant across k can be realized by tying the covariance matrices when estimating joint GMM distribution. This tying can effectively result in more robust estimates of the conditional covariances and can partially explain the superior results obtained using MMSE in the experiments.

4. EXPERIMENTAL RESULTS

Experiments are performed on large vocabulary spontaneous speech recognition system. Both clean and multi-style (MST) acoustic models are trained and tested. There are in total about 120 hours of clean data in the training set. In the clean acoustic model case, the clean acoustic model is trained on clean data. During decoding, the input speech is compensated and decoded by the clean acoustic model. In the MST model case, 15dB and 10dB noisy data are generated by adding humvee, tank and babble noise to the clean data. These three types of noise are chosen to match the military deployment environments in the DARPA Transtac Project. In the MST training, SNR-specific mappings are trained on 15dB and 10dB data separately by stacking the clean speech with the noisy speech. After the mappings are obtained, they are applied back to the noisy training data to yield noise-compensated speech data which are used for multi-style training. In decoding, one mapping is selected for the given environment (SNR) by a GMM-based environment classifier, which will be detailed later, and applied to compensate the incoming speech signal which is eventually decoded by the MST acoustic model. This is in the same spirit of using speaker-adaptive training (SAT) scheme, where some adaptation or compensation method is used in both training and decoding.

The feature space of the acoustic models is formed as follows. First, 24 dimensional Mel-frequency cepstrum coefficients (MFCC) including energy are calculated. The MFCC features are then mean and energy normalized. 9 vectors are stacked leading to a 216-dimensional parameter space. The feature space is finally reduced to 40 dimensions using a combination of linear discriminant analysis (LDA), and maximum likelihood linear transformation (MLLT). This 40-dimensional space is the final space for both training and decoding.

The acoustic model uses Gaussian mixture models associated to the leaves of a decision tree. The tree clustering is done by asking questions about phoneme context. The phoneme inventory has 54 phonemes for American English, and each phoneme is represented by 3 states. After aligning feature vectors to leaves, the GMMs for the leaves are first initialized, and then they are refined by running four iterations of the Forward-Backward algorithm. Rank distributions for each leaf are calculated using the resulting Gaussian mixture models. These discrete rank distributions are used to calculate acoustic scores in the decoding stage. The search uses a stack decoder which employs the rank distributions and trigram language models to find the most likely spoken utterance. The vocabulary has 30K words. The clean acoustic model in the experiments has 43K Gaussians and the MST models have 55K Gaussians.

In terms of noise compensation, a GMM with 1024 Gaussian components is estimated for (clean,15dB) and (clean,10dB) stereo features respectively. In the experiments, we assume the covariance matrices are diagonal. Therefore, the stochastic mapping is scalar and the mapping is performed dimension by dimension. During decoding, a GMM-based environment classifier is built for the detection of clean, 15dB and 10dB SNR environments. In this environment classifier, a GMM with 4 mixture of Gaussians is estimated for each environment using the first 10 frames of the utterances from that environment in the training data. Before decoding, the likelihood of the first 10 frames of the input utterances is computed against each GMM. The environment with the maximum likelihood is chosen as the environment of the input utterance and the mapping of that environment, i.e. clean, 15dB or 10dB, is applied to compensate the utterance.

In [9], the MAP mapping has been shown to obtain superior performance than SPLICE on the Bell-labs CARVUI database. In this paper, we focus on the performance comparison between the proposed MMSE mapping and the MAP mapping in [9]. In the following tables, MAP and MMSE denote the algorithm. SSM24 and SSM40 stand for the feature space to which the mapping is applied - 24 dimensional cepstral space in SSM24 and 40 dimensional space after MLLT in SSM40. The number of EM iterations of the MAP mapping is represented as "1iter" and "3iter" for one and three iterations, respectively.

The experiments are carried out on two test sets both of which are collected in the DARPA Transtac project. The first test set (Set A) has 11 male speakers and 2070 utterances in total recorded in the clean condition. The utterances are spontaneous speech which are corrupted artificially by adding humvee, tank and babble noise to produce 15dB and 10dB noisy test data. Table 1 shows the word error rate of the proposed MMSE mapping when being applied to the 24 dimensional cepstral space and the 40 dimensional MLLT space. Clearly, MMSE mapping in the MLLT space yields better performance which makes sense since the MLLT space is the final feature space where the acoustic models are estimated. This is also consistent with what was observed on MAP mapping in [9]. Therefore, in later experiments, the mappings are compared in the 40 dimensional MLLT space.

Condition	Clean	15 dB	10 dB
no compensation	15.96	31.97	40.72
MMSE-SSM24	14.84	31.21	40.58
MMSE-SSM40	14.70	28.74	35.47

Table 1. Word error rate (WER) with clean acoustic model on Set A when applying MMSE mapping to different domains.

Tables 2 and 3 compare the performance between the MMSE and MAP mappings with clean and MST acoustic models. From the Table 2 with clean acoustic model, the MAP mapping with 3 iterations obtains better performance than 1 iteration and the MMSE mapping gives better performance than the MAP with 3 iterations. When multi-style training is performed, both MAP MST and MMSE MST yield significant better performance compared to MST without noise compensation in 15dB and 10dB. MAP and MMSE deliver comparable WER in this test set with multi-style training. Since the MST model has more training data than the clean model, it has more Gaussians (i.e. 55K vs. 43K). That is the reason for the better performance of the MST model than that of the clean model in the clean condition without compensation in the tables.

Condition	Clean	15 dB	10 dB
no compensation	15.96	31.97	40.72
MAP-SSM40-1iter	14.77	30.63	39.23
MAP-SSM40-3iter	14.77	30.54	39.12
MMSE-SSM40	14.70	28.74	35.47

Table 2. Word error rate (WER) with clean acoustic model on Set A using MAP and MMSE mappings.

Condition	Clean	15 dB	10 dB
no compensation	10.48	20.16	27.15
MAP-SSM40-1iter	11.31	16.63	20.09
MAP-SSM40-3iter	10.96	17.10	20.58
MMSE-SSM40	11.25	16.94	20.24

Table 3. Word error rate (WER) with MST model on Set A using MAP and MMSE mappings.

The MAP and MMSE mappings are evaluated on another test set (Set B) in Table 4 which has 7 male speakers with 203 utterances from each. The utterances were recorded in the real-world environment with humvee and tank noise running in the background. This is a very noisy evaluation set and utterance SNRs are measured around 5dB to 8dB. In this real-world noisy test set, the MMSE mapping achieves 18% relative WER reduction compared to the MAP mappings in the clean model scenario. It also yields around 5% WER reduction when multi-style training is employed.

5. SUMMARY AND CONCLUSIONS

In this paper we investigated an MMSE-based stereo feature stochastic mapping approach for noise compensation. The MMSE mapping, which is estimated based on the GMM joint distribution of the stereo features, is shown to be a piece-wise linear function. We discussed the mathematical connections between the proposed MMSE mapping and other piece-wise linear algorithms known in noise robust speech recognition, i.e. SPLICE and MAP mapping.

model	clean model	MST model
no compensation	59.07	58.58
MAP-SSM40-1iter	56.48	44.67
MAP-SSM40-3iter	56.33	45.46
MMSE-SSM40	46.19	43.02

Table 4. Word error rate (WER) with clean and MST model on Set B using MAP and MMSE mapping.

We conducted experiments on two spontaneous speech test sets to compare the performance of the proposed MMSE mapping to the MAP mapping. From the computational standpoint, the MMSE mapping is advantageous over the MAP mapping since it needs no EM iterations and requires no run-time matrix inversion. Performance-wise, the MMSE mapping scheme also yields superior results when decoding with both clean acoustic model and multi-style trained acoustic model. Especially in the real-world noisy evaluation set, the MMSE mapping scheme yielded significantly better performance.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the DARPA Transtac project.

7. REFERENCES

[1] Y. Gao, B. Zhou, L. Gu, R. Sarikaya, H.-K. Kuo, A.-V.I. Rosti, M. Afify, and W. Zhu, "IBM MASTOR: Multilingual automatic speech-to-speech translator," *Proc. of ICASSP*, pp. 1205–1208, 2006.

[2] B. Zhou, D. Dechelotte, and Y. Gao, "Two-way speech-to-speech translation on handheld devices," *Proc. of ICSLP*, pp. 1637–1640, 2004.

[3] L. Gu, Y. Gao, F. Liu, and M. Picheny, "Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 377–392, 2006.

[4] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.

[5] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large vocabulary continuous speech recognition under adverse acoustic environments," *Proc. of ICSLP*, pp. 806–809, 2000.

[6] J. Droppo, L. Deng, and A. Acero, "Evaluation of the splice algorithm on the aurora 2 database," *Proc. of Eurospeech*, pp. 217–220, 2001.

[7] J. Droppo and A. Acero, "Maximum mutual information splice transform for seen and unseen conditions," *Proc. of Interspeech*, pp. 989–992, 2005.

[8] F. Liu, Y. Gao, L. Gu, and M. Picheny, "Noisy robustness in speech translation," *Proc. of Eurospeech*, pp. 2797–2800, 2003.

[9] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *Proc. of ICASSP*, pp. 377–380, 2007.