STEREO-BASED STOCHASTIC MAPPING WITH DISCRIMINATIVE TRAINING 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]
Orange Lab, Smart Village, Cairo, Egypt[2]
Emails: {cuix, yuqing}@us.ibm.com[1], mohamed_afify2001@yahoo.com[2]

ABSTRACT

This paper presents an enhanced stochastic mapping technique in the discriminative feature (fMPE) space that exploits stereo data for noise robust LVCSR. Both MMSE and MAP estimates of the mapping are given and the performance of the two is investigated. Due to the iterative nature of the MAP estimate, we show that combining MMSE and MAP estimates is possible and yields superior performance than each individual estimate. A multi-style discriminative training with minimum phone error (MPE) criterion is further applied to the compensated features and obtains significant performance improvement on real-world noisy test sets.

Index Terms— Stereo feature, stochastic mapping, discriminative training, noise robustness, automatic speech recognition

1. INTRODUCTION

Stereo-based stochastic mapping (SSM) proposed in our previous work [1][2] has been shown to significantly improve the performance of LVCSR systems in noisy environments. The SSM is a front-end data-driven technique for noise robustness. It assumes a joint Gaussian mixture model (GMM) in the stereo feature space and the mapping between clean and noisy features is estimated from the GMM to compensate the noisy features. In [1] and [2], cepstral and LDA domains were chosen to perform the feature compensation. In this paper, we extend our previous work by introducing the SSM to the discriminative feature space - fMPE space. fMPE [3] expands the traditional speech features, e.g. MFCC or PLP, into a very high dimensional and sparse space based on their posterior probabilities against a large set of Gaussians and projects them back to the original space by discriminatively estimating the projection matrix under the minimum phone error (MPE) criterion. Emerging as a powerful feature space discriminative training approach, fMPE has yielded impressive gains over the traditional features themselves. Therefore, fMPE space is a potentially better stage to apply SSM.

SSM can be estimated under various criteria among which maximum a posteriori (MAP) and minimum mean square error (MMSE) are the two typical choices. In this paper, both the MAP and MMSE estimates are given. The two estimates and fMPE can be considered belonging to a family of piece-wise linear estimators. Their mathematical connections are also discussed. Since the MAP estimate is iterative, it is also possible to combine the two estimates, i.e. MAP and MMSE, which will show to deliver superior performance than each individual estimate alone.

Given the SSM compensated features, they can be either directly decoded by clean acoustic models or used for an environment adaptive multi-style re-training. The re-trained multi-style model is able to capture the acoustic characteristics of the compensated features. It can be further improved by the MPE discriminative training. The performance of all the scenarios will be investigated in the paper on some real-world noisy test sets.

The remainder of the paper is organized as follows. Section 2 provides an overview of noise robust techniques in the front-end feature space. In Section 3, we give the mathematical formulation of SSM under the MAP and MMSE criteria and a discussion of the relationship between the piece-wise linear estimators. Experimental results are presented in Section 4 followed by a summary and discussions in Section 5.

2. NOISE ROBUSTNESS IN FEATURE SPACE

Noise robust techniques can be roughly categorized into two groups depending on whether they are applied in the feature space or model space. Compared to the model space techniques, feature space techniques have a low computational complexity and are easy to decouple from the acoustic model end, which makes them attractive for complicated LVCSR systems.

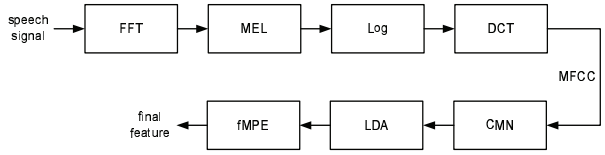

Fig. 1. Pipeline of an exemplary front-end feature computation scheme.

Fig.1 shows an exemplary front-end computation of an IBM LVCSR system with MFCC features. It can be easily extended to other features like PLP. There are multiple stages in such pipelines and the computation evolves through various feature spaces such as linear spectral space (FFT or power spectrum), Mel spectral space, cepstral space and discriminatively trained feature space (LDA and fMPE). Depending on the nature of the algorithm, feature space noise robust techniques apply compensation or enhancement at different feature spaces. For instance, spectral subtraction [6] performs noise suppression in the linear spectral space. The phase-sensitive feature enhancement in [7] is in the log Mel spectral space.

As a data-driven approach that does not rely on explicit model of feature extraction, SSM can be flexibly applied to different feature spaces. In [1] and [2], it was applied in both MFCC and LDA spaces. The results indicate that the LDA space was a better choice

than MFCC since it was closer to the final model space. In [3], the MPE based discriminative training was introduced to the LDA feature space and the resulted discriminatively trained fMPE space has shown to achieve significant improvement over the original LDA space. Therefore, in this paper we extend SSM to the fMPE space for hopefully better performance.

3. SSM AND DISCRIMINATIVE TRAINING

SSM is based on stereo features $\{(x, y)\}$ that are the concatenation of clean speech feature vectors x and noisy speech feature vectors y. In the most general case, y can be L_n noisy vectors used to predict L_c clean vectors in x.

Define $z \equiv (x, y)$ as the joint stereo feature vectors. A GMM is assumed and trained by the EM algorithm on the joint vectors z as shown in Eq.1

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

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. The trained GMM and the noisy features are used to estimate the clean features during testing.

3.1. SSM with fMPE

In this particular work, x and y are obtained by fMPE training on the LDA features according to the feature pipeline in Fig.1. From [3], the fMPE features are computed as Eq.4.

$$\zeta^{\text{fMPE}} = \zeta^{\text{LDA}} + M \cdot h \tag{4}$$

where ζ^{fMPE} and ζ^{LDA} are features in the fMPE and LDA spaces, h a vector in a high dimensional but sparse space consisting of posterior probability against a collection of Gaussians, M a projection matrix estimated under the MPE criterion.

The generation of the vector h uses the enhanced high-dimensional features presented in [8]. The Gaussian clusters is composed of 1024 Gaussians and the posterior probability of the nth Gaussian γ_n is supplemented with the γ_n-scaled offset of the LDA feature ζ^{LDA} from the mean of the nth Gaussian and normalized by the its standard deviation. Two-layered projection for computational efficiency discussed in [8] is also employed.

3.2. MMSE-based SSM

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

$$\hat{x} = E[x|y] \tag{5}$$

The solution of Eq.5 can be written in a piece-wise linear form as Eq.6 [2]

$$\hat{x} = \sum_k p(k|y)(A_k y + b_k) \tag{6}$$

where

$$A_k = \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \tag{7}$$

$$b_k = \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \tag{8}$$

and $p(k|y)$ is the posterior probability against $p(y)$, the marginal noisy speech distribution of the joint stereo distribution $p(x, y)$.

3.3. MAP-based SSM

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

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, p(x|y) \tag{9}$$

From [1], Eq.9 can be solved using the EM algorithm, which results in an iterative estimation process. In each iteration, the estimate can also be written in a piece-wise linear form as Eq.10.

$$\hat{x}^{(l)} = \sum_k p(k|\hat{x}^{(l-1)}, y)(C_k y + d_k) \tag{10}$$

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

$$C_k = \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma_{x|y,k}^{-1}\right)^{-1} \cdot \Sigma_{x|y,k}^{-1}\Sigma_{xy,k}\Sigma_{yy,k}^{-1} \tag{11}$$

$$d_k = \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma_{x|y,k}^{-1}\right)^{-1} \cdot \Sigma_{x|y,k}^{-1}\left(\mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k}\right) \tag{12}$$

and $p(k|\hat{x}^{(l-1)}, y)$ is the posterior probability against the joint stereo distribution.

3.4. Mathematical Connections

It was shown in [2] that the MMSE estimate of SSM is a special tying case of one iteration of the corresponding MAP estimate. In other words, it assumes all Gaussians in the GMM share the same conditional covariance matrix

$$\Sigma_{x|y,k} = \Sigma_{x|y} \tag{13}$$

which is a reasonable result of the "averaging" effect of the expectation function $E[x|y]$ in the MMSE estimate.

Due to the iterative nature of the MAP estimate of SSM, an initial guess has to be made about the clean speech feature $\hat{x}^{(0)}$. A natural choice would be the noisy speech feature y itself, which was used previously in [1] and [2] as a starting point. It is also interesting to combine the two estimates by setting the MMSE estimate as the starting clean feature for the MAP iteration as shown in Eq.14

$$\hat{x}^{(0)} = \hat{x}_{\text{MMSE}} \tag{14}$$

Therefore, the combination leads to a two-stage stochastic mapping strategy which performs the MMSE estimate first followed by the iterative MAP estimate.

A close investigation shows that the MMSE estimate in Eq.6, the MAP estimate in Eq.10, SPLICE and fMPE can be considered under a unified framework of piece-wise linear estimators weighted by posterior probabilities. In [2], it is shown that SPLICE in Eq.15

is a special case of the MMSE estimate of SSM in Eq.6 under the assumption that A_k is an identity matrix which is equivalent to x and y having a perfect correlation. In [5], a connection between SPLICE (Eq.15) and fMPE is discussed where the fMPE formula in Eq.4 is rewritten in the form in Eq.16 with m_k being the kth column vector of the projection matrix M and $p(k|y)$ the posterior probability in h.[1] From Eq.15 and Eq.16, both SPLICE and fMPE share a similar piece-wise linear structure with posterior probability. SPLICE has the bias term r_k estimated under the maximum likelihood criterion while fMPE has m_k estimated under the minimum phone error criterion.

$$\begin{aligned} x &= y + \sum_k p(k|y) r_k \\ &= \sum_k p(k|y)(y + r_k) \end{aligned} \tag{15}$$

$$\begin{aligned} x &= y + M \cdot h \\ &= y + \sum_k p(k|y) m_k \\ &= \sum_k p(k|y)(y + m_k) \end{aligned} \tag{16}$$

Therefore, if we define f_{fMPE}, f_{MMSE} and f_{MAP} as the piece-wise linear functions for fMPE, MMSE and MAP mappings respectively, the overall MAP-based SSM estimation in the fMPE space with the MMSE-based SSM estimate being the starting point can expressed as

$$\hat{x} = f_{\text{MAP}} \circ f_{\text{MMSE}} \circ f_{\text{fMPE}}(y^{\text{LDA}}) \tag{17}$$

where $\circ$ indicates composition. This amounts to applying a sequence of posterior probability weighted piece-wise linear mappings on the noisy LDA features to estimate the clean features.

3.5. Multi-style MPE re-training

After the stochastic mapping, the compensated features can be directly decoded by clean acoustic models. For better performance, an environment adaptive multi-style discriminative re-training can be further applied. In this case, the estimated mapping is applied back to the training data to train a new acoustic model with the SSM-compensated features under the MPE criterion [9] in Eq.18

$$\mathcal{F}_{\text{MPE}}(\lambda) = \sum_r^R \frac{\sum_s p_\lambda(\mathcal{O}_r|s)^\kappa p(s)^\kappa A(s, s_r)}{\sum_s p_\lambda(\mathcal{O}_r|s)^\kappa p(s)^\kappa} \tag{18}$$

where λ are the HMM parameters, $\mathcal{O}_r$ the feature sequence of the rth utterance, κ a probability scale and $p(s)$ the pre-scaled language model probability. It is an average of the "raw phone accuracy" in $A(s, s_r)$ of all possible sentences s, weighted by the sentence posterior probability.

The acoustic model obtained by the multi-style re-training is able to capture the characteristics of the compensated speech features and therefore is supposed to yield better performance than the clean acoustic model.

4. EXPERIMENTAL RESULTS

Experiments are conducted to evaluate the proposed technique on LVCSR tasks.

[1]The symbols in Eq.16 are chosen to be consistent with those in Eq.15 only for comparison purpose. x and y represent features in different spaces in this particular equation and should not be confused with clean and noisy features x and y in the rest of the paper.

The clean training data has 150 hours of speech from which an MPE clean acoustic model with 55k Gaussians and 4.5k states is built. The noisy data are generated by adding a mix of humvee, tank and babble noise to the clean data around 15 dB. These three types of noise are chosen to match the military deployment environments in the DARPA Transtac Project. Thus, there are in total 300 hours of training data in the multi-style training case and the multi-style acoustic model has 90k Gaussians and 5k states.

The feature space of the acoustic models is created as Fig.1. The MFCC features are composed of 24 dimensions. After utterance-based cepstral mean normalization, 9 vectors, including the current vector and its left and right neighbours, are stacked to form a 216-dimensional parameter space. The feature space is then reduced to 40 dimensions using a combination of linear discriminant analysis (LDA) and a global semi-tied covariance (STC) matrix. Finally fMPE training is performed to get the discriminative feature. The fMPE projection matrix is learned from the 150 hour clean data and later on applied to both clean and noisy data.

GMMs are trained on the noisy training data and the mapping is SNR-specific. In test, a GMM-based environment classifier is used to estimate the SNR of the utterances. The environment classifier has two sets of GMMs modeling clean and noisy environments with each having 4 Gaussian components. They are trained on the first 10 frames of each training utterance. The first 10 frames of test utterances are collected and the GMM with the higher likelihood is chosen as the environment of the utterance. SSM compensation is only applied to the noisy environment. The dimension-by-dimension compensation is employed in the experiments where Σ_{xx}, Σ_{yy} and Σ_{xy} are assumed diagonal. Hence, it only involves scalar operations.

A good noise robust technique should be able to improve performance under noisy conditions while maintaining decent performance in clean conditions. To this end, the proposed technique is evaluated on two test sets with continuous speech. Set A consists of 2070 utterances (around 1.7 hours) recorded in clean condition. Set B consists of 1421 utterances (around 1.2 hours) recorded in a real-world noisy condition with humvee noise running in the background. The estimated SNRs of the Set B are about 5-8dB. The test utterances are decoded by a Viterbi decoder on a finite state graph with a trigram language model and a vocabulary of 32k English words.

Table 1 shows the performance tested using the clean MPE acoustic model. In this table, word error rates (WERs) of various SSM estimates are presented where SSM_MAP stands for the MAP estimate starting from the noisy speech feature, SSM_MMSE for the MMSE estimate, SSM_MMSE_MAP for the MAP estimate starting with the MMSE estimate. All the MAP estimations are run for 3 iterations. The numbers in the parentheses are the number of Gaussians in the GMM for the SSM estimation. The baseline is tested without feature compensation. Since the clean acoustic model stays unchanged, SSM gives the same results for Set A after environment detection. As the acoustic model is discriminatively trained on clean speech, the baseline result on Set B noisy data is very poor. But SSM is able to significantly improve the results. From the table, SSM_MMSE_MAP yields the best performance on Set B, better than SSM_MAP and SSM_MMSE alone. Compared to the SSM_MAP, SSM_MMSE_MAP reduces the WER relatively by 50%. SSM with 2048 Gaussians in the GMM gives a slight gain over using 1024 Gaussians.

Table 2 shows the performance of acoustic models by multi-style training. The baseline model is trained with fMPE and MPE on the multi-style data including the 150 hours of clean data and the 150 hours of un-compensated noisy data. Other models are trained on the 150 hours of clean data and the 150 hours of the SSM-compensated

Condition	Set A	Set B
baseline	3.14	61.46
SSM_MAP(1024)	3.14	55.45
SSM_MAP(2048)	3.14	55.22
SSM_MMSE(1024)	3.14	26.41
SSM_MMSE(2048)	3.14	25.39
SSM_MMSE_MAP(1024)	3.14	24.77
SSM_MMSE_MAP(2048)	3.14	24.12

Table 1. Word error rate (WER) of SSM in fMPE space on Sets A and B against clean acoustic model.

noisy data. These acoustic models excluding the baseline are first trained under the ML criterion based on which the MPE training is then applied. The ML and MPE models are distinguished in the parentheses. The GMM in SSM is composed of 1024 Gaussians. The notations for the estimate, i.e. SSM_MAP, SSM_MMSE and SSM_MMSE_MAP, have the same definitions as Table 1. It is observed from the two tables that the baseline (original clean and noisy features without compensation) with multi-style training in Table 2 improves in the noisy condition (Set B) but degrades in the clean condition (Set A) compared to the baseline in Table 1. When using compensated feature for multi-style training, the performance improves for both Set A and Set B. SSM_MMSE_MAP again yields better performance than SSM_MAP and SSM_MMSE. MPE training gives additional gain over ML models on Set B but degrades a little bit on Set A for SSM_MAP and SSM_MMSE. SSM_MMSE_MAP with MPE training gives the overall best results on both Set A and Set B, as shown in the last row of Table 2. It significantly reduces WER in the noisy condition (Set B) while maintaining a decent performance in the clean condition (Set A).

Condition	Set A	Set B
baseline	5.74	27.07
SSM_MAP(ML)	3.24	27.39
SSM_MAP(MPE)	3.56	27.35
SSM_MMSE(ML)	2.87	24.17
SSM_MMSE(MPE)	3.43	23.56
SSM_MMSE_MAP(ML)	3.30	23.66
SSM_MMSE_MAP(MPE)	3.13	22.20

Table 2. Word error rate (WER) of SSM in fMPE space on Sets A and B with multi-style re-training.

5. SUMMARY AND DISCUSSIONS

In this paper, we extend SSM into the discriminatively trained feature space - fMPE space. Both MAP and MMSE estimates of the mapping are investigated. We show that the combination of the two estimates leads to a two-stage feature compensation process. It uses the MMSE estimate as the starting point to perform the iterative MAP estimation. It yields superior performance over individual MMSE or MAP estimate. With the compensated fMPE features, a multi-style MPE training is further applied and shown to get additional performance improvement. The experimental results indicate that the proposed technique significantly improve the performance of LVCSR systems under real-world noisy conditions while giving decent performance in the clean condition.

SSM is a data-driven feature space noise robust technique that exploits stereo data. Hence, it has its advantages and disadvantages. Since it is data-driven and does not rely on model for feature computation, it is quite flexible to apply to various speech features (e.g. MFCC or PLP) and various spaces (e.g. linear or Mel-spectral space, cepstral space, LDA and fMPE spaces, etc). SSM requires stereo data from a particular feature space. However, stereo data is usually expensive to collect. This is one shortcoming of SSM. A suboptimal alternative, as done in this paper, would be to artificially generate data for the noisy channel. This works well for the real-world noisy data in the experiments. Different from the model-based feature compensation techniques which typically estimate noise along with the compensation, SSM as a data-driven approach relies on the noise in the training data and may not handle the unseen noise very well. This is its another disadvantage. Coping with this kind of noise type mismatch problem is the focus of future work and could use well-known adaptation techniques.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Dan Povey and Dr. John Hershey for their comments and suggestions.

7. REFERENCES

[1] M. Afify, X. Cui and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," Proc. of ICASSP, 2007.

[2] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition," Proc. of ICASSP, 2008.

[3] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau and G. Zweig, "fMPE: Discriminatively Trained Features for Speech Recognition," Proc. of ICASSP, 2005

[4] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA 2 database," Proc. of Eurospeech, 2001.

[5] L. Deng, J. Wu, J. Droppo and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," IEEE Signal Processing Letters, Vol. 12, No. 6, Jun 2005.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoustics, Speech, Signal Processing, vol. ASSP-27, No. 2, 1979.

[7] L. Deng, J. Droppo, A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 3, 2004.

[8] D. Povey, "Improvements to fMPE for discriminative training of features," Proc. of Interspeech, 2005.

[9] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," Proc. of ICASSP, 2002.

[10] M. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Trans. on Speech and Audio Processing, Vol. 7, No. 3, 1999.