fMPE-MAP: Improved Discriminative Adaptation for Modeling New Domains

Jing Zheng[1] *Andreas Stolcke*[1,2]

[1] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025 USA
[2] International Computer Science Institute, Berkeley, CA 94704 USA
{zj,stolcke}@speech.sri.com

Abstract

Maximum a posteriori (MAP) adaptation and its discriminative variants, such as MMI-MAP (maximum mutual information MAP) and MPE-MAP (minimum phone error MAP), have been widely applied to acoustic model adaptation. This paper introduces a new adaptation approach, fMPE-MAP, which is an extension to the original fMPE (feature minimum phone error) algorithm, with the enhanced ability in porting Gaussian models and fMPE transforms to a new domain. We applied this approach to the SRI-ICSI 2007 NIST meeting recognition system, for which we ported our conversational telephone speech (CTS) and broadcast news (BN) models to the meeting domain. Experiments showed that the proposed fMPE-MAP approach has comparable or better performance than simply training the fMPE transform on combined data, in addition to the obvious speed advantage. In combination with MPE-MAP, we obtained about 20% relative word error rate reduction on a lecture meeting evaluation test set, over the models trained with the standard MAP approach.

Index Terms: adaptation, MAP, MPE, fMPE, meeting recognition

1. Introduction

Recent research has shown that discriminative training techniques, such as maximum mutual information (MMI) [1] and minimum phone error (MPE) [2], outperform the conventional maximum likelihood estimation (MLE) in large vocabulary speech recognition tasks. It has also been shown that discriminative criteria can be incorporated into the standard maximum a posteriori (MAP) adaptation scheme, leading to MPE-MAP and MMI-MAP [3], which has shown benefits compared to the standard MLE-based MAP in various different applications.

fMPE is a relatively new technique that uses MPE as a criterion to train a linear transform that applies to a very high dimensional conditioning feature to modify standard speech recognition features and to improve recognition accuracy [4]. In the original fMPE approach, the transform and the model parameters are updated iteratively to improve the MPE objective function, where the parameter update is based on the standard MLE approach. It has been shown that fMPE training followed by model-based MPE training can further improve recognition accuracy.

In this paper, we study the question: can we use the fMPE algorithm to adapt a hidden Markov model (HMM) to a new domain? We have practical needs for this kind of adaptation: In the SRI-ICSI meeting recognition system [6][7], we have based on our conversational telephone speech (CTS) and broadcast news (BN) models, since only a relatively small amount of meeting domain data is available. In the past, we have been using MLE-MAP (the standard MAP) and MMI-MAP for domain adaptation. With the advent of the fMPE technique, we also wanted to investigate if fMPE could be used for the same task.

The rest of the paper is organized as follows: Section 2 reviews the original fMPE algorithm; Section 3 describes fMPE-MAP estimation approach; Section 4 shows experimental results; Section 5 discusses other possible variants of fMPE-MAP; Section 6 concludes.

2. fMPE Review

Before introducing fMPE-MAP, let us first review the fMPE technique. We adopted most of the notation from the original paper [4]. The key idea of fMPE is to apply a linear transform M to a high-dimensional conditioning feature vector h_t, to modify the original recognition feature x_t at time t, and obtain the new feature vector y_t that improves the MPE objective function:

$$y_t = x_t + Mh_t \quad (1)$$

With transformation M, the model parameters, mainly the Gaussian parameters, will also be updated following the standard MLE scheme. The fMPE estimation procedure is to find M that maximizes the MPE objective function F_{MPE}, which is defined in [2].

$$M^* = \arg\max_M F_{MPE}(y, \lambda) \quad (2)$$

where y refers to transformed features, and λ to new model parameters, mainly Gaussian means and variances.

Since h_t can be very high dimensional, first order gradient decent optimization is applied to update each matrix element M_{ij} in row i column j:

$$M_{ij} = M_{ij} + \nu_{ij} \frac{\partial F}{\partial M_{ij}} \quad (3)$$

and it is easy to obtain:

$$\frac{\partial F}{\partial M_{ij}} = \sum_{t=1}^{T} \frac{\partial F}{\partial y_{ti}} h_{tj} = \sum_{t=1}^{T} \left[\frac{\partial F}{\partial y_{ti}}^{direct} + \frac{\partial F}{\partial y_{ti}}^{indirect} \right] h_{tj} \quad (4)$$

where y_{ti} refers to the i-th component of the feature vector y_t, h_{tj} to the j-th component of the conditioning vector h_t. The differential dF/dy_{ti} is decomposed into two parts: the direct differential, which is associated with the change of F directly caused by variation of y_{ti}, and the indirect differential with the change of F caused by variation of Gaussian means and variances that are indirectly related to the y_{ti} variation.

More detailed descriptions of the differential computation and parameter update can be found in the original paper [4], and are not repeated here.

3. fMPE-MAP Estimation

Assume we have a model λ_0 trained from a domain with abundant training data. Now we want to port this model to a new domain, with a relatively small amount of training data, which is not sufficient to train an independent model with competitive accuracy. Can we estimate an fMPE transform based on the new data but still respect the original model?

A simple and natural idea is to use λ_0 as a prior model to update model parameters in the standard MAP scheme. This can be easily integrated into an fMPE training procedure: when a new transform is estimated, we run an MLE-MAP update of model parameters, instead of the MLE update in fMPE training. In a diagonal-covariance HMM/GMM system, we can formulate the mean τ and variance τ^2 update as

$$\mu_{smi} = \frac{\theta(y) + \tau\mu_{smi,0}}{\gamma_{sm} + \tau} \tag{5}$$

$$\begin{aligned}\sigma^2_{smi} &= \frac{\theta(y^2_{smi}) + \tau(\mu^2_{smi,0} + \sigma^2_{smi,0})}{\gamma_{sm} + \tau} - \mu^2_{smi} \\ &= \frac{\theta\left[(y_{smi} - \mu_{smi})^2\right] + \tau\sigma^2_{smi,0}}{\gamma_{sm} + \tau} + \frac{\tau}{\gamma_{sm} + \tau}(\mu^2_{smi,0} - \mu^2_{smi}) \\ &\approx \frac{\theta\left[(y_{smi} - \mu_{smi})^2\right] + \tau\sigma^2_{smi,0}}{\gamma_{sm} + \tau}\end{aligned} \tag{6}$$

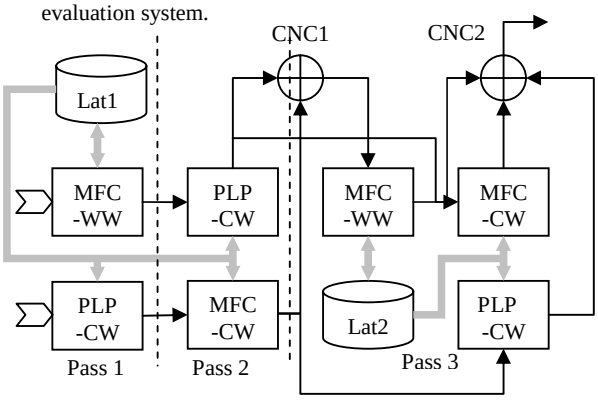
where s, m, and i are state, Gaussian component, and dimension index respectively; Δ_{sm} is the sum of posterior probabilities of the Gaussian component sm over time; $\tau(.)$ represents the function of posterior-probability-weighted summation over time; the subscript ",0" indicates the parameters being from the prior model λ_0; τ is a parameter controlling the rate of MAP update: when τ is 0, it is identical to MLE update; when τ is infinite, the new model will always equal λ_0.

With the MAP update in place of the MLE update, the differential of the MPE objective function will need to change. In fact, the direct differential computation remains the same, as it assumes model parameters being fixed. The indirect differential will need to take into account the factor of τ. Using Equations 5 and 6, we obtain

$$\frac{\partial F}{\partial y_{ti}}^{indirect} = \sum_s \sum_m \frac{\gamma_{sm}(t)}{\gamma_{sm} + \tau}\left[\frac{\partial F}{\partial \mu_{smi}} + 2\frac{\partial F}{\partial \sigma^2_{smi}}(y_{ti} - \mu_{smi})\right] \tag{7}$$

where $\tau_{sm}(t)$ is the posterior probability of m-th Gaussian of state s, and τ_{sm} is the sum $\tau_{sm}(t)$ over time. It is easy to see than when τ equals 0, (7) will degenerate into the standard fMPE indirect differential formula in [4].

Figure 1. Diagram of SRI-ICSI 2005-2006 meeting evaluation system.



In summary, the fMPE-MAP update formula differs from the fMPE update in only two places: first, using Equation 7 to compute the indirect differential of the MPE objective function w.r.t. the transformed feature component; second, using the MAP update (Equations 5, 6) for means and variances instead of the MLE update after the transform update. Other procedures, such as statistics collection and learning rate estimation remain the same. One can easily implement the fMPE-MAP on top of the original fMPE implementation.

In our implementation, we adopt a variant of the MPE objective function, minimum phone frame error (MPFE), which measures frame-level phone accuracy, and showed a slight advantage over the standard MPE in model-based MPE training [5]. As MPFE and the standard MPE differs only in the way of computing phone accuracy, with all other aspects identical, to reduce unnecessary confusion we keep the name fMPE unchanged despite actually using the MPFE criterion. This also applies to the MPE-MAP mentioned in later sections.

4. Experiments

All our experiments were conducted with the SRI-ICSI meeting recognition system as described in [6][7]. Figure 1 shows the system architecture, which is inherited from SRI's Fall 2004 Rich Transcription CTS evaluation system [8]. The system uses two acoustic front ends, MFCC and PLP, and three acoustic models. The MFCC front end generates 64-dimensional features, including 39 dimensions from heteroscedastic linear discriminant analysis (HLDA) [9] transformed MFCC and voicing features [10], and 25 dimensions from KLT transformed MLP posterior features [11][12]. The PLP front end generates 39-dimensional HLDA transformed PLP features. Three sets of acoustic models are trained using the two front ends: MFC-WW is a within-word gender-dependent model; MFC-CW is a cross-word gender-dependent model; PLP-CW is a cross-word gender-independent model. All the models are ported from other domains by adapting to around

200 hours of meeting data. The MFCC models were originally trained for the CTS domain on about 1400 hours of Switchboard and Fisher data; the PLP models were originally trained on about 800 hours of BN and TDT data. All models were trained with the MPFE criterion. Meeting training and test data are downsampled to 8kHz to match the MFCC front end.

The system employs three passes of decoding connected by cross-adaptation, involving two sets of lattice generation and two sets of confusion network combination (CNC) to generate the final output. While the NIST meeting recognition evaluation includes both close-talking and farfield microphone conditions, only results for close-talking microphones are reported here.

4.1. Different Adaptation Approaches

In this experiment, we used the MFC-WW model with multiword bigram decoding to compare different adaptation approaches: MLE-MAP, MPE-MAP, fMPE-MAP, fMPE-MAP+MPE-MAP. The fMPE-MAP used an fMPE transform with about 32M parameters, applied to a 100K-dimensional posterior feature vector generated by a two-layer PLP Gaussian mixture model with five contexts. Results on the NIST 2005 lecture meeting test set are shown in Table 1.

From the result it is clear that there is a serious mismatch between the original CTS models and the test data, and all adaptation approaches significantly improved recognition accuracy. For example, MPE-MAP itself reduced the word error rate (WER) by 8.9% absolute, which is 5.5% more than MLE-MAP. fMPE-MAP alone also reduced WER by 7.9% absolute (20.4% relative). Combing fMPE-MAP and MPE-MAP brings the best performance, 28.6% WER, which represents a 26.3% relative reduction from the original CTS model's 38.8%.

4.2. fMPE-MAP vs. fMPE

In this experiment, we used the PLP-CW model in a trigram lattice rescoring setup, to compare two different strategies: fMPE-MAP on meeting data and fMPE on combined training data. Since the PLP front end generates 39-dimensional features, the fMPE transform has about 19.5M parameters. The posterior features input to the fMPE transform are as in Section 4.1.

We pooled the training data from BN, TDT, and the meeting domain together into a 1000-hour training set, from which we trained an fMPE transform from scratch, using the MLE-MAP adapted old BN as the seed model. After fMPE training, we also ran four iterations of MPFE training on the combined 1000–hour training set. Then, using exactly the same setup, we trained another transform using fMPE-MAP on the 200 hours of meeting data only, followed by four iterations of model-based MPE-MAP training on the same data set. We applied maximum likelihood linear regression (MLLR) adaptation to each of the models and speaker adaptive training (SAT) normalization [13] to the input features, and compared the results of adapted models by rescoring multiword trigram lattices generated with the MFC-WW models. The experiments were performed on two data sets, the lecture and the conference meeting test data from the NIST 2005 meeting evaluation.

Table 1. MFC-WW model results on NIST 2005 lecture set

Models	WER	Rel.Δ
Original CTS model	38.8%	-
MLE-MAP	35.4%	-8.8%
MPE-MAP	29.9%	-22.9%
fMPE-MAP	30.9%	-20.4%
fMPE-MAP+MPE-MAP	28.6%	-26.3%

Table 2: PLP model results (% WER) on a NIST 2005 lecture and conference meeting test sets

	Lecture	Conference
Original BN model	29.4	27.8
MLE-MAP (200 hrs)	27.1	26.8
fMPE (1000 hrs)	26.9	25.7
fMPE-MAP (200 hrs)	26.4	25.4
fMPE + MPFE (1000 hrs)	25.4	25.1
fMPE-MAP+MPE-MAP (200 hrs)	24.8	25.0

Table 3: Full system results (% WER) NIST 2006 eval data

	Lecture	Conference
MLE-MAP	34.1	22.8
MPE-MAP	29.7	22.5
fMPE-MAP	28.7	22.3
fMPE-MAP+MPE-MAP	26.3	22.2

From Table 2 it is clear that both fMPE and fMPE-MAP led to significant improvement on both test sets. Nevertheless, fMPE-MAP and fMPE-MAP+MPE-MAP on 200 hours of in-domain training data gave slightly better results than fMPE and fMPE+MPFE on 1000 hours of combined training data, especially on the lecture meeting test set. One can argue that given appropriate weight to the in-domain portion, fMPE and fMPE+MPE on mixed data should be at least as good as the fMPE-MAP and fMPE-MAP+MPE-MAP. This is of course true, though the training time on the mixed data is about three to four times longer, also the choice of weighting is not obvious without experimentation. Both these issues argue for adaptation versus the full training approach.

It is worth pointing out that on the lecture test set, the fMPE result (26.9%) is only slightly better than the MLE-MAP result (27.1%), though the MLE-MAP model was used as the seed to train the fMPE transform. We think this can be explained by the fact that fMPE training uses the MLE update, which undid some of the effects of the MPFE training that was used to build the original BN model.

4.3. Full System Evaluation

Having compared the different model training techniques on 2005 evaluation data, we applied fMPE-MAP, optionally followed by MPE-MAP, to all the three acoustic models used in the system, and obtained results on 2006 meeting test data. We compared system-level results with baseline systems trained by MLE-MAP and MPE-MAP.

As Table 3 shows, fMPE-MAP with MPE-MAP outperforms MLE-MAP and MPE-MAP, especially on the lectures, giving 22% relative WER reduction. For conference meeting, improvement from discriminative adaptation is generally small, though fMPE-MAP gave the largest improvement.

5. Other fMPE-MAP Variants

The above fMPE-MAP descriptions and experiments used a prior model trained without the fMPE transform. What if the original model had also been trained with fMPE? For example, in gender-dependent model building, we may want to use a gender-independent model with an fMPE transform as the prior for gender dependent model training. Can we refine the gender-independent transform to be gender dependent? The answer is yes. In fact, in the former fMPE-MAP training, we initialize the transform to be zero. In this case, we can simply initialize the fMPE transform with the gender-independent one before applying fMPE-MAP, with everything else remaining the same. We can expect to get gender-dependent fMPE transforms and models in a few iterations. Of course, this initialization approach can also be applied to other adaptation scenarios where the prior model was trained with an fMPE transform in the first place.

6. Conclusions

We have described fMPE-MAP as an extension to the standard fMPE training, with the ability to perform discriminative adaptation. Under this approach, fMPE becomes a special case of fMPE-MAP with τ equal to zero. We applied fMPE-MAP to the NIST meeting recognition task, and obtained significant improvements over a system without fMPE-MAP. We also compared our approach to fMPE on combined in-domain and out-of-domain training. The fMPE-MAP approach yielded comparable or slightly better results, in addition to a substantial reduction in training time.

7. Acknowledgment

We thank our ICSI colleagues for numerous contributions to the meeting recognition system. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

References

[1] P.C. Woodland and D. Povey. “Large Scale MMIE Training for Conversational Telephone Speech Recognition,” *Proc. Speech Transcription Workshop*, College Park, 2000.

[2] D. Povey and P.C. Woodland. “Minimum Phone Error and I-Smoothing for Improved Discriminative Training,” *Proc. ICASSP*, Orlando, 2002.

[3] D. Povey, M.J.F. Gales, D.Y. Kim and P.C. Woodland, "MMI-MAP and MPE-MAP for Acoustic Model Adaptation," in *Proc. Eurospeech*, Geneva, 2003.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proc. ICASSP*, Philadelphia, 2005.

[5] J. Zheng and A. Stolcke, “Improved Discriminative Training Using Phone Lattices,” in *Proc. Eurospeech*, Lisbon, 2005.

[6] A. Janin, A. Stolcke, X. Anguera, K. Boakye, Ö. Cetin, J. Frankel, and J. Zheng, “The ICSI-SRI Spring 2006 Meeting Recognition System,” in *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI* 2006 (S. Renals, S. Bengio, and J. Fiscus, eds.), *Lecture Notes in Computer Science*, Springer, 2007.

[7] A. Stolcke, X. Anguera, K. Boakye, Ö. Cetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, “Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-text Evaluation System," *in Machine Learning for Multimodal Interaction Second International Workshop, MLMI 2005* (S. Renals and S. Bengio, eds.), vol. 3869 of *Lecture Notes in Computer Science*, Springer, 2006.

[8] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sönmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, “Recent Innovations in Speech-to-text Transcription at SRI-ICSI-UW,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1729-1744, September 2006.

[9] Nagendra Kumar, *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, Johns Hopkins University, 1997.

[10] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke, “Voicing Feature Integration in SRI's Decipher LVCSR System,” in *Proc. IEEE ICASSP*, vol. 1, (Montreal), May 2004.

[11] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, “Incorporating Tandem/HATs MLP Features into SRI's Conversational Speech Recognition System,” in *Proc. DARPA Rich Transcription Workshop*, Palisades, NY, 2004.

[12] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, “TRAPping Conversational Speech: Extending TRAP/Tandem Approaches to Conversational Telephone Speech Recognition,” in *Proc. ICASSP*, Montreal, 2004.

[13] M. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” *Computer Speech and Language*, vol. 12, 1998.