DISCRIMINATIVELY TRAINED REGION DEPENDENT FEATURE TRANSFORMS FOR SPEECH RECOGNITION

Bing Zhang[†]*, Spyros Matsoukas, Richard Schwartz*

BBN Technologies
50 Moulton St., Cambridge, MA 02138
{bzhang,smatsouk,schwartz}@bbn.com

ABSTRACT

Discriminatively trained feature transforms such as MPE-HLDA, fMPE and MMI-SPLICE have been shown to be effective in reducing recognition errors in today's state-of-the-art speech recognition systems. This paper introduces the concept of Region Dependent Linear Transform (RDLT), which unifies the above three types of feature transforms and provides a framework for the estimation of piece-wise linear feature projections, based on the Minimum Phoneme Error (MPE) criterion. Recognition results on English conversational telephone speech data show that RDLT offers consistent gains over the baseline systems, which are trained using the LDA+MLLT projection.

1. INTRODUCTION

Discriminative feature optimization has been a research focus in the last few years. Various works have been published, which share the idea of training a feature transform under discriminative criteria such as MPE and MMI. By using discriminative criteria, the feature optimization is better correlated with the reduction of recognition errors, hence offers better accuracy than standard feature transforms like LDA+MLLT [1] or HLDA [2].

In MPE-HLDA [3, 4], a global linear projection is optimized, selecting compact features from concatenated cepstral coefficients across several frames (long span features). Being a linear projection, it offers moderate gains over LDA+MLLT but is quite limited.

Feature space MPE (fMPE) [5, 6] is a nonlinear feature transform also trained using the MPE criterion. A global Gaussian mixture model (GMM) is trained, and the posterior vectors of those Gaussians are projected into a lower dimensional space, in order to correct some predefined features.

The idea of using Gaussian posteriors in fMPE to estimate offsets is not new. A similar algorithm called SPLICE [7] has been used previously for noise compensation, however it is better formulated as a piece-wise linear transform of the original features. A detailed comparison between fMPE and SPLICE is given in [8].

In this work, we combine the idea of piece-wise transforms, as in fMPE and SPLICE, together with the idea of using long span feature projections, as in MPE-HLDA, forming a kind of more general transformation, which is referred to as region dependent transform (RDT) in the rest of the paper. In RDT, either a linear or a nonlinear feature transform can be used in each region. As a special case of RDT, a linear projection of long span features is used for each region. We refer to it as region dependent linear transform (RDLT).

[†] Bing Zhang is a Ph.D. student at College of Computer & Information Science, Northeastern University.

The paper is organized as follows. After the review of MPE-HLDA and fMPE in Section 2, the concept of region dependent transform is formed in Section 3. In Section 4, the implementation is introduced in terms of a feature transform network. Section 5 describes the experimental conditions. In Section 6, results are presented, showing the effect of using linear projections versus offsets, and the effect of using different model size in the feature transform estimation. The paper ends with conclusions and discussion of future work.

2. MPE-HLDA AND FMPE REVIEWED

Both MPE-HLDA and fMPE aim at optimizing the objective function of MPE [9]. The feature transform in MPE-HLDA is a global linear projection:

$$F_{\text{MPE-HLDA}}(o_t) = Ao_t \tag{1}$$

where o_t is long span feature obtained by concatenating several frames of cepstral features from c_{t-k} to c_{t+k}, as showing below:

$$o_t = [c_{t-k}^T, ..., c_{t-1}^T, c_t^T, c_{t+1}^T, ..., c_{t+k}^T]^T \tag{2}$$

In fMPE, the feature transform is

$$F_{\text{fMPE}}(x_t) = x_t + M\gamma_t \tag{3}$$

where M is a projection matrix, and γ_t is an N-dimensional vector of Gaussian posterior probabilities, computed using a GMM trained in the same space of x_t. x_t is some low dimensional feature vector that can be used to train HMM models directly. Typically x_t is formed by linearly projecting o_t. In real fMPE system, posteriors from left and right frames of current frame t are added to γ_t to make the vector even longer, however, for simplicity, such detail is ignored in our analysis.

Eq. (3) can be rewritten as

$$F_{\text{fMPE}}(x_t) = \sum_{i=1}^{N} \gamma_t^{(i)}(x_t + M^{(i)}) \tag{4}$$

by expanding the matrix-vector multiplication. $M^{(i)}$ denotes the ith column of M, and $\gamma_t^{(i)}$ is the ith element of γ_t. For a fixed t, all the $\gamma_t^{(i)}$'s add up to 1.

Eq. (4) shows that fMPE can be viewed as weighted sum of some feature vectors, each obtained by shifting x_t by a constant term. Since the posterior tells which Gaussian the frame is close to, fMPE can be viewed as a region dependent feature correction function. Here the division of regions in the acoustic space is performed by the GMM.

3. REGION DEPENDENT TRANSFORM

It seems at a first glance that two very different functions are used in MPE-HLDA and fMPE, however, if we write Eq. (1) in a similar form as Eq. (4), we have

$$F_{\text{MPE-HLDA}}(o_t) = \sum_{i=1}^{1} \gamma_t^{(1)}(A_i o_t) \tag{5}$$

where there is only one Gaussian, whose posterior $\gamma_t^{(1)}$ is always 1.

From Eq. (4) and (5), it is natural to come up with a generalized form of region dependent linear transform (RDLT) in which both linear projection A_i and bias b_i are specific to region i (or equivalently to Gaussian i).

$$F_{\text{RDLT}}(o_t) = \sum_{i=1}^{N} \gamma_t^{(i)}(A_i o_t + b_i) \tag{6}$$

An expression similar to Eq. (6) can also be found in MMI-SPICE [10], however, in the latter the input feature o_t is a low dimensional feature vector, and A_i is a square matrix. By this means, MMI-SPLICE only applies linear transform of predefined features, while RDLT tries to reselect useful information from longer context.

Conceptually, Eq. (6) can be further generalized by removing the linear constraint, which leads to the general region dependent transform (RDT) as

$$F_{\text{RDT}}(o_t) = \sum_{i=1}^{N} \gamma_t^{(i)} f_i(o_t) \tag{7}$$

where f_i is a vector-to-vector mapping function whose parameters depend on region i. Other specializations of f_i could lead to hierarchical transforms, or general nonlinear transforms. However, in this paper we will only focus on the regionally linear transforms.

Since the linear projections A_i in Eq. (6) have a lot more parameters than bias vectors do, the total number of regions needed for RDLT can be much smaller than in the original fMPE.

As variations of RDLT, one can think of using parameter tying for A_i, since it is possible that fewer distinctive projections are really needed than biases. The tying could be implemented by clustering the Gaussians in the GMM into fewer groups, as we normally do for fast Gaussian computation. Within each group, parameters of the projections could be shared. The parameter tied RDLT (tRDLT) can be expressed as

$$F_{\text{tRDLT}}(o_t) = \sum_{i=1}^{N} \gamma_t^{(i)}(A_{r(i)} o_t + b_i) \tag{8}$$

where $r(i)$ gives the group index of Gaussian i. In one extreme case, if one group is used for all regions, it is equivalent to adding fMPE and MPE-HLDA together.

4. FEATURE TRANSFORM NETWORK

The above analysis shows several variations of region dependent feature transforms. Besides them, context expansion and speaker dependent transforms are usually used in training and decoding. In order to handle the increasing complexity in feature extraction and adaptation, the feature transform network was developed, which has the following features:

- The network is a directed acyclic graph in topology, in which vector-to-vector mapping functions are associated with edges and vertices.
- Two types of functions are supported, temporal or non-temporal. For example, context expansion is a temporal function since it takes several frames of input to produce the output, while linear projection is a non-temporal function.
- The framework provides services such as forward processing, back-propagation of derivatives, memory optimization, caching and integrity checking of the topology and interconnections.
- Nested network is supported for easy combination of multiple transforms.
- In addition, a configuration file is used to specify the network topology and parameters of all feature transforms inside.

Having the feature transform framework simplifies research and development on speech feature processing, as well as the acoustic training in general.

5. EXPERIMENTAL SETUP

5.1. Training and testing corpora

We evaluated the performance of region dependent linear transforms on the 2300-hour EARS RT04 CTS training corpus, consisting of 370 hours of Switchboard and Callhome data, plus 1930 hours of Fisher data. Language Model (LM) training made use of 530M words of web data released by the University of Washington (UW) [11], 141M words from BN data, 47M words of archived text from CNN and PBS, and 2M words from the TDT4 database. Testing was performed on two sets: the RT03 evaluation set (Eval03), consisting of 3 hours of Switchboard-II and 3 hours of Fisher data, and the RT04 development set (Dev04), consisting of 3 hours of Fisher data.

5.2. Baseline System

The baseline system uses a Vocal Tract Length Normalized (VTLN) PLP front-end, computing 14 cepstral coefficients and normalized energy per frame of speech (25 msec window length, 10 msec frame step). Mean and covariance normalization are applied to the cepstra on a conversation side basis, to reduce variability due to the channel/speaker. The actual 60-dimensional features used in acoustic model training are produced by applying LDA+MLLT on sets of 15 contiguous cepstral frames (225 dimensions).

Recognition is carried out in three passes. The first pass is a fast-match search performed in the forward direction, using a bigram language model and a composite within-word triphone State Tied Mixture (STM) HMM. The output of the forward pass consists of the most likely word ends per frame along with their partial forward likelihood scores. This set of choices is used in a subsequent backward pass to restrict the search space, allowing for less expensive decoding with more detailed acoustic and language models. The backward pass is a time-synchronous beam search, employing an approximate trigram LM and within-word quinphone State Clustered Tied Mixture (SCTM) HMMs. The output of the backward pass is a word lattice, which is subsequently rescored in a third pass, using a crossword quinphone SCTM model and an exact trigram LM.

The baseline crossword SCTM model is speaker and gender independent, consisting of approximately 223K tied states, sharing

Gaussian parameters in 7K codebooks. The average number of Gaussians within a codebook was varied to produce three model configurations with 12, 44 and 120 components per codebook.

MPE training of the baseline models is carried out on unigram lattices, generated on the 2300-hour corpus using the ML models. A particular form of MPE training, in which the objective function is smoothed with an MMI prior [12], was found to give optimal results.

5.3. RDLT training procedure

The storage efficient procedure of MPE-HLDA [4] was slightly modified to enable computing derivatives of MPE with respect to any parameters in the feature transform network. Based on the chain rule, the derivative of the MPE objective function with respect to the transformed feature for each frame was computed first. Then it was back-propagated through the feature transform network to get the derivatives of parameters to be updated. The limited memory BFGS [13] algorithm was used for the numerical optimization, typically converging in about 10-14 iterations.

In all RDLT experiments, Gaussian posteriors were obtained using a GMM trained directly from the LDA+MLLT features via unsupervised clustering. The MPE objective function was evaluated on the same unigram lattices used in the baseline MPE training.

The performance of RDLT models was measured on the 9-hour combined Eval03+Dev04 test set, by crossword rescoring of word lattices generated by the baseline system.

6. RESULTS

6.1. Projections vs. Offsets

In order to show the effect of using linear projections versus offsets in region dependent transforms, variations of RDLT are compared to the LDA+MLLT projection in Table 1. For this purpose, only small SCTM crossword HMMs were used, having 12 Gaussians per state (12-GPS) on average.

For convenience, we use the notation $\text{RDLT}_{M,N}$ for the transform that has M linear projections and N offsets. For instance, $\text{RDLT}_{1,0}$ is the MPE-HLDA transform, and $\text{RDLT}_{0,N}$ is the fMPE transform without its context expansion.

Transform	# proj. (M)	# offset (N)	WER (SI-ML)
LDA+MLLT	N/A	N/A	25.9
$\text{RDLT}_{1,0}$	1	0	24.9
$\text{RDLT}_{0,N}$	0	1000	24.6
$\text{RDLT}_{1,N}$	1	1000	24.0
$\text{RDLT}_{M,0}$	1000	0	22.3
$\text{RDLT}_{M,N}$	1000	1000	22.3

Table 1. Variations of RDLT with different number of discriminatively trained projections and offsets. Unadapted decoding results on the Eval03+Dev04 test set.

$\text{RDLT}_{1,N}$ is equivalent to estimating a linear projection of the joint feature set $[o_t^T \ \gamma_t^T]^T$. In practice, the derivatives of the MPE objective function on the two parts can be very different in magnitude, which could lead to bad convergence rate of the numerical optimization algorithm. Scaling of the posteriors was used in our experiments to overcome the problem. In one approach, we used a single scaling factor on the posteriors, which was determined by examining the magnitude of the derivatives. However, this required pre-computing the derivatives. In another approach that did not need the derivative pre-computation, we normalized the joint vector $[o_t^T \ \gamma_t^T]^T$ before the optimization process, to have the same average variance in all dimensions. Both approaches improved the convergence rate of the L-BFGS algorithm.

For RDLTs that had multiple linear projections, we did not observe significant difference between using 0 or 1000 offsets in this comparison. This is reasonable since the number of parameters in 1000 offsets is too small compared to that in the 1000 projections.

We also found that increasing the number of projections even further may not offer significant gains. We have run another experiment with 2500 projections, getting almost the same results as we did with 1000 projections. On the other hand, increasing the number of offsets could possibly offer more gains, given that fMPE was benefited from a large number of offsets.

It would be interesting to obtain more intermediate results in order to find the optimal balance of projections and offsets, but we were not able to investigate alternate configurations due to time constraints.

6.2. Results before and after MPE training of HMM

We estimated two sets of RDLT transforms, one using a small HMM (12-GPS), and another using a medium-sized HMM (44-GPS). In what follows, we refer to these transforms as 12-GPS RDLT and 44-GPS RDLT, respectively. Notice that in both cases, the RDLT had 1000 projections and no offsets. After the feature optimization, larger ML models were retrained using the optimized feature transforms. Finally we ran 6-8 iterations of regular MPE training, on top of these ML models, based on the same unigram lattices that were used in the training of the RDLTs.

Tables 2 and 3 show the results before and after MPE training, where the rows correspond to different feature transforms, and the columns to different final model sizes. In these decoding experiments, only an SCTM crossword HMM was used to rescore trigram lattices generated from the backward pass of the baseline recognition experiment.

Transform	ML Model WER(%)		
	12-GPS	44-GPS	120-GPS
LDA+MLLT	25.9	23.7	22.5
12-GPS RDLT	22.3	22.1	21.9
44-GPS RDLT	-	21.6	$20.8^{\ddagger}$

Table 2. Unadapted Eval03+Dev04 decoding results of ML models with different feature transforms.

The following two observations can be made from the ML results of Table 2:

- A 12-GPS HMM trained on the 12-GPS RDLT features provides a large gain (14% relative) compared to the 12-GPS LDA+MLLT baseline. However, building larger HMMs in the 12-GPS RDLT feature space results in only a small improvement in accuracy (0.4%), while the improvement from increasing the HMM size in the LDA+MLLT feature space is dramatic (3.4% absolute). As a result, the relative gain from the 12-GPS RDLT shrinks down to 2.7%.
- When a 44-GPS model is used in the RDLT estimation, a consistent relative gain of about 8% is obtained compared to the LDA+MLLT baselines for both 44-GPS and 120-GPS final model configurations.

It appears that the performance of RDLT depends heavily on the configuration of the HMM used during the optimization process. If the resolution of this HMM is too low, the discriminative optimization process will concentrate on fixing a lot of errors that do not normally occur in a higher resolution HMM. The RDLT will still provide features with better discriminative properties compared to those of standard LDA+MLLT; but to maximize the gain from this technique, the size of the RDLT HMM should be close to the size of the final HMM used in recognition. A similar observation was reported in [14], where the authors showed that fMPE trained with a medium sized HMM performed better than one trained with a very small HMM.

Transform	**MPE Model WER(%)**		
	12-GPS	**44-GPS**	**120-GPS**
LDA+MLLT	22.1	21.1	20.4
12-GPS RDLT	21.2	20.8	20.4
44-GPS RDLT	-	20.3	19.6‡

Table 3. Unadapted Eval03+Dev04 decoding results of MPE models with different feature transforms.

Similar observations can be drawn from the MPE results of Table 3. In this case, using a larger HMM in the RDLT estimation is critical in order to preserve the gain after retraining of larger models. We can see that the 12-GPS RDLT features offer no improvement in the 120-GPS final MPE HMM. However, using the 44-GPS RDLT gives a 4% relative gain.

It is normal to see a reduced overall gain from the RDLT when using MPE training of the final HMMs, since both the feature optimization and final HMM Gaussian reestimation are done based on the MPE criterion, operating on the same lattices. It would be interesting to see whether there is an additional gain from regenerating the lattices on the training data after the RDLT estimation, to be used in the final MPE training.

Finally, it's worth mentioning that we also reran the experiments marked with "‡" in the tables by recreating decoding lattices using within-word ML STM and SCTM models that were trained with the 44-GPS RDLT features. By using RDLT features in early stages of the decoding, the WERs were further reduced to 20.4% for the ML model and 19.2% for the MPE model, i.e., 9.3% and 5.8% relative WER reductions compared to the ML and MPE baselines, respectively.

7. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we have introduced the concept of region dependent transform by extending MPE-HLDA with the idea of dividing the acoustic space into regions and having a feature transform for each region. Then we discussed the relationship between RDT and other feature extraction techniques, showing that both fMPE and MPE-HLDA are special cases of the region dependent linear transform. We also suggested that the number of projections and biases in this transform could be optimized further by tying the parameters of the projections.

2300 hours of English CTS data were used to train the HMM and the feature transform. We have obtained 9.3% and 5.8% relative WER reductions to the ML and MPE baseline unadapted systems. Training issues especially the effect of HMM size have been analyzed, and the results show that medium sized HMMs should be preferred over small ones in the feature training.

Further research directions may include an investigation of parameter tying schemes, and the integration of RDLT with speaker adaptation.

8. REFERENCES

[1] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. IEEE, June 2000, vol. 2, pp. II1 129–II1 132.

[2] N. Kumar and A. G. Andreou, "A generalization of linear discriminant analysis in maximum likelihood framework," Tech. Rep. JHU-CLSP Technical Report No. 16, Johns Hopkins University, Aug. 1996.

[3] B. Zhang, S. Matsoukas, J. Ma, and R. Schwartz, "Long span features and minimum phoneme error heteroscedastic linear discriminant analysis," In *Proceedings of EARS RT-04 Workshop* [15].

[4] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," In *Proceedings of ICASSP* [16], pp. 925–929.

[5] D. Povey and et al., "fMPE: discriminatively trained features for speech recognition," In *Proceedings of EARS RT-04 Workshop* [15].

[6] D. Povey, "fMPE: discriminatively trained features for speech recognition," In *Proceedings of ICASSP* [16], pp. 961–964.

[7] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proceedings of Interspeech*, Aalborg, Denmark, Sept. 2001, ISCA.

[8] 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.

[9] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of ICASSP*, Orlando, FL, May 2002, IEEE.

[10] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," In *Proceedings of Interspeech* [17].

[11] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. HLT/NAACL*, 2003, pp. 7–9.

[12] D. Povey et al., "EARS Progress Update," presentation in EARS STT meeting, Nov. 2003.

[13] D. C. Liu and J. Nocedal, "On the limited memory BFGS methods for large scale optimization," *Mathematical Programming 45*, pp. 503–528, 1989.

[14] D. Povey, "Improvements to fMPE for discriminative training of features," In *Proceedings of Interspeech* [17].

[15] DARPA, *Proceedings of EARS RT-04 Workshop*, Palisades, NY, Nov. 2004.

[16] IEEE, *Proceedings of ICASSP*, Philadelphia, PA, Mar. 2005.

[17] ISCA, *Proceedings of Interspeech*, Lisbon, Portugal, Sept. 2005.