# COMBINING DISCRIMINATIVE FEATURE, TRANSFORM, AND MODEL TRAINING FOR LARGE VOCABULARY SPEECH RECOGNITION

*Jing Zheng*[1], *Ozgur Cetin*[3], *Mei-Yuh Hwang*[2], *Xin Lei*[2], *Andreas Stolcke*[1,3], *Nelson Morgan*[3]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025 USA
[2]Dept. of Electrical Engineering, University of Washington, Seattle, WA 98195 USA
[3]International Computer Science Institute, Berkeley, CA 94704 USA
Contact Email: zj@speech.sri.com

## ABSTRACT

Recent developments in large vocabulary continuous speech recognition (LVCSR) have shown the effectiveness of discriminative training approaches, employing the following three representative techniques: discriminative Gaussian training using the minimum phone error (MPE) criterion, discriminately trained features estimated by multilayer perceptrons (MLPs); and discriminative feature transforms such as feature-level MPE (fMPE). Although MLP features, MPE models, and fMPE transforms have each been shown to improve recognition accuracy, no previous work has applied all three in a single LVCSR system. This paper uses a state-of-the-art Mandarin recognition system as a platform to study the interaction of all three techniques. Experiments in the broadcast news and broadcast conversation domains show that the contribution of each technique is nonredundant, and that the full combination yields the best performance and has good domain generalization.

***Index Terms***— MLP, MPE, fMPE, Mandarin LVCSR

## 1. INTRODUCTION

In recent years, discriminative acoustic training techniques have led to significant accuracy improvements on many LVCSR tasks. Among these approaches, the following three categories of techniques have seen widespread use because of their proven effectiveness.

*Discriminatively trained models*. This category includes model parameter estimation techniques based on maximum mutual information (MMI) [1,2], minimum phone error (MPE) [3], minimum classification error (MCE) [4], and other relevant criteria. Compared to traditional maximum likelihood (ML) training, discriminative training better addresses the model incorrectness problem, which is a clear theoretical limitation for the hidden Markov model (HMM)-based recognition systems. So far in LVCSR, the MPE criterion has been the most widely used, because of its superior performance and generalization.

*Discriminatively trained features*. State-of-the-art LVCSR systems are mostly based on phoneme units, and therefore rely on phone discrimination. However, commonly used LVCSR features, such as Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) coefficients, are not explicitly optimized for phone discrimination. One approach that has proven effective involved multilayer perceptrons (MLP)s estimating some form of posterior phone probabilities at the frame level, to be used as HMM observation features. This approach is known as Tandem acoustic modeling [5]. MLP features are then typically used in combination with regular features, such as MFCC and PLP, to obtain maximum benefit. The resulting combined feature typically has high dimensionality.

*Discriminatively trained transforms*. The idea of these approaches is to use discriminative criteria to estimate linear feature transforms, which make corrections to standard features to improve discriminative power. Typical approaches include fMPE [6] and MPE-RDLT (region dependent linear transforms) [7]. Compared to MLP features, which are usually appended to regular features, these approaches modify the standard features themselves through transformation based on certain conditions.

These three categories of discriminative training methods attack the speech modeling problem at different levels. Can they work together? Are their contributions nonredundant? In the remainder of the paper, we answer this question by applying MLP features, fMPE transforms, and MPE training to a large vocabulary Mandarin recognition task, and by studying how the methods interact. Section 2 briefly introduces the Mandarin system used as the platform for this study. Section 3 reviews the three approaches, proposes a combined approach, and presents some comparative results using a simple recognition system. Section 4 shows experimental results on a full state-of-the-art Mandarin recognition system. Section 5 summarizes the paper.

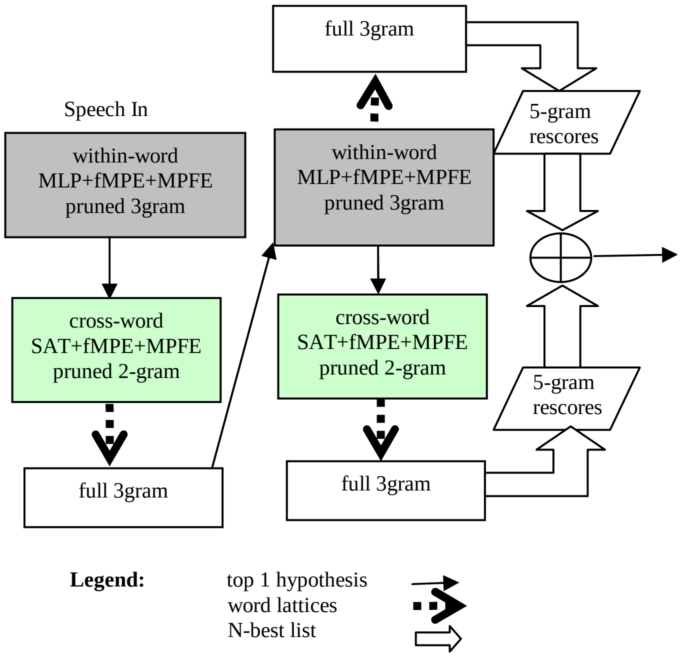

Figure 1: Mandarin LVCSR system illustration.

## 2. THE MANDARIN LVCSR SYSTEM

This section summarizes the Mandarin LVCSR system used as the testbed for this work. More detailed information is available in [16].

Acoustic models were trained on 465 hours of speech, including LDC Mandarin Hub4, Mandarin TDT4 closed captions filtered by flexible alignment [13], and the data released in the first two quarters of the DARPA GALE program. Two sets of models were trained: one was a crossword triphone SAT [14] normalized model based on MFCC+pitch front-end 42-dimensional features, trained using fMPE and MPFE; the second model was similar except that it used MFCC+pitch+MLP 74-dimensional features without SAT normalization, and within-word triphones.

The language model (LM) training corpora comprised 946 million segmented words, including transcriptions of the acoustic training data, text corpora available under the GALE program, and 195M words of Web data. We used a unigram ML-based word segmentation algorithm to segment the training text into multi-character words. Two large 5-gram LMs were developed for N-best rescoring: a pruned word 5-gram LM interpolated with two class-based 5-gram LMs, and an unpruned counts-based word 5-gram LM.

The search structure of the system is depicted in Figure 1. The two acoustic models were each applied twice in four decoding passes in an interleaved way: after the first pass, the remaining three passes performed adapted decoding based on hypotheses generated from the previous pass. The last two passes also generated N-best lists, which were rescored by the 5-gram LMs mentioned earlier. A character-level confusion network combination was then applied to the two sets of N-best lists to produce the final results.

## 3. COMBINED DISCRIMINATIVE MODELING

We first review the three techniques we used in this study: MLP features, MPE models, and fMPE transforms. Then we propose the combined approach and compare results between different configurations.

### 3.1 MLP features

We built our system following the Tandem approach [5], i.e., features generated by the connectionist (MLP) system were used as HMM observation data. This approach produced significant word error rate reduction, about 10% relatively in an English Conversation Telephone Speech system [8], and was shown to generalize across domains and even languages, making it even more attractive [9].

We used two types of MLPs to generate features. A PLP/MLP, which focuses on medium-term information, was trained on 9 consecutive frames of PLP features, as well as their first and second order differences. Hidden activation temporal pattern-MLPs (HATs) [10, 11], which focuses on long-term information, extract information from 500 ms windows of critical band energies. Both PLP/MLP and HATs systems generated phone posteriors, which were combined using inverse-entropy weighting and then projected to 32-dimensional features via a *Karhunen-Loeve* transform.

We appended the 32-d MLP features with the 39-d MFCC features and the 3-d pitch features (consisting of log pitch and its first and second differences) features to form a 74-dimensional feature vector, which was then used to train the HMM system.

### 3.2 MPE models

MPE model training was first proposed in [3]. Gaussian parameters are estimated to optimize the following objective function:

$$F_{MPE}(\lambda) = \sum_{r=1}^{R} \sum_{s} P_k(s \mid O_r, \lambda)\, RawPhoneAccuracy(s) \qquad (1)$$

where $P_k(s \mid O_r, \lambda)$ is the posterior probability of hypothesis *s* for utterance *r* given observation $O_r$, current parameter set $\lambda$, and acoustic scaling factor *k*. *RawPhoneAccuracy*(*s*) is a measure of the number of correctly transcribed phones in hypothesis *s*. So the MPE objective function is the weighted average phone accuracy in the lattices generated by a LVCSR system.

Optimization of the MPE criterion employs an adapted extended Baum-Welch algorithm, which was initially used for MMI training. MPE training aims to minimize phone error rates, which in turn leads to lower word error rates. MPE training is shown to produce more accurate models than MMI training, and also appears to have superior generalization from training to test data, compared with direct word error rate minimization (MWE) [3], and therefore has been adopted widely.

In this study, we used a variant of MPE training, optimizing the weighted average of frame-level phone accuracies of phone lattices generated from a fast decoding system [12]. We had shown that this approach, also known as MPFE, had comparable or better results than the standard MPE in English and Mandarin CTS system, when combined with a MMI model prior and *I*-smoothing [12].

### 3.3 fMPE transforms

fMPE transforms were first proposed in [6]. The idea is to use a very large linear transform that projects a high-dimensional feature vector into the standard feature space, as a correction to the original feature to optimize the MPE objective function of equation (1):

$$y_t = x_t + Mh_t \tag{2}$$

where $x_t$ is the original feature vector at time $t$; $h_t$ is a high-dimensional feature vector; $M$ is the transform matrix that needs to be estimated, and $y_t$ is the corrected feature that has improved discriminative power according to the MPE criterion. fMPE optimization uses first-order gradient descent, with iterative transform estimation and ML model updates. It has been shown that subsequent MPE training after the fMPE transform is fixed can achieve additional word error rate reductions.

Similar to [6], we constructed $h_t$ by computing Gaussian posteriors from evaluating a Gaussian mixture model (GMM), and with neighboring context expansion. The GMM used for this purpose had 102K Gaussians trained on MFCC+pitch features (42 dimensions). We also used a second layer of 3210 Gaussians clustered from the 102K Gaussians for fast computation.

### 3.4. The combined approach

Although both MLP features and fMPE transforms aim at improving the discriminative power of ASR input, they have very different implementations and internal principles, which leave room for effective combination. In this study, we used the 74-d MFCC+pitch+MLP features as input, a 102K-dimensional posterior vector with context width 3, resulting in an fMPE transform $M$ of size 306K x 74. We typically ran 3-4 iterations of fMPE optimization. Then, with the fMPE transform fixed, we ran 4-5 iterations of MPFE model updates. Thus, the system would incorporate all three discriminative training approaches.

Table 1: CERs on dev06bn test set with 1-pass speaker-independent bigram decoding.

| Sysid | MLP | MPFE | fMPE | WER | Rel. Δ |
|---|---|---|---|---|---|
| S0 | | | | 17.1% | - |
| S1 | Yes | | | 15.3% | -10.5% |
| S2 | | Yes | | 14.6% | -14.6% |
| S3 | | | Yes | 15.6% | -8.8% |
| S4 | Yes | Yes | | 13.4% | -21.6% |
| S5 | Yes | | Yes | 14.7% | -14.0% |
| S6 | | Yes | Yes | 13.9% | -18.7% |
| S7 | Yes | Yes | Yes | 13.1% | -23.3% |

To have a comparison with the combined approach, we also trained fMPE transforms without the MLP features (i.e., with the 42-d MFCC+pitch features only). In this setup, because of the lower input feature dimension, we increased the context width to 5 to have an fMPE transform of size closer to the previous one: 510K x 42. After the fMPE transform was learned, we ran MPFE training. We also tested the other 6 possible combinations, such as the MLP features with MPFE training but no fMPE transforms, etc.

Instead of running the full system shown in Figure 1, we ran the comparison with a speaker-independent (SI) single-pass bigram decoder using acoustic models of within-word triphones without SAT. The training data was as described in Section 2. The acoustic model contained about 200K Gaussians. The bigram was the highly pruned bigram described in [16]. Testing was performed on the dev06bn test set, which consists of 3.5 hours of Mandarin broadcast news.

As Table 1 shows, the contributions from the three discriminative training approaches were nonredundant, and the fully combined approach (S7) worked the best among all configurations. Compared to the baseline S0, which used none of the three discriminative training approaches, the CER reduction was quite significant: 4.0% absolute or 23.3% relative.

## 4. FULL SYSTEM RESULTS

We tested the full Mandarin LVCSR system on three different test sets: eval04, which is the one-hour 2004 EARS broadcast news (BN) evaluation test set; dev06gale, which is the GALE portion (1.43 hours) of dev06bn; and dev05bc, which contains 2.7 hours of broadcast conversations (BC) with non-Mandarin speech. For comparison, we also ran a system that shared the same search structure but without using any of the discriminative training approaches, as the baseline. The CER results of these two configurations are compiled in Table 2.

As Table 2 shows, in all cases the system with discriminative training preformed significantly better than

Table 2: CERs of the full system on dev06gale, eval04 and dev06bc.

| | dev06gale | eval04 | dev05bc |
|---|---|---|---|
| Baseline | 6.1% | 13.8% | 28.3% |
| Eval system | 5.2% | 12.2% | 22.5% |
| Rel. Δ (%) | -14.7% | -13.1% | -20.5% |

the baseline system. Nevertheless, in the broadcast news domain, the improvement from discriminative training was considerably smaller compared to the results in Table 1. We believe that this was mainly due to the strong language models employed in the full system, which were tuned on BN-typed text already. The reduced improvement could also come from the fact that adaptation might wash out some of the improvement. In broadcast conversations, we see a bigger improvement (20.5% relatively), similar to those in Table 1, probably because the language models were less well-matched to the test data, while the discriminative training approaches had good generalization across domains.

## 5. CONCLUSIONS

We have presented a discriminative training approach combining MLP features, fMPE transforms, and MPFE training. Experimental results showed that the contributions from these three individual approaches were nonredundant. Applying the combined approach in a state-of-the-art Mandarin LVCSR system led to significant character error rate reductions on three different test sets in different domains. The biggest improvement is in the broadcast conversations domain, which matched training data the least, demonstrating good cross-domain generalization of the proposed method.

## 6. ACKNOWLEDGMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

## 7. REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. “Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition.” *Proc. ICASSP*, Tokyo, 1986.

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

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

[4] E. McDermott, *Discriminative Training for Speech Recognition*, Ph.D. Thesis, Waseda, Japan, 1997.

[5] D.P.W. Ellis, R. Singh, and S. Sivadas, “Tandem Acoustic Modeling in Large-Vocabulary Recognition,” in *Proc. ICASSP*, Salt Lake City, 2001.

[6] 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.

[7] B. Zhang, S. Matsoukas, and R. Schwartz, “Discriminatively Trained Region Dependent Feature Transforms for Speech Recognition,” in *Proc. ICASSP*, Toulouse, 2006.

[8] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, “Using MLP Features in SRI's Conversational Speech Recognition System,” in *Proc. Eurospeech,* Lisbon, 2005.

[9] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and Cross-language Portability of Acoustic Features Estimated by Multilayer Perceptrons," in *Proc. ICASSP*, Toulouse, 2006.

[10] 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.

[11] 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.

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

[13] A. Venkataraman, A. Stolcke, W. Wen, D. Vergyri, V. Gadde, and J. Zheng, “An Efficient Repair Procedure for Quick Transcriptions,” in *Proc. ICSLP*, Jeju Island, Korea, 2004.

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

[15] C.J. Chen, H. Li, L. Shen, G.K. Fu, "Recognize tone languages using pitch information on the main vowel of each syllable", in *Proc. ICASSP*, Salt Lake City, 2001.

[16] M.Y. Hwang, X. Lei, J. Zheng, O. Cetin, W. Wang, G. Peng, A. Stolcke, “Advances on Mandarin broadcast speech recognition,” submitted to *ICASSP* 2007.