# CORRECTED TANDEM FEATURES FOR ACOUSTIC MODEL TRAINING

*Arlo Faria and Nelson Morgan*

International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
{arlo,morgan}@icsi.berkeley.edu

## ABSTRACT

This paper describes a simple method for significantly improving Tandem features used to train acoustic models for large-vocabulary speech recognition. The linear activations at the outputs of an MLP classifier were modified according to known reference labels: where necessary, the activation of the output unit corresponding to the correct phone label was increased in order to make an accurate classification. This technique was inspired by another experiment that determined a lower error bound on ASR performance within the Tandem framework. By simulating an idealized classifier with forward-backward phone posterior probabilities, we observed a best-case scenario in which nearly all errors were eliminated. Although this performance is not practically achievable, the experiment demonstrated the validity of the Tandem processing approach and suggested that considerable gains are possible by improving the MLP phone classifier.

***Index Terms***— speech recognition, feature extraction, multilayer perceptrons, Hidden Markov models

## 1. INTRODUCTION

The predominant and successful framework for automatic speech recognition (ASR) utilizes Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) parameterizing continuous distributions of acoustic features based on a short-term spectral envelope. Tandem acoustic feature extraction [1] was introduced to leverage the discriminative power of a multi-layer perceptron (MLP) classifier, producing an alternative feature representation based on local estimates of phone posterior probabilities. Such MLP-derived features have been used for large-vocabulary ASR [2, 3], complementing other discriminative methods such as MPE parameter estimation [4] and fMPE feature transforms [5].

The feature extraction front-end can be decoupled from sophisticated back-end modeling and decoding, so a system designer can conveniently view Tandem processing as a modular unit to be optimized independently. To this end, we first devised an exploratory experiment in which the MLP was simulated to be at its optimum, providing essentially perfect classification of phonetic speech units. The outputs of the simulated classifier were replaced with forward-backward probabilities by aligning HMM models composed from reference word sequences. Such *idealized* Tandem features allowed for a tremendous gain in ASR performance; slight procedural modifications virtually eliminated all errors.

Of course, while reference word transcriptions form part of the labeled training data, they should not have been available for test data. To avoid cheating we therefore tried using the simulated classifier exclusively during training and applied the normal MLP on test data. Surprisingly, this mismatch did not deteriorate performance but instead provided considerable improvement over the standard Tandem procedure. Inspired by this result, we developed a simple technique for preparing *corrected* Tandem features based on linear output activations of an MLP, and demonstrated significant improvement on a Mandarin broadcast news ASR task.

## 2. MANDARIN BROADCAST NEWS ASR SYSTEM

Our experiments are based on the Mandarin broadcast news ASR system described in [6], simpler than our state-of-the-art implementation [7] developed as a multi-site collaboration for the DARPA GALE project. SRI's DECIPHER recognizer was configured for word-based modeling, although all ASR results are reported as character error rates (CER).

The training set (Mandarin Hub4) comprised 30 hours of television shows, carefully transcribed including speaker labels. Test data are from the DARPA EARS RT-04 evaluation (eval04) and the DARPA GALE 2006 evaluation (eval06).

Automatically-segmented utterances were clustered and assigned pseudo-speaker labels. Standard acoustic features were based on mel-frequency cepstral coefficients, warped with vocal tract length normalization and mean-and-variance normalization applied on a per-speaker basis. Since Mandarin is a tonal language, it was useful to additionally include a smoothed log-pitch estimate [6]. Adding two temporal derivatives resulted in a 42-dimensional acoustic feature vector, which we will simply reference as “MFCC”.

Within-word triphone HMM models were based on a 72-phone inventory comprising consonants and tonal vowels. Parameters were shared across 2000 states clustered with a phonetic decision tree, and observation distributions were mod-

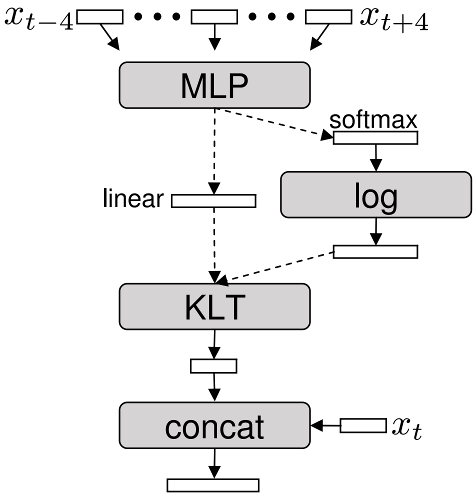

**Fig. 1**. Tandem feature extraction: a multi-layer perceptron estimates phone posterior probabilities, which are transformed for better Gaussian modeling, then concatenated with a standard ASR feature vector to serve as an HMM's acoustic observations. The softmax-logarithm transformation may be omitted, using linear activations at MLP outputs.

eled by a diagonal covariance GMM with 32 mixture components. Viterbi re-alignment of the training data was used for maximum-likelihood parameter estimation.

Recognition networks were compiled from trigram language models trained on over one billion words, with a 60K lexicon [7]. Two decoding passes were separated by 3-class MLLR speaker adaptation, all operating in under 5x real time.

## 3. TANDEM FEATURE EXTRACTION

Figure 1 depicts the general procedure for preparing Tandem acoustic features. This section describes the specific configurations used for experimentation.

The MLP input layer had 378 units, representing 9 context frames of 42-dimensional features similar to those described in the previous section – except based on PLP analysis. Training examples were taken from an 870-hour corpus of television broadcasts (flexibly aligned to closed-captioned transcriptions [8]), mapping HMM states to 71 phone output targets – excluding the *reject* phone. A fully-connected hidden layer of 15,000 units contained nearly 7 milion weights, trained with a quasi-online backpropagation algorithm.

Applying a softmax nonlinearity at the MLP's output layer approximated $P_{\mathrm{mlp}}(Q_t|X_{t\pm4})$: the posterior probability distribution over phones $Q_t$ given the local acoustic evidence $X_{t\pm4}$ centered at the current time $t$ and its 8 neighboring frames of temporal context. Subsequent conversion to the logarithmic domain was intended to better Gaussianize this distribution. The experiments in the Section 4 use the softmax transformation to enable a probabilistic interpretation; however, we have found it is generally better to use the MLP outputs prior to this nonlinearity, as in Section 5. Note that in both cases the MLP was trained using a softmax nonlinearity to determine the cross-entropy error criterion.

**Table 1**. Comparison of Tandem features from two phone classifiers: an idealized simulation and a trained MLP. Character error rate reported on the CCTV subset of eval04.

| Feature | Train | Test | CCTV CER |
|---|---|---|---|
| MFCC | – | – | 11.7 |
| Tandem | MLP | MLP | 9.1 |
| Tandem | idealized | MLP | 8.6 |
| Tandem | idealized | idealized | 4.7 |

Because the HMM-GMM acoustic models operated under an assumption of diagonal covariance, a Karhunen-Loeve Transform (Principal Components Analysis) was applied for orthogonalization and also to rank and reduce the dimensionality to 32. The resulting vector of transformed MLP outputs was then concatenated with the MFCC features described previously, resulting in a 74-dimensional Tandem feature.

Due to practical considerations, the HMM-GMM models used a relatively small training set compared to the MLP training; in our experience, the gains due to MLP features are still consistent – albeit smaller– when the HMM-GMM models are trained on the same amount of data as the MLP.

## 4. IDEALIZED TANDEM FEATURES

Our first experiment sought to determine a lower error bound on ASR performance using Tandem features. Idealized Tandem features were prepared by replacing the MLP outputs (after softmax) with forward-backward phone posterior probabilities to simulate a classifier with "perfect" accuracy.

To simulate this ideal phone classification, we defined the outputs of the hypothetical classifier to be $P_{\mathrm{fb}}(Q_t|X,\hat{W})$: the posterior probability distribution over phones $Q_t$ given the entire acoustic utterance $X$ and its corresponding word transcription $\hat{W}$. This was computed with forward-backward HMM inference, where the model structure was defined by composition of elementary phone models as specified by the word sequence and a pronunciation dictionary. To avoid numerical complications due to artificial zeros in the pruned forward-backward distributions, we lightly interpolated with the MLP-derived probability distribution:

$$P_{\mathrm{ideal}}(Q_t|X,\hat{W}) \doteq P_{\mathrm{fb}}(Q_t|X,\hat{W}) + 0.01P_{\mathrm{mlp}}(Q_t|X_{t\pm4})$$

The MLP-derived distribution was chosen for interpolation to introduce realistic errors rather than arbitrary noise.

The simulation of $P_{\mathrm{ideal}}(Q_t|X,\hat{W})$ for idealized Tandem features required forced alignment to the reference word tran-

**Table 2**. Eliminating the MFCC concatenation and applying a full-rank KLT orthogonalization improved idealized Tandem features in a cheating scenario; however, the opposite effect was observed for MLP-derived features.

| Train & Test | +MFCC | KLT | CCTV CER |
|---|---|---|---|
| MLP | yes | reduced | 9.1 |
| | no | reduced | 9.2 |
| | no | full-rank | 9.7 |
| idealized | yes | reduced | 4.7 |
| | no | reduced | 3.4 |
| | no | full-rank | 1.8 |

scriptions. Due to difficulty in obtaining proper alignments for all of the test data, in this section results are reported only on the relatively easy CCTV subset of eval04. The MLP classifier was able to achieve 79.7% frame-level phone accuracy on this data, scored relative to labels from aligned reference transcriptions. The simulated “perfect” classifier achieved 99.2% accuracy, a less than perfect score due to slight discrepancies between maxima of its forward-backward distributions and the Viterbi-aligned reference labels.

Table 1 summarizes the results of our exploratory experiment. Tandem features provided a gain in ASR performance relative to standard MFCC features. The simulation of an idealized classifier provided a very good result, albeit cheating on the test data. Interestingly, the non-cheating scenario in which idealized features were only used for training data was better than the standard Tandem procedure, despite the expected negative effect due to mismatch of conditions.

In further experiments with idealized Tandem features, we note that it was possible to achieve even better results for the cheating scenario by slightly modifying the Tandem feature extraction process. We eliminated the concatenation step, removing the MFCC components from the Tandem feature vector. Then instead of a dimensionality reduction, we applied a full-rank KLT orthogonalization. Table 2 shows that this greatly decreased the ASR error for idealized Tandem features derived from a simulated perfect classifier; however, performance worsened when using a real MLP classifier.

## 5. CORRECTED TANDEM FEATURES

The experiments of the previous section demonstrated the potential benefit of training acoustic models with idealized Tandem features, for which phone posteriors from an MLP (via softmax at the output layer) were replaced by forward-backward probabilities. However, in our experience we often find it best to use linear activations for the MLP outputs, so it would be desirable to apply an analogous technique in this situation. Yet it is not trivial to convert a forward-backward distribution into a vector of simulated linear activations.

**Table 3**. Comparison of standard and corrected Tandem features derived from an MLP’s linear output activations.

| Feature | Train | Test | eval04 | eval06 |
|---|---|---|---|---|
| MFCC | – | – | 19.2 | 30.6 |
| Tandem | MLP | MLP | 15.5 | 24.2 |
| Tandem | corrected | MLP | 15.1 | 23.9 |

We resolved this with a simple technique for correcting the linear activation outputs of an MLP. Using the Viterbi-aligned reference labels for the training data, we determined for each frame whether the MLP’s classification was correct. If the MLP’s maximal output correctly related to the aligned phone, we left all the outputs unmodified for that frame. If the MLP’s classification was incorrect, we changed the value at the output unit that should have had the maximal activation; we increased it to have the same value as the maximal activation over the other output units. Unlike the preparation for idealized Tandem features, this correction was a relatively minimal modification to the MLP outputs: it was applied only to frames which were incorrectly classified – about 20% of the training set – and affected just one of the MLP output units.

Table 3 shows the experimental results using corrected Tandem features for acoustic model training. For both the eval04 and eval06 test sets, the corrected training features provided modest improvement over the features from unmodified MLP outputs. Over the two sets, the statistical significance of the systems’ difference was verified by a two-tailed MAPSSWE test [9]: $p = 0.015$ (179 vs. 135 unique errors).

## 6. DISCUSSION

### 6.1. Training with corrected Tandem features

The most important result in this work is the observation that an ASR system using Tandem features was significantly improved by applying a small correction to the training features.

The correction procedure is very simple to implement and relies only on having Viterbi-aligned reference labels for the training data; this information is often readily available as it is typically used to prepare the one-hot encoded targets for MLP training. By contrast, an alternative procedure using forward-backward alignments – e.g. for preparing idealized Tandem features – might require a considerable amount of extra computation and storage space. This practicality of the approach provides an easy way to improve existing systems using MLP-derived features; we expect to soon demonstrate results on larger systems and other tasks.

The method might be refined with a principled approach to determine the magnitude of correction. Rather than arbitrarily increasing the correct activation to equal the maximal activation, perhaps a larger increment would be better. However, it is also possible that large corrections could exaggerate the mismatch between the train and test features.

### 6.2. Towards perfect feature extraction

Our cheating experiments with idealized Tandem features demonstrated that such an ASR front-end could reduce the error rate as low as 1.8%. Analyzing this small amount of remaining error, we determined that in half of the cases the automatic utterance segmentation was directly responsible for deletion errors – this problem in our system has since been addressed [10]. We have therefore demonstrated that virtually perfect ASR performance can be achieved with little more than a front-end modification.

To claim that perfect features lead to perfect performance may at first seem obvious, and some researchers have commented that "if you put in the answer at the beginning, of course you'll get it back at the end". However, this is precisely the objective of Tandem feature extraction: a framework for easily exploiting a rich phonetic information stream within the constraints of a very complicated system. That the various manipulations of Tandem processing do not corrupt the idealized input is a validation of the approach.

It is also telling that the standard Tandem procedure had to be modified slightly in order to greatly reduce the error from 4.7% to 1.8% CER. In a general pattern recognition view, the MFCC concatenation should add information and the KLT reduction should remove noise. With idealized Tandem features, however, the added MFCC components were noisy and the truncated KLT dimensions were informative. Though not currently applicable, this suggests that special considerations might need to be examined when designing Tandem systems with extremely accurate classifiers.

Lastly, these experiments might suggest alternative approaches for efficient ASR decoding, considering that the MLP forward pass can be much faster than real-time. With more accurate classifiers, it may be possible to utilize less sophisticated back-end architectures for ASR; in experiments with idealized features, we observed that performance did not degrade even when the GMM models contained fewer mixtures and were trained on less data. Reviewers have suggested another interesting experiment: to decode directly from the idealized posteriors with a hybrid HMM/ANN system [11].

## 7. CONCLUSION

This paper has described a method to improve a large vocabulary speech recognition system using *corrected* Tandem features for acoustic model training. We also demonstrated a hypothetical system using *idealized* Tandem features to determine a bound on ASR performance within this framework, indicating that further front-end improvements have the potential to greatly benefit the overall system.

## 8. ACKNOWLEDGMENTS

Thanks to Andreas Stolcke for permission to use SRI's DECIPHER recognizer. Mei-Yuh Hwang, Gang Peng, and Xin Lei shared UW's expertise with Mandarin and provided the pitch-based features. 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.

## 9. REFERENCES

[1] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proc. ICASSP*, 2000.

[2] Q. Zhu, A. Stolcke, B.Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," *Proc. Interspeech*, 2005.

[3] J. Zheng, O. Cetin, M-Y. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining Discriminative Feature, Transform, and Model Training for Large Vocabulary Speech Recognition," *Proc. ICASSP*, 2007.

[4] D. Povey and PC Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *Proc. ICASSP*, 2002.

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

[6] X. Lei, M. Siu, M-Y. Hwang, M. Ostendorf, and T. Lee, "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition," *Proc. ICSLP*, 2006.

[7] M-Y. Hwang, G. Peng, W. Wang, A. Faria, and A. Heidel, "Building a Highly Accurate Mandarin Speech Recognizer," *Proc. ASRU*, 2007.

[8] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V.R.R. Gadde, and J. Zheng, "An Efficient Repair Procedure For Quick Transcriptions," *Proc. ICSLP*, 2004.

[9] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *Proc. ICASSP*, 1989.

[10] G. Peng, M-Y. Hwang, and M. Ostendorf, "Automatic acoustic segmentation for speech recognition on broadcast recordings," *Proc. Interspeech*, 2007.

[11] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, Kluwer Academic Publishers Boston, 1994.