When a Mismatch Can Be Good:

Large vocabulary speech recognition trained with idealized Tandem features

Arlo Faria[*]
University of California at Berkeley
EECS Department
Berkeley, CA 94720, U.S.A.
arlo@cs.berkeley.edu

Nelson Morgan[†]
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, U.S.A.
morgan@icsi.berkeley.edu

ABSTRACT

This paper explores Tandem feature extraction used in a large-vocabulary speech recognition system. In this framework a multi-layer perceptron estimates phone probabilities which are treated as acoustic observations in a traditional HMM-GMM system. To determine a lower error bound, we simulated an idealized classifier based on alignment of reference transcriptions. This cheating experiment demonstrated a best-case scenario for Tandem feature extraction, highlighting the potential for dramatic system improvement. More importantly, we discovered a way to exploit the result without cheating: using the simulated classifier during training and a MLP classifier at test, the performance improved despite the mismatched Tandem features.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms

Experimentation

Keywords

speech recognition, multi-layer perceptron, Tandem, feature extraction, acoustic modeling

1. INTRODUCTION

Tandem feature extraction [1] was developed to incorporate the flexibility and discriminative power of a multi-layer perceptron (MLP) classifier into the predominant and successful framework for automatic speech recognition (ASR) using Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). Based on inputs comprising several frames of acoustic evidence, a feed-forward network generates outputs that estimate local phone posterior probabilities, which are then treated as feature representations of the acoustic observations. The nonlinear MLP offers several advantages in comparison to typical Gaussian mixture classifiers: it is less restricted by parametric modeling assumptions, straightforwardly exploits temporal correlations, and is discriminatively trained.

Tandem and other MLP features have been successfully applied to large-vocabulary ASR tasks [8]. Their contribution was shown to be similar to other discriminative methods such as MPE training [6] and fMPE feature transforms [5] – as well as complementary in combination [7].

From the viewpoint of a system designer, Tandem feature extraction is a modular unit that can be optimized independently. For example, the MLP phone classifier could be trained on a different data set from that used to train the HMM-GMM models. In general, an MLP trained on more data will achieve higher classification accuracy, providing better Tandem features which improve the overall ASR performance. Extending this observation, we devised a novel experiment in which the MLP was simulated to be at its optimum, providing essentially perfect classification of the desired speech units. This allowed us to investigate the special considerations that are evident when the MLP performs at such a high level.

The outputs of the simulated classifier were defined as probabilities computed via the forward-backward algorithm with HMM models composed from reference transcriptions. Although these word sequences were considered part of the labeled training data, they could not be assumed for unseen test data. Therefore, to avoid cheating we used the simulated classifier exclusively during training and applied the normal MLP at test. Surprisingly, this mismatch did not deteriorate performance but instead provided considerable improvement over the standard Tandem procedure.

2. ASR SYSTEM FOR MANDARIN BN

Collaborating in the DARPA GALE project, researchers from the Univ. of Washington helped us replicate a baseline system for Mandarin broadcast news [3]. SRI's DECIPHER recognizer was configured for word-based modeling (with automatic text segmentation [2]), although all ASR results are reported as character error rates (CER).

[*] Also affiliated with ICSI.

[†] Also affiliated with UC Berkeley.

SAC'08 March 16-20, 2008, Fortaleza, Ceará, Brazil
Copyright 2008 ACM 978-1-59593-753-7/08/0003 ...$5.00.

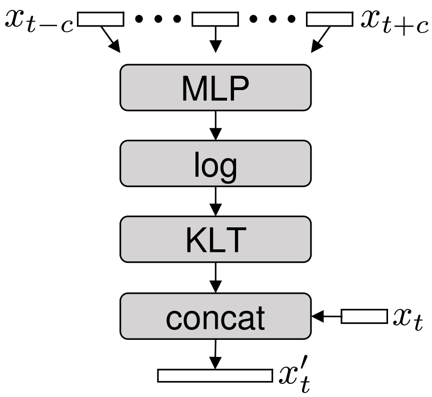

Figure 1: Tandem feature extraction: a multi-layer perceptron estimates phone posterior probabilities, which are transformed for better Gaussian modeling, then concatenated with a standard ASR feature vector to serve as an HMM's acoustic observations.

The Hub4 training set was fairly small: 30 hours of television shows, carefully transcribed including speaker labels. 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 modeled by a diagonal covariance GMM with 32 mixture components. Viterbi re-alignment of the training data was used for maximum-likelihood parameter estimation.

The test set considered in this paper is the CCTV subset of the RT04 evaluation set. Automatically-segmented utterances were clustered and assigned pseudo-speaker labels. The recognition network was compiled from a bigram language model trained on 121M words, with a 49K lexicon. Two decoding passes were separated by 3-class MLLR speaker adaptation, all operating in under 5x real time.

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 [3]. Adding two temporal derivatives resulted in a 42-dimensional acoustic feature, which we will simply reference as "MFCC".

3. TANDEM FEATURE EXTRACTION

Figure 1 depicts the general procedure for preparing Tandem features. This section describes the specific configuration used as a baseline for experimentation.

The input layer of the MLP had 378 units, representing 9 consecutive frames of 42-dimensional acoustic features similar to those described in the previous section – except based on PLP analysis. A quasi-online backpropagation algorithm trained the network's weights from labeled data derived by alignment of the reference transcriptions, as described in Sec. 4.1. Aligned HMM states were mapped to 71 phone output targets, excluding the *reject* phone.

Applying a softmax within the MLP's output layer approximated $P_{\mathrm{mlp}}(q|X)$, the posterior probability of phone q given the local acoustic evidence X. Conversion to the logarithmic domain was intended to better Gaussianize this distribution. A Karhunen-Loeve Transform (Principal Components Analysis) was then applied to orthogonalize and reduce features to 32 dimensions. Lastly, this vector was concatenated with the MFCC features described previously, resulting in a 74-dimensional Tandem feature.

MLP networks were trained on the 30-hour Hub4 set, as well as a larger 870-hr set of closed-captioned television broadcasts. For the Hub4 set, cross-validation accuracies determined the learning rates and early stopping point for training a network with 1150 hidden units. A network of 15,000 hidden units was trained from the larger data set with a fast partitioned approach [8].

4. IDEALIZED TANDEM FEATURES

4.1 Simulating perfect classification

Idealized Tandem features were prepared by replacing the MLP outputs with forward-backward phone posterior probabilities to simulate a "perfect" classifier.

Since the data had not been manually annotated at the phone level, we generated reference labels with a forced alignment technique. A pronunciation dictionary and the baseline MFCC acoustic models were used to compile HMM networks corresponding to the reference word transcriptions. Viterbi decoding determined the most likely sequences of states and corresponding phones. Although these labels might not be as faithful as a human annotation, this procedure is perhaps better for our purposes because it is based on the state and phone sequences required for our ASR systems to achieve optimal recognition performance.

To simulate perfect phone classification, we could have used the Viterbi-aligned labels to generate a sequence of one-hot vectors assigning "hard" probabilities of 1 to the correct phones. Instead, we used phone posterior probabilities to account for the labeling uncertainty at phone boundaries. In these experiments we chose the vectors to be "soft" distributions $P_{\mathrm{fb}}(q|X)$ derived from forward-backward HMM inference, where the model structure was defined by reference word-level transcriptions.

To avoid numerical precision complications due to artificial zeros in the pruned forward-backward distributions, we interpolated with the actual distributions from the MLP:

$$P_{\mathrm{ideal}}(q|X) = 0.99P_{\mathrm{fb}}(q|X) + 0.01P_{\mathrm{mlp}}(q|X)$$

The interpolation could have used arbitrary noise, but we chose $P_{\mathrm{mlp}}(q|X)$ to introduce realistic error distributions.

4.2 Cheating versus mismatched correction

The simulation of a nearly ideal classifier required knowledge of the reference word transcriptions. These labels are generally considered part of the training set, but for test data it would be unrealistic to assume their availability – it would be a cheating experiment. We describe this hypothetical system nonetheless because it is of theoretical interest.

To avoid cheating, we experimented with a system in which training and test features were mismatched: we simulated idealized Tandem features for the training data, but at test time we used the MLP as in the usual application of Tandem feature extraction. This training procedure can be viewed as a correction in which the Tandem features were adjusted to coincide with the desired classification.

Table 1: Frame-level phone classification accuracy and character error rate of Tandem ASR systems using different phone classifiers.

Feature	Train	Test	Acc.	CER
MFCC	–	–	–	11.7
Tandem	30hr MLP	30hr MLP	70.4	10.6
Tandem	870hr MLP	870hr MLP	79.7	9.1
Tandem	idealized	30hr MLP	70.4	10.5
Tandem	idealized	870hr MLP	79.7	8.6
Tandem	idealized	idealized	99.2	4.7

5. EXPERIMENTAL RESULTS

Table 1 summarizes the results of our experiments. Tandem features were computed with various phone classifiers: two MLPs trained on different quantities of data and the simulated idealized classifier. Classification accuracy was measured as the percentage of correct frames, scored against reference labels from aligned test data transcriptions.[1]

Tandem features provided a gain in ASR performance relative to standard MFCC features, even for the small MLP trained on Hub4 data. The more powerful MLP trained on a larger data set classified phones more accurately, which directly translated to a decrease in the overall system's character error rate. Interestingly, the mismatched non-cheating scenarios were better than the standard Tandem procedure, especially with the more accurate large MLP. The simulation of an idealized classifier provided the best result, albeit cheating on the test data.

Table 2 shows the effect of the speaker-based MLLR model adaptation between the first and second decoding passes of the recognizer. When idealized Tandem training features were mismatched with MLP-derived Tandem features for test data, the first-pass hypotheses were worse than for the matched condition. However, the results were reversed for the second-pass hypotheses after model adaptation.

In further experiments with idealized Tandem features used in the cheating scenario, it was possible to achieve even better results by slightly modifying the standard 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 3 shows that this greatly decreased the ASR error for Tandem features derived from the idealized classifier; however, performance worsened when using an MLP classifier.

6. DISCUSSION

6.1 Training with idealized Tandem features

The most important result in this work is the observation that an ASR system using Tandem features was improved by correcting the training features with probabilities from a forward-backward alignment.

Our expectation had been that the mismatched use of an error-prone MLP at test-time would deteriorate performance. As seen in Table 2, this is indeed the case for the first-pass recognition hypotheses. However, it seems that the negative effects of the mismatch are mitigated by model adaptation prior to the second decoding pass. In an attempt to explain these unexpected results, we consider the distribution of features modeled by a diagonal covariance GMM for a given state of an HMM triphone.

Table 2: Speaker-based model adaptation improves the second-pass hypotheses more significantly for the mismatched scenarios.

Train	Test	First-pass	Spkr-adapt
30hr MLP	30hr MLP	11.1	10.6
idealized	30hr MLP	12.3	10.5
870hr MLP	870hr MLP	9.5	9.1
idealized	870hr MLP	9.7	8.6

Table 3: With "perfect" classifiers, Tandem features were improved by eliminating the concatenation with MFCC and applying a full-rank KLT. For real classifiers, the MFCC components and dimensionality reduction seemed to provide robustness.

Classifier type	MFCC concat?	KLT reduce?	CER
30hr MLP	yes	yes	10.6
	no	yes	11.6
	no	no	12.2
870hr MLP	yes	yes	9.1
	no	yes	9.2
	no	no	9.7
"perfect"	yes	yes	4.7
	no	yes	3.4
	no	no	1.8

In the case of a simulated idealized classifier, the distribution over MLP outputs should be unimodally determined by the single phone output which corresponds to the given triphone HMM, allowing the MFCC components to be jointly modeled by all 32 distinct mixtures. Although idealized Tandem features might allow for improved modeling of the MFCC components, the effect would be overshadowed in the likelihood computation because the distributions over the MLP outputs would be learned with extremely low variance. When performing the first decoding pass with these models, performance could be worse because the mismatched MLP outputs contribute too much to the likelihood. Prior to the second decoding pass, model adaptation would increase the variance along these dimensions to allow for the contribution of the well modeled MFCC components.

These results and our interpretation suggest possible improvements to our system's design. To reduce the effect of the mismatch in the first-pass decoding, the acoustic models could be adapted to the MLP-derived Tandem features as a final step of training. As an alternative to idealized Tandem features, a simple solution might have been to increase the number of mixtures in the GMM. In addition, it might be interesting to modify Tandem features for the test data as well as the training data, using posteriors derived from a phone lattice produced in an earlier recognition pass.

[1] Despite cheating, the idealized classifier was not 100% accurate, perhaps due to differences between maxima of the forward-backward distributions and the Viterbi-aligned phone sequence which determined reference scoring labels.

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.[2] 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 reversal indicates special considerations to be examined when designing Tandem systems with extremely accurate classifiers.

Lastly, these experiments were meaningful in suggesting an economical allocation of resources. With more accurate classifiers, it may be possible to utilize less sophisticated back-end architectures for ASR. In further cheating experiments, we observed that similarly excellent performance was possible even when the GMM models contained fewer mixtures and were trained on less data. This could prove to be a great practical advantage of Tandem feature extraction: it may be worthwhile to spend more time and effort to develop and train a better MLP classifier, which would allow one to use simpler ASR models trained on less data.

7. CONCLUSION

This paper has described a method to improve a large vocabulary speech recognition system using idealized Tandem features for acoustic model training. We also demonstrated a hypothetical system to determine a bound on ASR performance within the framework of Tandem feature extraction, 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. David Johnson optimized software and maintained the computing infrastructure for high-performance MLP training at ICSI. 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. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. *Proc. ICASSP*, 2000.

[2] M. Hwang, X. Lei, W. Wang, and T. Shinozaki. Investigation on Mandarin Broadcast News Speech Recognition. *Proc. Interspeech*, 2006.

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

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

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

[7] 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.

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

[2] This problem in our system was recently addressed in [4].