1   Introduction

The traditional statistical formulation of automatic speech recognition (ASR) presents a generative model in which training and decoding operate under the maximum likelihood criterion. Several approaches have introduced more discriminative techniques: model parameter estimation based on alternative criteria (MMI [1], MPE [2, 3]); linear feature transforms (fMPE [4], MPE-HLDA [5]); and posterior-based features (Tandem [6], TRAP [7], Tandem/HAT [8]). This work is only concerned with the last of these, particularly Tandem features produced with a multi-layer perceptron (MLP), although a recent result demonstrates that these three discriminative approaches can be combined to provide complementary improvements [9].

In the predominant Hidden Markov Model paradigm for large-vocabulary ASR, acoustic observation likelihoods are computed from a Gaussian Mixture Model. Due to the assumptions of the HMM-GMM framework, distributions are most accurately modeled for acoustic features which are relatively low-dimensional, somewhat decorrelated, and fairly localized to a single centisecond frame. By contrast, neural network structures have been used to classify inputs which are high-dimensional and temporally correlated by the inclusion of neighboring frames as context. Hybrid connectionist HMM-ANN systems [10, 11] were developed as an alternative to the HMM-GMM, taking advantage of the MLP's model accuracy, context sensitivity, and parsimonious use of parameters. However, incremental enhancements such as speaker adaptation and discriminative parameter estimation were more easily implemented in conventional systems [12], which exhibited much better performance in an evaluation of large-vocabulary broadcast news recognition [13].

A subsequent evaluation, the ETSI Aurora task [14], compared performance of feature extraction front-ends in noisy background conditions. Since all participants were required to use the same HMM-GMM architecture for acoustic modeling, groups with a connectionist background were forced to compromise. Tandem acoustic modeling [15, 16] was thus developed to conveniently enable conventional systems to exploit the advantages of the MLP, by simply treating a MLP-derived phone posterior probability distribution as if it were an ordinary acoustic feature vector. In noisy conditions, these Tandem features were shown to be more robust than standard ASR features. Tandem features and other MLP-derived variants have since been successfully employed in large-vocabulary ASR systems [6, 8, 17]. A prevailing explanation for the positive effect of these features is that the MLP performs a nonlinear feature transformation into a space which is explicitly oriented for discriminability of phones, the underlying structural units of the HMM architecture. Such an interpretation has been investigated in [16, 18], and in similar context by [19].

This paper first describes Tandem acoustic modeling and other related MLP features. The experimental objective is to investigate the factors which influence performance on a Mandarin broadcast news task. In controlled experiments designed to isolate the variables involved, none of the modifications has a particularly large effect although some general trends are noticeable. Guided by these results, the most promising modifications are then selected and applied together to synthesize a significantly improved system. Additionally, a series of cheating experiments simulate what is perhaps the "best-case scenario" for features based on phone posteriors – and which might also be considered an upper-bound on acoustic modeling techniques for ASR. Observations and insights from these experiments are discussed in the final section of this paper, which considers future research possibilities for MLP feature extraction.
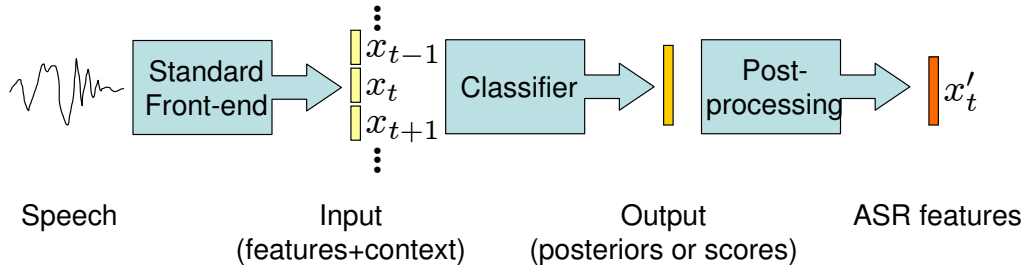
Figure 1: The Tandem acoustic modeling approach.

2   The Tandem approach for producing MLP features

Unfortunately, a confusing nomenclature has previously been used to describe the Tandem acoustic modeling approach, its associated features, and their variants. For clarity of exposition, we will therefore observe the following distinctions:

The Tandem approach to feature extraction is a data processing paradigm in which a speech signal is transformed into a sequence of acoustic observation features with the use of a classifier trained to discriminate speech units.

MLP features are the product of the Tandem approach when the classifier is implemented as a multi-layer perceptron. This term can often be used to describe what are also known as *posterior* or *probabilistic* features, but also includes other types of non-linear transformations.

Tandem features are specifically MLP features produced from a fairly standard configuration, described in Sec. 3.2. Experiments in this paper explore variations on this *Tandem baseline*.

  The central component of Tandem acoustic modeling is a mechanism which is able to classify pre-processed acoustic inputs according to a finite set of categorical units. The outputs of this classifier are then post-processed and presented to a conventional ASR system as if they were acoustic observation features. Figure 1 depicts the general procedure.

  From the vantage point of a system builder, the process can be viewed as a separate feature extraction module that is independent of the modeling architecture, typically a HMM-GMM. Such black-box modularity is generally a desirable characteristic in complex ASR systems. For example, vocal tract length normalization can be implemented simply and efficiently using a generic speech reference model, as described by [20]: this allows the procedure to be entirely contained within the feature extraction module, as opposed to more complex approaches which integrate warp factor estimation into acoustic model training and multi-stage decoding.

2.1   Tandem features

The most common choice for the classifier is a three-layer MLP, a feed-forward neural network structure. The acoustic signal is usually processed by a standard ASR front-end to produce a

centisecond stream of PLP features; however, the input to the network is usually defined as a 9-frame context comprising the current frame plus four consecutive frames preceding and following. Output units of the MLP are determined by the phonetic inventory of a given language; training targets used to learn the network structure are derived from data which has been manually labeled or automatically aligned to reference transcriptions. The network's weight parameters are learned from training data using the back-propagation algorithm with a cross-entropy error function.

Distributions produced by the MLP are vectors in a probability space which is not easily modeled by a GMM; thus, probabilities are approximately Gaussianized by conversion to the logarithmic domain. The features are further processed by applying a Karhunen-Loeve Transform (i.e., Principal Components Analysis), which orthogonalizes the features to satisfy the typical diagonal covariance GMM assumption; additionally, this procedure enables a dimensionality reduction by retaining only the feature components which contribute most to the overall variance of the data. Finally, the resulting vector is concatenated with a standard ASR feature vector, typically MFCC, to serve as higher-dimensional observations for the acoustic models.

2.2 Other MLP features

Whereas Tandem features are derived from a relatively medium-term context of cepstral feature inputs, another common class of MLP features considers an input representation of logarithmic critical band energies (via PLP analysis [21]) over a much longer temporal context, up to a full second. Influenced by Fletcher's experiments suggesting that speech information is independently conveyed in separate critical bands, TRAP processing [7] first attempts to classify phones using each critical band in isolation. Outputs from these Neural TRAP classifiers are then presented as inputs to a merger network, which is able to provide more accurate estimates. A similar hierarchical approach is used for HATs features [22], where the input to the merger network is composed from the hidden layer activations of the independent critical band classifiers.

A practical alternative to HATs is the Tonotopic MLP [23], a four-layer network with a partially connected input structure, allowing joint learning of the critical band and merger network weights. Four-layer network structures and other HATs alternatives were investigated in [24], which also discussed a method for reducing the input's large dimensionality due to the long temporal context. Another possible remedy is explored as the Split Temporal Context approach of [25].

The MLP does not necessarily have to be trained to classify phones. Hierarchical expert networks have been used in [26] to distinguish speech versus non-speech sounds, as well as voicing states. More recently, researchers have investigated the use of articulatory features, using broad phonological classes as MLP targets [27]. It is hoped that a trained network might be language-independent, and could be applied to resource-poor languages. Portability across languages and domains is discussed in [28], and a solution is proposed in the form of a multi-task MLP by [29].

Due to the variety of MLP features, there has also been considerable interest in feature combination strategies. Simple addition of feature vectors can work well, but an inverse entropy-weighted combination [10, 30] is often preferred because of the probabilistic interpretation of the phone posteriors. Recent experiments have also demonstrated that a better combination rule might be possible [31], inspired by the Dempster-Shafer theory of evidence and also appealing to Fletcher's observation of the product-of-errors phenomenon in human speech perception.

The application of the Tandem approach for generating MLP features is not restricted to processing of posterior probability distributions. Bottleneck features [17] can be derived from a very

narrow middle layer of a five-layer MLP; interestingly, the dataflow from the inputs through such a constricted layer does not greatly deteriorate phone discrimination at the output layer. It is hypothesized that the hidden activations of the middle layer project the input to a low-dimensional discriminable space, obviating the need for further dimensionality reduction techniques.

Although most MLP classifiers consider inputs from a temporal context of multiple frames, a speech signal carries a much greater amount of information at the utterance level. An alternative and possibly complementary method for estimating posteriors is to perform the top-down inference afforded by a HMM structure – e.g., using the forward-backward algorithm. A goal of such a research agenda [32] is to allow the feature extraction module to exploit all possible acoustic evidence, including high-level word and phonological constraints encoded in the HMM's topology.

3   Experiments

At the International Computer Science Institute, a recent research effort has focused on Tandem features, HATs features, articulatory features, and their combination for speech recognition of Mandarin and Arabic broadcast news. For the experiments in this paper we consider only Tandem features, primarily for practical reasons: the hierarchical structures of HATs and articulatory systems are implemented relatively inefficiently, making it extremely cumbersome to perform multiple experiments in a short period of time. Moreover, many of the experimental modifications are generally applicable to the Tandem approach, so it is presumed that they will also benefit other MLP features. Experiments are performed on Mandarin broadcast news because we have access to resources for developing a state-of-the-art baseline system for this task, though it is also similar results may reasonably be expected for other domains.

3.1   A baseline system for Mandarin broadcast news

The baseline ASR system is based on the Mandarin broadcast news recognizers developed by the Univ. of Washington, using SRI's DECIPHER system. Thanks to a close collaboration between ICSI, SRI, and UW, it was possible to replicate a system presented in [33, 34].

Systems were trained on a relatively small data set of 29 hours of Hub-4 broadcast news shows, accurately transcribed along with speaker labels. The acoustic features were based on standard 39-dimensional MFCC plus the first two derivatives, prepared with fast GMM-based maximum-likelihood vocal tract length normalization [20] and mean-and-variance normalization applied on a per-speaker basis. Because Mandarin is a tonal language, it has been found that better recognition can be attained using a smoothed log-pitch estimate [33] and its two temporal derivatives, which were appended to the MFCC features to result in a 42-dimensional acoustic feature vector. Acoustic models were based on a 72-unit phoneset comprising consonants and tonal vowels at four levels, with tone 5 mapped to tone 3. Elementary HMM units with a 3-state no-skip Bakis topology were used to represent within-word triphones. Model parameters were tied across 2000 states, clustered using a phonetic decision tree [35]. Each state's observation distribution was modeled by a diagonal covariance GMM with 32 mixture components. Maximum-likelihood parameter estimation was used to train on data which were iteratively re-aligned with the Viterbi algorithm.

The testset considered in this paper is the hour-long RT04 evaluation set (eval04), although many experiments were conducted on a one-third subset of this data (CCTV) which was more similar in style to the training data. The recognition decoder first applied an automatic segmentation

| | CCTV | | eval04 |
Feature type	1st-pass	Spkr-adapt	Spkr-adapt
MFCC	15.5	13.9	24.2
MFCC (UW)	15.5	13.9	24.1
MFCC+F0	13.0	11.7	21.0
MFCC+F0 (UW)	13.0	11.7	21.4
Tandem	11.1	10.6	19.7

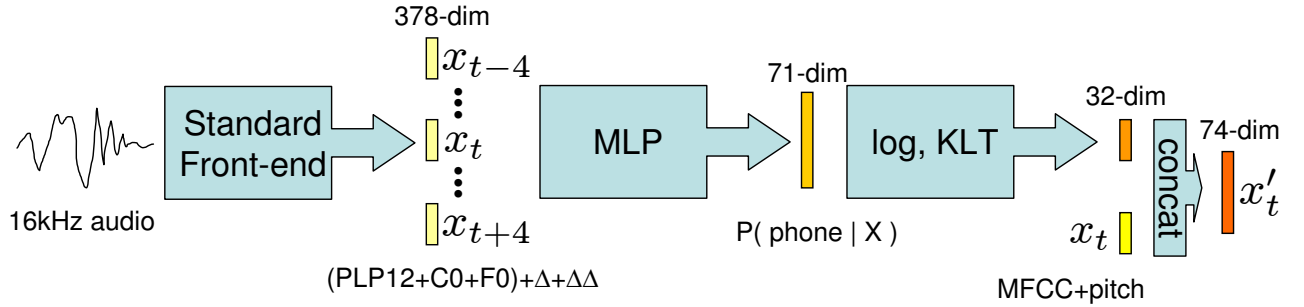Table 1: Performance (CER) of the baseline systems

Figure 2: The baseline configuration for producing Tandem features.

of the test data into 5-10s utterances, which were then assigned pseudo-speaker cluster labels. The recognition network was compiled from a word-level bigram language model trained on 121M words, with a lexicon of 49K words. Two heavily pruned forward passes of the decoder were separated by a round of 3-class MLLR speaker adaptation.

There were a few aspects of the language which required special attention, such as the specially processed pitch features described in [33]. Additionally, an automatic maximum-likelihood segmentation procedure [34] was devised to pre-process the text of the training transcriptions, since Chinese is written without whitespace to delimit word boundaries. While the pronunciation of a string of Chinese characters is to some extent independent of the segmentation, it is nonetheless important to have an accurate word transcription because the ASR system's vocabulary and language modeling operate on a consistent decomposition of word units. It should be noted, however, that the output of Mandarin ASR systems is typically scored against a reference transcription using the metric of character error rate (CER), rather than the usual word error rate.

Table 1 shows the performance of the baseline system, detailing the relative improvements due to modeling pitch features (15%) and performing a second recognition pass after speaker adaptation (10%). Although the baseline system created for these experiments was intended to be identical to [33, 34], there were slight differences – for reasons which are unknown but likely trivial.

5

3.2 A baseline system for Tandem features

The baseline system used for Tandem features is based on a recent setup used at ICSI for the 2007 GALE evaluation. The input to the MLP comprises 378 units, representing 9 frames of PLP and pitch features (and two derivatives). The outputs are 71 phone units – identical to the phoneset used for the baseline HMM-GMM system described previously, minus the "reject" phone. The labeled data used to train the network were derived from a forced-alignment of the reference transcriptions using the baseline MFCC+F0 acoustic models, and mapping triphone states to monophone targets. The network was constructed with 1150 hidden units, or about 500K learnable weight parameters. This setting was determined in reference to [36], with a 20:1 ratio of training data frames to weights. After applying a softmax non-linearity to the activations at the output layer, a logarithm was applied, followed by a KLT orthogonalization and dimensionality reduction to 32 feature dimensions. The resulting vector was then concatenated with the 42-dimensional MFCC+F0 features used in the baseline system described above, resulting in a 74-dimensional Tandem feature. This process is summarized by Figure 2.

Table 1 shows the performance of this baseline Tandem system. While the Tandem features provide an improvement over the standard features, the effect is more dramatic in the first pass than after speaker adaptation, perhaps because the Tandem features are more speaker-independent.

In the following experiments, we will attempt to modify this Tandem baseline system along one dimension at a time. This will serve to guide intuitions about how to improve the system.

3.3 Modifying the input representation

The first set of experiments considers the nature of the input to the MLP. While there are some potentially confounding factors, the experiments can be justified from a practical perspective.

For optimal performance, it would be reasonable to make the input representation as rich as possible, in order to allow the MLP to learn any necessary patterns while ignoring less informative or redundant aspects. The expressiveness of the input is usually reflected in the size of the input layer to the MLP, which in turn affects the number of weight parameters which must be updated during training. Because training time is an overriding practical constraint, for a reasonably fair comparison we can decide to hold the training time constant. When varying the input layer size, the number of hidden units was also adjusted so that the network always maintained about 500K weight parameters, creating a trade-off between the expressiveness of the input versus the discriminative power of the network.

Table 2 displays results from several systems using different input representations. Character error rate is shown for both the eval04 testset and its CCTV subset. For the CCTV data, it was possible to evaluate the frame-level phone accuracy of the MLP classifier on the test data, using labels from the alignment procedure described in Sec. 4. The eval04 data was only used in cases where we wanted to further investigate a promising result from the CCTV data. We display p-values from a two-tailed MAPSSWE significance test [37, 38] relative to the Tandem baseline system; assuming independence of errorful segments separated by two or more correctly recognized characters (rather than words, as usually applied), the MAPSSWE test typically generated sample sizes of over one hundred segments. The two-tailed test was used because it was unclear how to predict the direction of system differences.

It is clear that pitch features are very helpful, since ther removal substantially decreases phone recognition accuracy and increases CER. This may appear obvious given the contributions seen

Feature type	CCTV		eval04	
	CER	Acc.	CER	Signif.
Tandem baseline	10.6	70.4	19.7	–
... w/o F0 inputs	11.2	67.7		
... w/ MFCC inputs	10.9	69.9		
... w/o context	11.2	60.9		
... ± 1 frame	10.7	67.6		
... ± 2 frames	10.2	69.4	19.6	0.78
... ± 3 frames	10.4	70.2	19.6	0.42
... ± 5 frames	11.0	70.6		
... ± 6 frames	10.7	70.6		
... w/o Δ or $\Delta\Delta$	10.3	66.0	19.9	0.94
... w/o $\Delta\Delta$	10.1	69.1	19.7	0.39

Table 2: Effects of varying the represenation of inputs to the MLP. Character error rate is presented in the first and third columns; test set frame-level phone recognition accuracy is given in the second column; p-values from a two-tailed MAPSSWE significance test are in the final column.

in the baseline HMM-GMM system, but it is a recent innovation for Tandem systems; it was also explored in a related context by [39].

It also helps to use PLP features rather than MFCC features for the inputs. In addition to giving providing better phone recognition accuracy, there may also be complementarity provided in the concatenation with the MFCC features at the final step of the Tandem process. Such crossing of MFCC and PLP feature streams is a common characteristic of the SRI system.

A further investigation considered the temporal context of the inputs. Varying the amount of context had a surprisingly minor effect, perhaps because the dependence of the hidden layer size and its discriminative power created an approximately equal and opposite tradeoff. A visual inspection of the input-to-hidden layer connections showed that weights had a greater magnitude for inputs that were centered on the current frame and were closer to zero for farther contexts.

Instead of varying the amount of temporal context, we could also remove the temporal derivative components of the input, since in theory this information can be learned by the network using only the static components. The removal of this redundant information did not seem to greatly affect the overall CER performance much, but it did have a considerable negative effect on the phone recognition accuracy. This phenomenon is not very well understood, but underscores the importance of basing design decisions upon the CER metric – rather than simply observing phone accuracies on cross-validation or development data and extrapolating ASR performance.

Yet another approach to decrease the size of the input is to apply a DCT along the temporal dimension of the input [24]. Due to difficulty of implementation, this was not attempted.

None of these modifications seemed to have significant effect, leading us to conclude that it is probably best to keep the Tandem baseline's 9-frame context and inclusion of temporal derivatives. Modifying the input representation should most likely be viewed as a task-specific design decision in which one must consider the tradeoffs between training time and performance. If training time presents a significant cost, it does not seem unreasonable to use a smaller input representation by decreasing the temporal context and/or removing derivative components.

Feature type	CCTV	eval04	Signif.
Tandem baseline	10.6	19.7	–
... w/ middle-state-only outputs	11.3		
... w/ monophone state outputs	10.5		
... w/ 50 triphone state outputs	11.4		
... w/ 100 triphone state outputs	11.5		
... w/ 200 triphone state outputs	11.0		
... w/ 400 triphone state outputs	10.7		
... w/ phone posterior soft targets	10.7	19.7	0.30
... w/ state posterior soft targets	10.2	19.6	0.34
... w/ monophone state targets (1150 HU)		19.5	0.19
... w/ state posterior soft targets (1150 HU)		19.3	0.17

Table 3: Varying the output representation of the MLP. CER is presented for the CCTV and eval04 testsets, along with a two-tailed MAPSSWE significance test relative to the Tandem baseline.

3.4   Modifying the output representation

We can also vary the representation of the output of the MLP, as summarized in Table 3.

Rather than phone targets, some researchers have experimented using HMM states. An articulatory feature system is presented in [40], where the articulatory target labels were derived only from the middle state of three-state phone HMM unit. While this allows better characterization of a phone-based target, it has the disadvantage of discarding about two-thirds of the available data; and in terms of implementation, this did not result in any significant decrease in training time.

It has also been suggested that the HMM states themselves can be used as targets [41, 42], reflecting the concern that the inner structure of a phone may need to be more carefully discerned. Using monophone states as targets has the unfortunate side-effect of expanding the output layer by a factor of three, so the the hidden layer should be decreased to maintain a constant number of total weight parameters. Despite the reduced discriminative power of the network, the performance does not seem to be negatively affected.

A similar situation arises when using triphones or triphone states as targets, for which the output dimensionality is cubically enlarged. As a tractable approximation, we thus used aggressively clustered triphone states. It was not possible to achieve a promising result with these outputs, suggesting perhaps a better solution is needed for working with such large output layers.

It is not clear that Viterbi-aligned one-hot unary "hard" targets are the best way to represent training labels for the data. Alternatively, state or phone posteriors from forward-backward HMM alignments can serve as "soft" targets. This allows for a representation of the uncertainty of labels near unit boundaries, although in practice most of the posterior distributions have very low entropy. Soft targets seem to be more helpful for discrimination of outputs represented as HMM states rather than phones, perhaps because there is more uncertainty in the labels.

In the results presented thus far, we have kept the total number of parameter weights constant, as we did when varying the input layer. However, the situation is subtly different from before since one cannot argue that more output units should provide better performance given an unlimited number of hidden units. Because large output layers are not inherently favored, it may not be fair

Feature type	CTV	eval04	Signif.
Tandem baseline	10.6	19.7	–
... w/ linear output layer	10.2	19.7	0.15
... w/o concat	11.6		
... w/ LDA	10.9		
... w/ Δ before KLT	13.5		
... w/ Δ after KLT (no concat)	12.5		

Table 4: Evaluation of several post-processing options for Tandem features.

to decrease the MLP's discriminative power by reducing the number of hidden units. An argument can be made that the number of hidden units should be constant, so that each output unit performs an integration over the same number of incoming connections. Assuming such a stance, where the hidden layer is fixed at 1150 units, we see that larger output layers perform better.

The conclusion drawn from this set of experiments is that HMM states seem to be a most promising representation for the outputs of the MLP. Additionally, using state posteriors as soft targets allows the network to be more accurately trained.

3.5   Post-processing options

Table 4 shows the results of varying the post-processing of the output of the MLP, in order to generate suitable Tandem features for a HMM-GMM system.

Early implementations of Tandem features did not apply a logarithm to Gaussianize the softmax posterior; instead, linear output layer activations were were used for feature preparation. Some have claimed that in theory, the two approaches are equivalent because the activations are simply scaled logarithmic posteriors; vice-versa, the logarithmic posterior is a normalization of the linear activations. However, the scaling factor might convey useful information, perhaps relating to confidence of the classifier, which is lost during the softmax normalization. Although the overall performance of the linear outputs appears to be identical to the baseline for the eval04 set, the MAPSSWE test uncovered a large number of differing segments on which the system outperformed the Tandem baseline, providing one of the most significantly different results of these investigations.

Several other post-processing options were tried, without success. Omitting the final concatenation with MFCC features, the 32-dimensional MLP-derived features were insufficient by themselves. Applying a Linear Discriminant Analysis as dimensionality reduction did not help, perhaps because the resulting features were not properly orthogonalized for the GMM. Other researchers have demonstrated success in computing deltas before or after the KLT dimensionality reduction [41, 43], although the investigations in this work failed to replicate those results.

It would have been interesting to investigate the effect of retaining a different number of KLT-reduced components. However, in our experience, this number usually covers at least 95% of the data's variance. Furthermore, because a GMM's likelihoods can depend on the dimensionality of the data, our systems have been tuned to features of a certain size; adjusting the Gaussian weighting factor would not be feasible for a large number of experiments.

3.6 Putting it all together

The results of the previous experiments should be interpreted as giving insight to guide a designer in combining the various modifications. It was determined that modifying the input representation is a tricky issue, so no change was made to the Tandem baseline's 9-frame context and use of derivatives. Using HMM states for the outputs seemed to be good, especially when trained using soft posteriors as targets, so this change was accepted. Finally, there is reason to believe that the post-processing should replace the softmax-logarithm step with linear activations.

The result of these combined modifications: the new system achieves 19.3% CER on eval04.

At this point it should be noted that a hypothesis was made prior to running this final experiment: that these modifications would improve the Tandem baseline system. Therefore it is perhaps acceptable to use a one-tailed test in this situation, which demonstrates significance at the 0.01 level (p = 0.006, one-tailed MAPSSWE test, N=123).

It was also argued that it is perhaps acceptable to keep the number of hidden units constant rather than total number of weight parameters. Using 1150 hidden units did not affect training time much, but improved performance to 19.0% CER on eval04.

4 More data and a best-case scenario

A maxim of the machine learning research community has long been that "there's no data like more data". While complicated system modifications can provide small improvements, sometimes a better use of time and resources is to simply obtain more training data and allow the learning mechanism to have more degrees of freedom to model that data. Such an attitude has recently been adopted, in which large MLPs have been trained for weeks on enormous data sets.

To investigate the effect of using a better-trained classifier, we used a MLP that was trained for the 2007 GALE evaluation. It is similar to the MLP described for the Tandem baseline system with the following exceptions: there were 15000 hidden units, rather than 1150; the training data comprised 870 hours of broadcast news, rather than 29 hours; the target alignments were derived with a procedure for flexible alignment of quickly transcribed closed-caption annotations [44], rather than careful annotation with speaker labels. Even though the larger MLP was trained on more data, we trained the HMM-GMM acoustic models using the same 29hr training set.

Table 5 displays the results of this experiment using a larger MLP. The phone recognition accuracy of the test set is dramatically improved, which leads to a sizeable gain in ASR performance as well. This highlights the importance of having a high-quality MLP classifier, and the desirability of networks which can be ported or adapted to domains which lack a suitable quantity of training data. It is interesting to note that omitting the final concatenation with MFCC features did not have such a drastic effect when the posterior-based features were of higher quality (cf. Table 4).

Given the results of using a better classifier, it is worthwhile to ask what would happen if we had access to a perfect phone classifier. To set up this cheating experiment, we use the phone posteriors from forward-backward alignment as if they were the output of such a classifier. This was straightforward for the training data, but much more complicated for the test data. First the reference transcriptions were Viterbi-aligned to the reference segmentation, then mapped to the automatic segmentation, and finally re-aligned. Good results were obtained for the CCTV subset of eval04, with the 0.9% CER representing deletions due to the segmentation; other subsets of eval04 had about 4% CER. This forced-alignment procedure provided the reference labels for all

Feature type	CCTV CER	CCTV acc.
Tandem baseline (29hr MLP training)	10.6	70.4
... w/ 870hr MLP training	9.0	78.2
... w/ 870hr MLP training w/o concat	9.1	78.2
... w/ gold phone posteriors	5.9	99.2
... w/ gold phone posteriors w/o concat	4.8	99.2
... w/ gold phone posteriors w/o concat (1st-pass)	4.4	99.2
Viterbi-align automatic segmentation to reference	0.9	

Table 5: Character error rate and frame-level phone recognition accuracy on the CCTV test set. Experiments using a MLP trained on more data, and using nearly perfect phone posteriors.

reported results of frame-level phone recognition accuracy of the CCTV test set. Producing phone posteriors with a forward-backward alignment, the maximum value did not always coincide with the Viterbi labeling, hence the 99.2% phone recognition accuracy of the gold posteriors.

A perfect phone classifier was simulated by using the forward-backward alignments of the training and test data. Because the pruned posteriors had many zero values, the vector was interpolated with a lightly-weighted vector produced by the 870hr-trained MLP; the effect was probably only slightly better than adding a small amount of random noise. Interpolating with different weights would have enabled the phone accuracy rate to be manipulated to simulate less-perfect classifiers, though this was not tried.

The results in Table 5 show that virtually perfect phone classification can result in a very good Tandem system – although still far from perfection. It is interesting to note that concatenation with the MFCC features and speaker adaptation can actually degrade the quality, suggesting that this could become a salient issue as MLP classifiers approach this performance limit.

More interesting questions can be asked about the remaining 4% error given perfect phone classification. An analysis of the errors shows that there are very few insertions compared to substitutions and deletions. It is possible to view this experiment as a "best-case scenario" for acoustic modeling, in which the decoder should always know which phone unit to be in at any given time. Perhaps a detailed analysis would determine that some of the remaining errors could be attributed to the language modeling or word segmentation. It might also be interesting to use this style of cheating experiment to compare MLP output units, such as using states rather than phones – or more radically, alternative subword units such as syllables, diphones, or whole words.

5   Discussion and Conclusion

This work has explored Tandem features and demonstrated that it is possible to achieve small gains by modifying the standard configuration along certain dimensions, and larger gains by combining these modifications. However, it is shown that much greater gains can be simply achieved by obtaining more training data to improve the classifier. This improvement can progress up to a certain point – which we are still far from reaching, though we can simulate it to observe some intriguing implications. (For example, it may not be wise to discuss the "best-case scenario" with a funding agency if they are setting unrealistic goals...) This investigation of MLP features has provided insights from which we can hypothesize future directions which are likely to lead us closer

to our goal of better speech discrimination for improved ASR performance.

Although the exploration of input representations was fruitless, it suggests that certain parts of the input may be more relevant than others. However, the fully-connected MLP initially gives equal consideration to all inputs, and may eventually be wasting resources by setting a large quantity of irrelevant weights to zero. It would be desirable for a network to achieve the same performance with fewer weight parameters to allow faster training. One solution could be to perform weight pruning; however, it is more likely that efficiency could be implemented with a pre-defined sparsely-connected structure. In the style of the Tonotopic MLP [23], structured input-to-hidden layer connections could allow a system designer to allocate more hidden units to regions that need them, such as the current context frame. It might be possible to to restrict hidden units to operate on very localized input regions, such as spectro-temporal tiles. Also, it might help to have certain hidden units emulate delta computations by striping their inputs across one feature component only, similar in style to TRAP/HATs processing.

The experiments with output representations suggest that using more expressive outputs also can help, although we need a way to deal with the larger output layer induced by units such as triphones. Intuitively, it also seems odd that so much input context is used to classify an output unit with no context. One possibility is to use a variant of the multi-task MLP [29], in which multiple targets might be used to specify the output at different times.

The benefit of posterior soft targets could become more significant in situations where the distributions are of higher entropy. For example, it could be useful to use soft targets for speaker-adaptive networks in which state posteriors for a test utterance are derived from a first-pass N-best list or recognition lattice.

The post-processing stage of the Tandem approach is necessary to massage posterior distributions into a form that is more easily modeled by a GMM. An alternative might be to view the MLP as performing a sort of vector quantization into a discrete phonetic space; taking the single maximum value of the posterior distribution is akin to finding the nearest codebook value. It may be worthwhile to revisit HMM systems in which discrete observations are modeled, especially considering the potential computational speed advantages of such architectures.

Lastly, the "best-case scenario" experiment presents a novel way of looking at the ASR problem which may be of interest to researchers who wish to decouple the acoustic model and focus on the other parts of the recognition system.

References

[1] PC Woodland and D. Povey. Large scale MMIE training for conversational telephone speech recognition. *Proc. Speech Transcription Workshop*, 2000.

[2] D. Povey and PC Woodland. Minimum phone error and I-smoothing for improved discriminative training. *Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*, 1, 2002.

[3] J. Zheng and A. Stolcke. Improved discriminative training using phone lattices. *EUROSPEECH*, 2005.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. fMPE: Discriminatively Trained Features for Speech Recognition. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 1, 2005.

[5] B. Zhang and S. Matsoukas. Minimum Phoneme Error Based Heteroscedastic Linear Discriminant Analysis for Speech Recognition. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 1, 2005.

[6] DPW Ellis, R. Singh, and S. Sivadas. Tandem acoustic modeling in large-vocabulary recognition. *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 1, 2001.

[7] H. Hermansky and S. Sharma. Temporal patterns (TRAPs) in ASR of noisy speech. *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, 1, 1999.

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

[9] 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. *Acoustics, Speech, and Signal Processing, 2007. Proceedings.(ICASSP'01). 2007 IEEE International Conference on*, 1, 2007.

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

[11] AJ Robinson, GD Cook, DPW Ellis, E. Fosler-Lussier, SJ Renals, and DAG Williams. Connectionist speech recognition of broadcast news. *Speech Communication*, 37(1-2):27–45, 2002.

[12] M.J.F. Gales, DY Kim, P.C. Woodland, HY Chan, D. Mrva, R. Sinha, and S.E. Tranter. Progress in the CU-HTK Broadcast News Transcription System. *IEEE Transactions on Audio, Speech, and Language Processing*, 14, 2006.

[13] D.S. Pallett, J. Fiscuss, J. Garofolo, A. Martin, and M. Przybocki. 1998 Broadcast News benchmark test results: English and non-English word error rate performance measures. *Proc. DARPA Broadcast News Workshop*, pages 5–12, 1999.

[14] D. Pearce. Aurora Project: Experimental framework for the performance evaluation of distributed speech recognition front-ends. *ISCA ITRW ASR2000*, 2000.

[15] H. Hermansky, DPW Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMMsystems. *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 3, 2000.

[16] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky. Feature extraction using nonlinear transformation for robustspeech recognition on the Aurora database. *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2, 2000.

[17] F. Grezl, M. Karafiat, K. Stanislav, and J. Cernocky. Probabilistic and Bottle-neck Features for LVCSR of Meetings. *Acoustics, Speech, and Signal Processing, 2007. Proceedings.(ICASSP'01). 2007 IEEE International Conference on*, 1, 2007.

[18] S. Sivadas and H. Hermansky. Generalized tandem feature extraction. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 1, 2003.

[19] V. Fontaine, C. Ris, and J.M. Boite. Nonlinear discriminant analysis for improved speech recognition. *Proc. Eurospeech*, 97:2071–2074, 1997.

[20] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, D.S. Inc, and MA Newton. Speaker normalization on conversational telephone speech. *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1, 1996.

[21] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990.

[22] B. Chen, S. Chang, and S. Sivadas. Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers. *Proc. EUROSPEECH*, 2001.

[23] BY Chen, Q. Zhu, and N. Morgan. Tonotopic Multi-Layered Perceptron: A Neural Network for Learning Long-Term Temporal Features for Speech Recognition. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 1, 2005.

[24] Q. Zhu, B. Chen, F. Grezl, and N. Morgan. Improved MLP Structures for Data-Driven Feature Extraction for ASR. *Proceedings of European Conference on Speech Communication and Technology*, 2005.

[25] P. Schwarz, P. Matejka, and J. Cernocky. Hierarchical Structures of Neural Networks for Phoneme Recognition. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 1, 2006.

[26] S. Sivadas and H. Hermansky. Hierarchical tandem feature extraction. *Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*, 1, 2002.

[27] K. Livescu, O. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, et al. ARTICULATORY FEATURE-BASED METHODS FOR ACOUSTIC AND AUDIO-VISUAL SPEECH RECOGNITION: SUMMARY FROM THE 2006 JHU SUMMER WORKSHOP. *ICASSP 2007*, 2007.

[28] A. Stolcke, F. Grezl, M.Y. Hwang, X. Lei, N. Morgan, and D. Vergyri. CROSS-DOMAIN AND CROSS-LANGUAGE PORTABILITY OF ACOUSTIC FEATURES ESTIMATED BY MULTILAYER PERCEPTRONS. *ICASSP 2006*, 2006.

[29] J. Frankel, O. Cetin, and N. Morgan. Transfer Learning for MLP-derived feature transforms. *NOLISP: ISCA Tutorial and Workshop on NOnLInear Speech Processing*, 2007.

[30] H. Misra, H. Bourlard, and V. Tyagi. New entropy based combination rules in HMM/ANN multi-stream ASR. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2, 2003.

[31] Fabio Valente and Hynek Hermansky. Combination of Acoustic Classifiers based on Dempster-Shafer Theory of evidence. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[32] H. Bourlard, S. Bengio, M.M. Doss, Q. Zhu, B. Mesot, and N. Morgan. Towards using hierarchical posteriors for flexible automatic speech recognition systems. *Proc. of the DARPA EARS RT04 Workshop*, 2004.

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

[34] M.Y. Hwang, X. Lei, W. Wang, and T. Shinozaki. Investigation on Mandarin Broadcast News Speech Recognition. *ICSLP*, 2006.

[35] SJ Young, JJ Odell, and PC Woodland. Tree-based state tying for high accuracy acoustic modelling. *Proceedings of the workshop on Human Language Technology*, pages 307–312, 1994.

[36] D. Ellis and N. Morgan. Size matters: an empirical study of neural network training forlarge vocabulary continuous speech recognition. *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, 2, 1999.

[37] L. Gillick, SJ Cox, D.S. Inc, and MA Newton. Some statistical issues in the comparison of speech recognition algorithms. *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 532–535, 1989.

[38] DS Pallet, WM Fisher, and JG Fiscus. Tools for the analysis of benchmark speech recognition tests. *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 97–100, 1990.

[39] X. Lei, M.Y. Hwang, and M. Ostendorf. Incorporating Tone-related MLP Posteriors in the Feature Representation for Mandarin ASR. *Proc. Interspeech*, pages 2981–2984, 2005.

[40] F. Metze and A. Waibel. A flexible stream architecture for ASR using articulatory features. *Proc. ICSLP*, 2002.

[41] D.P.W. Ellis and M.J.R. Gomez. Investigations into tandem acoustic modeling for the Aurora task. *Proc. Eurospeech-2001, Special Event on Noise Robust Recognition*, 2001.

[42] P. SCHWARZ, P. MATEJKA, and J. CERNOCKY. Towards lower error rates in phoneme recognition. *Lecture notes in computer science*, pages 465–472, 2004.

[43] M.J. Reyes-Gomez and D.P.W. Ellis. Error visualization for tandem acoustic modeling on the Aurora task. *ICASSP 2002*, 2002.

[44] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V.R.R. Gadde, and J. Zheng. An Efficient Repair Procedure For Quick Transcriptions. *Proc. ICSLP*, pages 2002–2005, 2004.