On Using MLP Features in LVCSR

Qifeng Zhu[1], Barry Chen[1,2], Nelson Morgan[1,2], Andreas Stolcke[1,3]

[1]International Computer Science Institute, [2]University of California, Berkeley, [3]SRI International
{qifeng, byc, morgan, stolcke}@icsi.berkeley.edu

Abstract

One of the major research thrusts in the speech group at ICSI is to use Multi-Layer Perceptron (MLP) based features in automatic speech recognition (ASR). This paper presents a study of three aspects of this effort: 1) the properties of the MLP features which make them useful, 2) incorporating MLP features together with PLP features in ASR, and 3) possible redundancy between MLP features and more conventional system refinements such as discriminative training and system combination. The paper shows that MLP transformations yield variables that have regular distributions, which can be further modified by using logarithm to make the distribution easier to model by a Gaussian-HMM. Two or more vectors of these features can easily be combined without increasing the feature dimension. Recognition results show that MLP features can significantly improve recognition performance in large vocabulary continuous speech recognition (LVCSR) tasks for the NIST 2001 Hub-5 evaluation set with models trained on the Switchboard Corpus, even when discriminative training and system combination are used.

1. Introduction

MLPs have been successfully used for pattern classification in many application areas. With large enough training data and MLP size, and using 1-of-c binary coded class targets, an MLP can learn the posterior probability of a class given an observation, *P(c|o)*. This was effectively used in *acoustic modeling* in hybrid MLP-HMM systems [6], where the scaled likelihood of a frame given a phone state in an HMM is computed by the posterior probability, *P(c|o)*, scaled by the phone prior probability, *P(c)*. This inherently discriminative approach worked well for many tasks, and was particularly useful for combinations of features with different statistical properties (e.g., continuous and binary features) [9].

On the other hand, the dominant paradigm for speech recognition has incorporated mixtures of Gaussians to represent emission distributions for HMMs. Within this framework, many performance-enhancing refinements have been developed, such as feature adaptation and discriminative training. To benefit from the strengths of both MLP-HMM and Gaussian-HMM techniques, the Tandem solution was proposed in 2001 using MLP outputs as observations for a Gaussian-HMM [5]. An error analysis of Tandem MLP features [8] showed that the errors of the system using MLP features are different from the errors of a system using cepstral features. This suggested that a combination of both feature styles might be even better. This was first applied in the Aurora task [1], and later in the EARS project [7] on large vocabulary continuous telephone speech recognition. Here we continue this work, applying the combination techniques to increasingly more advanced systems, and noting the properties of the features that might be responsible for the virtues of the MLP-based features.

In Section 2.1 properties of MLP features are discussed. Section 2.2 shows some technical details on how to combine MLP features with PLP features to achieve good ASR results. Section 2.3 presents results with systems incorporating other techniques that provide error reduction that could be redundant with that provided by MLP features; in particular, discriminative training and better language model rescoring, and system combination (ROVER).

2. Using MLP-based Feature in LVCSR

2.1. Properties of MLP Features

Currently short-term spectral-based (typically cepstral) features are used in ASR, such as MFCC, or PLP. These features are typically non-Gaussian, and are most often modeled by mixtures of Gaussians. When diagonal covariance matrices are used, many Gaussian mixtures can be needed to effectively model the feature distribution.

MLPs are effective at modeling unknown distributions. Cepstral features can be used as inputs to train an MLP with phoneme classes as targets. The MLP outputs, which are approximations to phone posterior probabilities given input features, can also be used as features for HMM. This MLP can then also be regarded as a nonlinear feature transform. There have been many kinds of linear feature transforms, such as LDA or HLDA [3], that make the transformed feature better for modeling by Gaussian mixtures for an HMM. This then suggests a question: when an MLP is used as a feature transform, i.e., when posterior approximations are used as features, what properties of this approach make it useful?

For the work reported here, the MLPs are trained using 46 mono-phones as targets. Thus, the MLP outputs have 46 components. For each phone class, one out of the 46 components corresponds to the underlying phone class. We call this component the “in-line” component for the class, and the rest are “off-line” components.

The distributions of MLP outputs are more regular than those for the PLP features. Figure 1 shows the feature distribution of three phone classes /ah/(triangle), /ao/ (star), and /aw/ (circle) in feature space spanned by the first three components of PLP feature. Figure 2 shows the feature space of the three MLP outputs corresponding to the same three phone classes, trained using PLP feature. The feature distributions for the three classes are more regular in the MLP feature space than PLP feature space. This is because MLP is able to discriminatively learn the irregular class boundaries and transform the features within the boundary close to their class target used in training, and transform the irregular class boundaries to equal posterior hyper-planes in the feature space of posteriors. The outputs are not always good estimates of the

class posteriors. Errors can happen, as seen in the figures, and the frame accuracy we can achieve is about 70% based on the highest posterior. But since the outputs are complemented by PLP features, they are rarely altogether bad.

While the distributions of MLP outputs shown in Figure 2 are regular, they are difficult to model using Gaussian mixtures due to the sharp shape of the distribution. In-line components distribute roughly uniformly between 0 and 1 and taper away near 0, and off-line components distribute very narrowly around 0. To Gaussianize the feature distributions, a simple approach is to take the log of the MLP outputs. Figure 3 shows the feature distributions of the log MLP outputs for the same three classes of the same three MLP output components. In the remainder of this paper we will refer to the log MLP outputs as the MLP features.

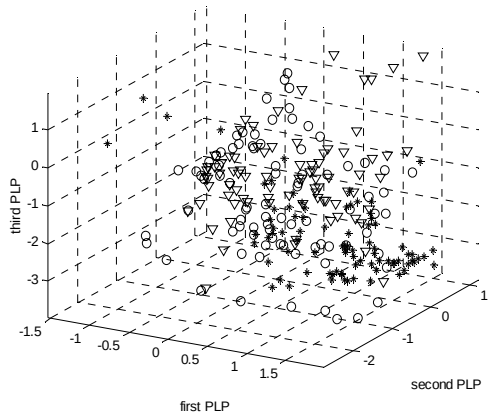

Figure 1: Feature distributions of the first three PLP components for three classes, /ah/(triangle), /ao/ (star), and /aw/ (circle).

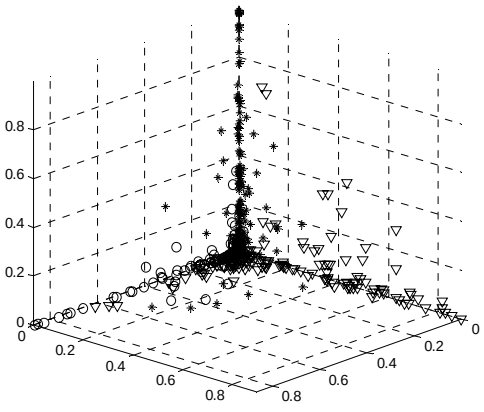

Figure 2: Feature distributions of the three MLP components corresponding to the three classes, /ah/(triangle), /ao/ (star), and /aw/ (circle).

The typical distribution of the in-line component of the MLP feature is concentrated close to 0 and tapers away gradually, which is due to the compression of the high posteriors close to 1 and the expansion of low posteriors by the logarithm. The typical distribution of an off-line component is close to a single Gaussian centered at a high negative number, which is expanded by the logarithm from the narrow posterior distribution close to zero. Two typical distributions of the log in-line and off-line components are shown in Figure 4. These distributions in Figures 3 and 4 should be easier to model with Gaussian mixture models.

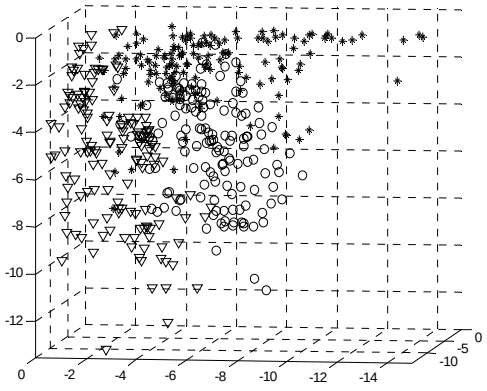

Figure 3: Feature distributions of the log MLP outputs.

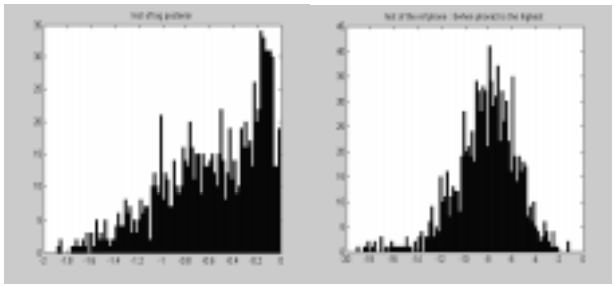

Figure 4: Typical distributions (histogram) of an in-line (left) and an off-line (right) MLP feature component.

Besides making the feature space more regular, the MLP feature can reduce the variation among speakers. Speaker variation is one of the major sources of within-class variation that degrades acoustic models. We compute PLP features with per-speaker (in practice, per-conversation-side) vocal tract normalization, where piece-wise linear frequency warping for each speaker is used to reduce speaker variation [10]. This VTLN is important but still leaves much speaker variability due to factors other than frequency warping. MLP features, trained with different speakers for the same target, can decrease this. PLP features for the same class from different speakers may be located in a different point in the PLP feature space, but may have similar class posterior probability, and thus be transformed to the same point after MLP transformation. In other words, posteriors are by nature speaker independent, if the training is speaker balanced instead of biased. A way to show this is by looking at the variances of speaker adaptive training (SAT) transforms among speakers, since differences on SAT transform could be used to represent speaker differences [4]. An SAT adaptation matrix was computed for each speaker on the concatenation of PLP features of 39 dimensions (13 static and their first and second derivatives) and MLP features (orthogonalized and truncated) of 25 dimensions, where all the feature components are normalized

to zero mean and unit variance. If there is no speaker variation, then all the SAT transforms should be the same; otherwise, they differ. Figure 5 shows the variances of every component of the SAT transform matrix among different speakers. The first 39 by 39 block of the SAT transform matrix has high variances, which means more variations among speakers in the PLP feature, but the next 25 by 25 block has much smaller variances, which means small variation among speakers. The ratio of the average variance of PLP block and the MLP feature block is 1.6.

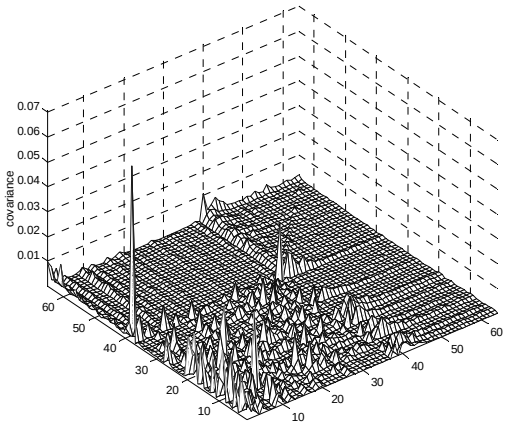

Figure 5: Variances of all elements in the SAT transform among different speakers.

While MLP features are being used here as HMM observations, they still have the properties of (log) phone posterior estimates. This makes it possible to combine different MLP outputs trained with same targets but different inputs to improve the features without increasing the total feature dimension. These MLPs may emphasize different aspects of the target class, so a combination may yield further improvement. In practice, we use two types of MLPs and combine MLP outputs using a weighted sum, where the weights are a normalized version of the inverse entropy [7]. The two types of MLP features are the PLP-MLP feature and the TRAP (or HATs) feature [2], which offer complementary information on the phone class. The first of these incorporates inputs from roughly 100 ms of speech (using 9 sequential PLP cepstral vectors as input), and the second uses 500 ms of input (from critical band energy trajectories).

2.2. Using MLP Features with PLP Features

Using MLP feature alone as in the tandem scheme has been discussed in papers such as [5]. To fully take advantage of the benefit together with regular features such as PLP feature, we combine PLP features with MLP features.

A simple way to combine PLP and MLP features is to concatenate them. The resulting feature can be as long as 39+46=85, where 39 is the dimension of PLP feature and 46 is that of MLP feature. The resulting feature vector could contain significant redundancies, and so a dimensionality reduction was considered. Two approaches were tried. KLT was applied to the MLP features and the components corresponding to the smallest eigenvalues were truncated. Another approach was to apply a more complicated linear feature transform, HLDA, to search for the best non-trivial feature mapping direction and discard the nuisance dimensions [3][10]. With KLT, by keeping the components corresponding to the highest 17 eigenvalues, the remaining dimensions had 95% of the total original feature variances. Keeping the most significant 25 dimensions after KLT covered 98% of the total original variance. Including all dimensions doesn't always improve ASR, but a truncated feature often gave better results. Further decreasing the dimension of the MLP feature to less than 15 was found to hurt ASR performance.

A critical detail that we found to be important for ASR with long feature vectors was to modify the acoustic modeling parameters. In particular, we found it useful to optimize a scale factor on the log likelihood associated with the individual Gaussians in the mixture (Gaussian weight). For a longer feature, log likelihood has a larger dynamic range, and the Gaussian weight should be tuned to a much lower number such as 0.3 instead of 0.8, which was the best tuned value to fit PLP features with 39 dimensions.

When VTL normalized PLP is used as an HMM feature, per-speaker mean and variance normalization is still helpful. We also use per-speaker mean and variance normalized PLP feature to train MLPs and to generate MLP features. The MLP feature is then normalized globally to zero mean and unit variance after KLT or HLDA before being used as an observation for the HMMs.

An extra per speaker normalization after MLP transformation was still helpful, apparently reducing some speaker variation left after the mean and variance normalization before the MLP based nonlinear transform. This is not a property of linear transforms, for which a previous per-speaker mean and variance normalized feature will still be per-speaker normalized, if the transform matrix is properly scaled.

The SRI Decipher system [10] was used to conduct recognition experiments. Table 1 shows the ASR results in word error rate on the NIST 2001 Hub-5 test set. The training set contained about 68 hours of conversational telephone speech (largely Switchboard) data. Gender dependent HMMs were trained with a maximum likelihood criterion, and a bigram language model was used in the decoding. The MLP feature dimension was reduced from 46 to 25 using KLT or HLDA. Baselines are PLP with first 2 derivatives (PLP) and PLP with the first three derivatives followed by HLDA to reduce to 39 dimension (referred to as PLP2); previous experience had suggested that incorporating more than two derivatives without HLDA was not useful. Results show 6-10% error reduction by adding MLP features, and an extra per-speaker mean and variance normalization after the KLT on the MLP feature further reduced errors. The last two rows show results by adding MLP feature trained with PLP (PLPMLP) and HATs as individual rather than combined MLP features. Clearly the combination of MLPs focusing on long term and short term gives lower WER.

The features were further tested with MLLR adaptation, where three adaptation transform matrices were computed for each speaker to reduce the difference between the testing condition of the speaker and the trained HMMs. Table 2 shows that the MLP features work well with MLLR. We were not able to make HLDA work as well as the simpler KLT with MLLR. Thus, KLT based truncation became our choice for MLP features, which is used in the experiments in Section 2.3.

Feature	Word Error Rate (Relative error reduction)
PLP baseline	39.1
PLP + MLP-KLT25	35.2 (10%)
PLP2 (*) baseline	37.2
PLP2 + MLP-HLDA25	34.4 (7.8%)
PLP2 + MLP-KLT25	34.8 (6.5%)
PLP2 + MLP-KLT25-spknorm	34.0 (8.6%)
PLP2+PLPMLP-KLT25-spknorm	36.1 (3.0%)
PLP2+HATs-KLT25-spknorm	36.0 (3.2%)

Table 1: Word error rate on NIST 2001 Hub-5 test set with PLP baseline and PLP plus different MLP features. (* PLP2 is PLP with three derivatives plus HLDA.)

Feature	WER (Error reduction)
PLP2 baseline	35.8
PLP2+ MLP-HLDA25	33.9 (5.3%)
PLP2+ MLP-KLT25	33.4 (6.7%)
PLP2 + MLP-KLT25-spknorm	32.6 (8.9%)

Table 2: Word error rate on NIST 2001 Hub-5 test set. MLLR adaptation is used on PLP2+MLP feature.

2.3. Using a Better LVCSR System

A key question is whether the front-end based techniques described here provide error reductions that are redundant or complementary with more advanced backend techniques, i.e., more powerful acoustic modeling and decoding techniques, which are used in state-of-the-art LVCSR systems. Impressive improvements provided in simple systems can often disappear with more powerful systems. To address this concern, we used versions of the SRI system that included discriminative training using a maximum mutual information (MMI) criterion, a 4-gram language model and duration model based rescoring (4G-LM), and system combination (ROVER) with MFCC-based ASR [10]. A related concern was whether improvements could still be observed when more training data is used.

Table 3 shows the recognition results with the improved system. To save time, the gender-dependent experiment was only conducted on male data (although spot checks with females showed similar results). A full-fledged LVCSR system was trained using 200 hours of male Switchboard data, but MLPs were trained with 128 hours of male speech. Results show that the relative error reduction due to adding MLP features can carry through to the improved system.

3. Summary and Conclusion

Using MLP features can improve ASR performance on a large conversational telephone speech recognition task, even when large amounts of training data, discriminative training (MMI) and other system enhancements are used.

MLP features provide a data-driven front-end approach to feature extraction that improves discrimination and ease of modeling. They are optimized to approximate phone posteriors. The MLP is trained discriminatively, can reduce speaker variability that is irrelevant to word recognition, and can generate feature distributions that are easily modeled by Gaussian mixture-based HMMs.

Experiments show that MLP features offer unique benefits that appear to be complementary to those provided by other techniques such as MMI based discriminative training and system combination. When properly used, MLP features can improve ASR performance in the conversational telephone LVCSR task significantly by reducing errors from 5% to 9%.

ASR System	PLP2	+ MLP	Error reduction
MMI	30.8	28.6	7.1%
MMI+ 4G-LM	25.6	23.5	8.2%
+System ROVER	24.5	23.0	6.1%

Table 3: Male WERs and relative error reductions on NIST 2001 Hub-5 set with the improved system.

4. Acknowledgements

This work was made possible by funding from the DARPA EARS Novel Approaches Grant No. MDA972-02-1-0024. Thanks also to Hynek Hermansky, Ozgur Cetin, and Pratibha Jain for many relevant discussions.

5. References

[1] Benitez, C., Burget, L., Chen, B., Dupont, S., Garudadri, H., Hermanskey, H., Jain, P., Kajarekar, S., Morgan, N., and Sivadas, S., "Robust ASR front-end using spectral based and discriminant features: experiments on the Aurora task", *Eurospeech 2001*.

[2] Chen, B., Zhu, Q., Morgan, N. "Learning long term temporal features in LVCSR using neural networks", *submitted to ICSLP 2004.*

[3] Gales, M.J.F., "Semi-tied covariance matrices for hidden Markov models", *IEEE Trans. Speech and Audio Processing, vol 7, pp. 272-281, 1999.*

[4] Gao, X., Zhu, W., and Shi, Q., "The IBM LVCSR System Used for 1998 Mandarin Broadcast News Transcription Evaluation", *Proc. DARPA Broadcast News Workshop, 1999.*

[5] Hermansky, H., Ellis, D.P.W. and Sharma, S. "Tandem connectionist feature extraction for conventional HMM systems", *ICASSP 2000, pp. 1635-1638.*

[6] Morgan, N. and Boulard, H., "Continuous speech recognition", *IEEE Signal Processong Magazine, vol. 12, no. 3, pp. 24, May 1995.*

[7] Morgan, N., Chen, B., Zhu, Q. and Stolcke, A., "TRAPping Conversational Speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition", *ICASSP 2004.*

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

[9] Robinson, A.J., Cook, G.D., Ellis, D.P.W., Fosler-Lussier, E., Renals, S.J., and Willianms D.A.G., "Connectionist speech recognition of Broadcast News", *Speech Communication, vol. 37, no. 1-2, pp. 27-45. 2002*.

[10] Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde V.R., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F., and Zheng, J., "The SRI march 2005 Hub-5 conversational speech transcription system", *Proc. NIST Transcription Workshop 2000.*