Improved MLP Structures for Data-Driven Feature Extraction for ASR

Qifeng Zhu[1], Barry Chen[1,2], Frantisek Grezl[1,3], Nelson Morgan[1,2]

[1]International Computer Science Institute, [2]University of California, Berkeley,
[3]Brno University of Technology

{qifeng,byc,franta,morgan}@icsi.berkeley.edu

Abstract

In this paper, we present our recent progress on multi-layer perceptron (MLP) based data-driven feature extraction using improved MLP structures. Four-layer MLPs are used in this study. Different signal processing methods are applied before the input layer of the MLP. We show that the first hidden layer of a four-layer MLP is able to detect some basic patterns from the time-frequency plane. KLT-based dimension reduction along time is applied as a modulation frequency filter. The new feature extraction was tested on a large vocabulary continuous speech recognition (LVCSR) task using the NIST 2001 evaluation set. We achieved 11.6% relative word error rate (WER) reduction compared to the traditional PLP-based baseline feature. This is also a significant improvement compared to our previously published results on the same task using MLP-based features with three-layer MLPs.

1. Introduction

Spectral-based features, such as MFCC or PLP, have been used as the dominant feature in automatic speech recognition (ASR) systems for more than a decade. Currently we are conducting research on a novel feature extraction method using MLP-based data-driven approaches and its application in LVCSR tasks [3]. In this scheme, the spectral information across several frames is transformed by a MLP, which is trained to classify different phones. The MLP outputs are further processed and used as the front-end feature for HMM-based speech recognition systems.

Our previous papers have shown significant improvements using MLP-based features obtained with this approach [2][5] on different LVCSR tasks, including the NIST 2004 Evaluation on Hub5 Continuous Telephone Speech (CTS) [6]. The topics of our previous papers cover long-term feature extraction, MLP output combination for merging long-term and short-term cues, using MLP-based feature with HMM for ASR, and the properties and performance of the feature in different tasks.

This paper presents our recent work and progress in exploring the structures and constraints of a MLP for data-driven feature extraction. Previously we used single or concatenated three-layer MLPs with specific constraints. In this paper we used a four-layer MLP structure. At the same time, we relaxed the narrow-frequency band constraints used in our previous long-term MLP systems. We also tested different signal processing methods before the input of a MLP. With these explorations, we found three MLP structures that can effectively extract different and complementary information from different views of the speech spectrogram. By combining the three MLPs, we are able to get more significant word error rate reduction using the MLP-based features. Compared with the baseline feature, PLP with first three derivatives followed by HLDA-based dimensionality reduction, we are able to reduced WER by 11.6% relative by using the MLP-based feature concatenated with the PLP baseline feature.

2. Feature extraction using four-layer MLP

2.1 Using MLP to learn phone posterior

MLPs are a class of popular mechanisms in the machine learning world. A MLP can be trained so that the output approximate class posteriors [1]. Our MLPs are trained with 46 mono-phones as the targets, and MLP outputs approximate phone posterior probabilities.

One important practical issue with MLPs is the choice of structure. Without knowing the detailed properties of the problem, the three-layer MLP is the most frequently used structure, since it can theoretically model any class boundary, if the hidden layer size is large enough. It maps the input feature (the input layer) nonlinearly, through a sigmoid function, to a hidden layer that often has a higher dimensionality. Hopefully, in the space spanned by the nodes of the hidden layer, different classes are linearly separable. A linear combination of the hidden layer outputs is applied to form linear boundaries among classes, further followed by a normalization using a softmax function so that the sum of the MLP outputs equals one.

In our previous work, we mostly use single or concatenated three-layer MLPs for extracting phone posterior from long or short time region of a spectrogram. Posteriors from different MLPs can be further combined for higher accuracy using simple or weighted average of different MLP posteriors. The combined posteriors are processed by taking logarithm and applying KLT-based dimensionality reduction. The MLP-based features are then concatenated with regular PLP features to make a final feature vector, which is used as the input to the HMM-based ASR system. The details are shown in [5].

2.2 From three-layer MLPs to four-layer MLPs

Instead of using a generic three-layer MLP structure, it might be possible to explore different details of the MLP structure so that the MLP might better fit with the underlying problem. It is known that the detailed structure of a model is important for modeling a problem. Unfortunately, there is often no systematic way in finding the best structure. This is true with many learning machines, for example, determining the detailed structure of an HMM with state-jump constraints. The exploration is often motivated or inspired by some observations or thoughts, and verified by experiments. In this

sense, this paper doesn't show a mathematical proof that the proposed structure is better than the structures we used before, but rather reports on new findings using different effective MLP structures, our motivations and our observations with the new structures, and possible explanations for the better results in ASR tasks.

We have several motivations for using four-layer MLPs. The initial motivation is that, since we already used concatenated three-layer MLPs, (for example in HATS [2]), we could try to combine them to form a four-layer MLP so that they have roughly the same structure but optimization can be performed jointly in one step. Another motivation is that, in our previous concatenated MLP structure, we applied certain constraints, for example, the band-separation constraint. With a four-layer MLP we want to experiment with relaxing these constraints so that the MLP can learn from the entire input. Another important motivation is that we guess there might be a limited number of basic "patterns" in the time-frequency plane. It might be more effective to first extract these patterns out of the time-frequency plane, then learn the phone posteriors from these basic patterns. For this reason, one might want to add another layer after the input layer to only learn these patterns, before they are further transformed to another layer where different classes are expected to be more linear separable.

With these motivations, we constructed our four-layer MLPs. There are two hidden layers, the first hidden layer is often smaller in size compared to the second hidden layer. Figure 1 shows the MLP structure. A slice of the time-frequency plane goes through a signal processing block, for example, log critical-band filter analysis, PLP analysis, or log critical-band analysis followed by DCT-based dimensionality reduction.

With such a MLP structure, the second layer bears the same meaning as the hidden layer at the three-layer MLP: it is optimized so that different classes are linearly separable. The first layer, however, can be viewed as a pattern detector. Looking at one node in the first hidden layer and all the connections from the input layer to this node, as illustrated by the circled node in Figure 1, the node gives maximal output when the MLP input matches the weights connecting to this node. If a weight is zero, then the output of that node is not sensitive to the input via that connection. The output of a node is sensitive to the pattern corresponding to the connection weights of large absolute values. For this reason, we can call the first layer a "pattern detector layer", and the second layer a "classification layer".

3. The Four-Layer MLPs

One advantage of using MLPs in feature extraction is that we are able to use different information sources as the input to different MLPs and later combine them. In our previous papers, we used two information sources: the log critical-band energy from long time period of a half second, and PLP with first two derivatives from short time period of 9 frames. We found that MLP outputs using these two information sources are complementary and can be effectively combined.

In this paper, we use three types of MLPs with different information sources as the input. All these MLPs have the basic structure shown in Figure 1. Their detailed sizes are tuned using the training data to maximize the frame accuracy of a held-out cross-validation set. The details of each of the three types of MLPs are presented below. MLPs are trained in a gender dependent manner. Each MLP has about 500K parameters, and is trained with 32 hours of gender dependent speech data.

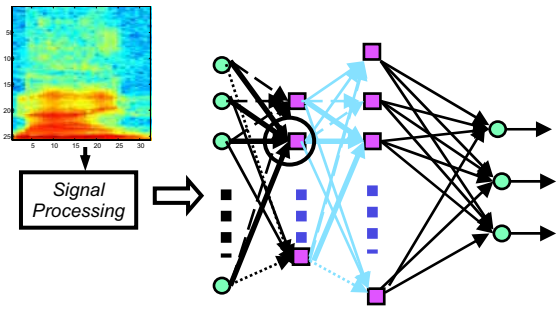

Figure 1: The general structure of a four-layer MLP for extracting phone posteriors from the time-frequency plane.

3.1 MLP with Short-term log-critical-band energy as the input (Short LCBE MLP4)

This is a new MLP structure compared to our previously published systems. The input to the MLP contains log critical-band energy of 9 consecutive frames. We use 15 Mel scale critical-bands for each frame. Thus the input vector of the MLP has 9*15=135 dimensions. The first hidden layer has 600 nodes and the second hidden layer has 672 nodes, and there are 46 output nodes corresponding to 46 different mono-phones.

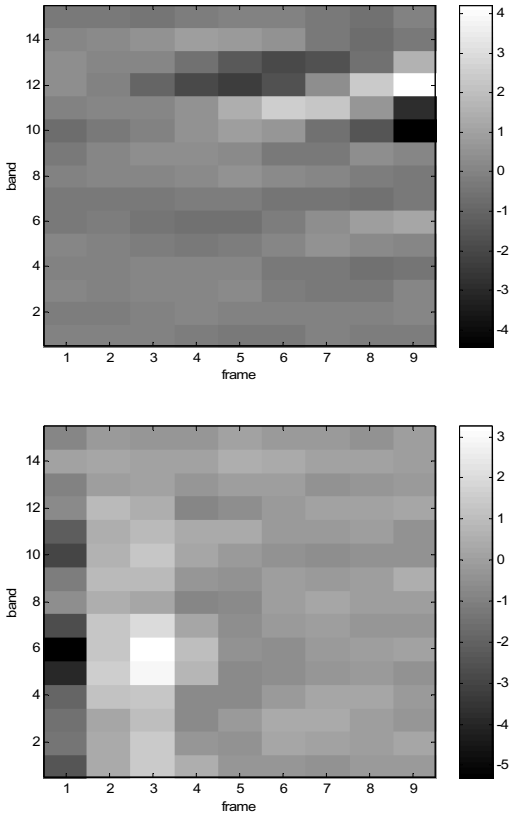

Figure 2: Illustration of the pattern corresponding to the weights connecting to a certain hidden node in the first

hidden layer. This hidden node is sensitive to a particular localized pattern in time-frequency plane.

By looking the patterns of the weights connecting the input to the first hidden layer, we often see:

- Only a small number of weights connecting to a node have values far away from zero.
- These nonzero weights often form localized patterns.

This means that the nodes in the first hidden layer are sensitive to certain localized input patterns. Some examples of the localized patterns represented by the weights are illustrated in Figure 2. The x-axis is the frame index, where the 5th frame is the center frame, and frames 1-4 are the previous frames and frames 6-9 are the future frames, and the y-axis is the critical-band index, which corresponds to frequency change. Some of these self-organized patterns seem to be tracking energy change, or some formant change in the time-frequency plane.

3.2 MLP with Short-term PLPs as the input (Short PLP-MLP4)

The input to this MLP is the 9 frame concatenation of regular PLP cepustal features plus energy, and their first two derivatives. The static PLP feature is a vector of dimension 12. Thus the total input size to MLP is (12+1)*3*9=351. The first hidden layer has 300 nodes, and the second hidden layer has 1187 nodes.

Again, the majority of the weights connecting to each node in the first hidden layer have values close to zero, and those values far from zero form some patterns. The patterns on the PLP feature are not localized across PLP coefficients, but are localized in time. Figure 3 shows a pattern from the weights of the PLP-MLP4. The y-axis are the 39 PLP plus derivatives.

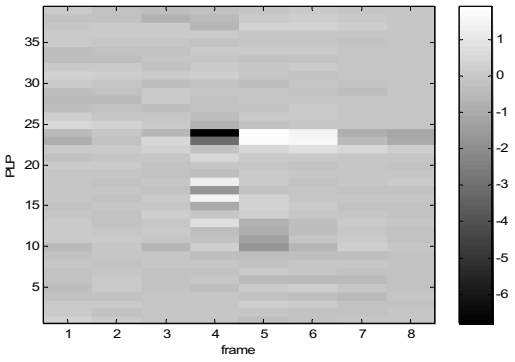

Figure 3: Illustration of a pattern corresponding to the weights connecting to a certain hidden node in the first hidden layer of a PLP-MLP4.

3.3 MLP with long-term log-critical-band energy as the input (Long LCBE MLP4)

In previous work we extracted long-term information using a MLP structure called HATS [2]. In this paper, we use four-layer MLPs. Besides the change in MLP structure, we also changed the signal processing before the MLP input. The spectrogram of 51 frames (about half second) first goes through a log critical-band analysis similar to that in section 3.1 with 15 bands. Then for each band a Hamming window is applied on the band energy from 51 frames to decrease the contribution of frames far away from the center frame.

It has been found that different modulation frequencies have different importance in speech perception [4]. It is often helpful to apply a modulation frequency filter to remove some high modulation frequency components. We use a DCT-based modulation frequency filter on the windowed log critical-band energy for each band. The 51-point DCT outputs are truncated to only keep the first 26 points. The input to the MLP has dimension of 26*15=390. The two hidden layers both have dimension of 520.

It is hard to discern patterns from the weights to the first hidden layer of this MLP, since the time index has been warped by the DCT. But the weights are larger at low DCT indexes, corresponding to lower modulation frequencies, as shown in Figure 3. This agrees with human recognition experiments showing that lower modulation frequencies are more important.

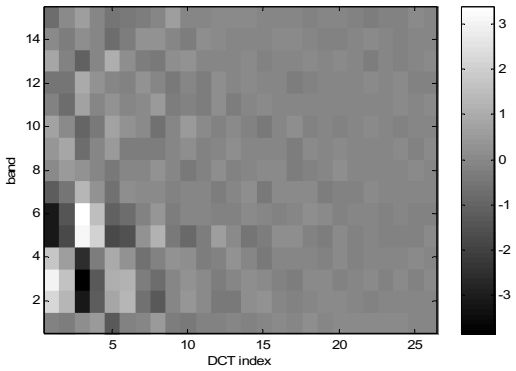

Figure 4: Illustration of a pattern corresponding to the weights connecting to a hidden node in the first hidden layer of a Long LCBE MLP4.

3.4 MLP combination

There are many ways to combine MLP outputs, for example, taking the average, or taking the product of the posteriors from different MLPs. We tested different ways in combining MLP outputs in the past. For those experiments, the combination method was not a critical factor and the results were similar. We have tended to use inverse entropy weighting-based posterior combination [5], which was less sensitive to some bad MLP outputs. We re-visited the combination methods in combining the three MLP outputs.

4. Experiments and Results

The experimental setting is the same as used in [5] . The SRI Decipher system was used to conduct recognition experiments. The training set contained about 64 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 test set is the NIST 2001 evaluation set.

The baseline feature is PLP with first three derivatives of 52 dimensions, followed by HLDA-based dimensionality reduction to 39. The MLP-based feature, of dimensionality 25, is appended to the PLP feature to form a super feature vector. The MLP-based feature can be derived from one single MLP or from combinations of two or three MLPs.

Table 1 shows the results of the MLP-based feature based on the combination of the three MLPs in Section 3. The first row shows the baseline WER using the PLP baseline. In the second row, three MLPs are combined using inverse entropy weighting-based combination (comb1). In the third row, the MLPs are combined by taking the average of the log posterior from each MLP (comb2). The last row shows the WER of using the 25-dimension MLP-based feature alone from the third row.

A significant improvement is achieved by appending MLP-based feature with the PLP baseline features with a WER reduction of 11.6%. Using the MLP-based feature alone (25 dimension) is also significantly better than the PLP feature (39 dimension) by 3.5%. In this experiment, all of the three MLPs have similar good performance, and it turns out that taking the average of the log posterior gives better performance.

Feature	Word Error Rate (Relative error reduction)
PLP baseline	37.2
PLP + Three MLP comb1	33.3 (7.8%)
PLP + Three MLP comb2	32.9 (11.6%)
Only, Three MLP comb2	35.9 (3.5%)

Table 1: The word error rate of the MLP-based feature with three MLP output combined. The test set is NIST 2001 Evaluation set.

Table 2 shows the WER using the MLP-based feature on each single MLP proposed in this paper, and the MLPs used in our previous papers, the HATS and the PLP-MLP3, where three-layer MLPs or concatenated three-layer MLPs are used. The previous structures, HATS and PLP-MLP3, have WER around 35.6% in this test set. Each of the new MLP structures from this paper is significantly better than all the old versions by WER reductions more than 1% absolution. All the three MLPs proposed in this paper have similar performance.

Feature	Word Error Rate (Relative error reduction)
PLP baseline	37.2
+ Long LCBE-MLP4	34.5 (7.3%)
+ HATS	35.6 (4.3%)
+ Short PLP-MLP4	34.4 (7.5%)
+ Short PLP-MLP3	35.6 (4.3%)
+ Short LCBE-MLP4	34.6 (7.0%)

Table 2: The word error rate using MLP-based feature from each MLP in this paper and the previous MLP structures.

We know based on our previous experience that MLPs complementary to each other can give large improvements through combination. Table 3 shows the WER results of the combination of every pair of MLPs in this paper to show how complementary these MLPs are to each other. The most complementary pair is the MLP with long term log critical-band energy and the MLP with input of short term PLP inputs in the last row. The least complementary is the MLP with long term log critical-band energy and the short term log critical-band energy inputs (LCBE combination in the second row), but even this case, the combined MLPs is still significantly better than using any single MLP alone.

5. Conclusion

We saw significant improvements with the three new MLP structures in this paper, used alone or in combination, compared with the previous MLP structures we used. We argued some possible reasons that the four-layer MLP fits better in this task. The first hidden layer, especially in the case when using log critical-band energy directly as the MLP input, is able to catch some basic patterns, and further classification can be more effective-based on these patterns than-based directly on the time-frequency plane. For extracting the long term feature, applying DCT-based truncation is found effective, by removing the high modulation frequencies.

MLPs are often used as black-boxes. But we found the four-layer MLP is able to extract some basic patterns from the spectrogram. Besides better performance in WER, such a self-organization behavior opens a door to our future research.

Feature	Word Error Rate (Relative error reduction)
PLP baseline	37.2
+ Short-term nets combination	33.5 (9.9%)
+ LCBE combination	33.9 (8.9%)
+ Long LCBE-MLP4 and PLP-MLP4 combination	33.4 (10.2%)

Table 3: The word error rate of the MLP-based feature by combining every two MLPs in this paper.

6. Acknowledgements

This research is supported by the DARPA EARS Novel Approaches Grant: No. MDA972-02-1-0024.

7. References

[1] Bishop, C., "Neural Networks for Pattern Recognition", Oxford University Press (1995).

[2] Chen, B., Zhu, Q., Morgan, N., "Learning Long Term Temporal Feature in LVCSR Using Neural Networks", ICSLP 2004.

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

[4] Houtgast, T., and Steeneken, H. J. M., "The modulation transfer function in room acoustics as a predictor of speech intelligibility.

[5] Zhu, Q., Chen, B., Morgan, N., and Stolcke, A., "On using MLP-based features in LVCSR," Proc. ICSLP, October 2004.

[6] Zhu, Q., Stolcke, A., Chen, B., Morgan, N., "Using MLP Features in SRI's Conversational Speech Recognition System", Proc. Interspeech 2005, accepted.