interspeech 2008 incorporating SST 08

# Transcribing Broadcast Data Using MLP Features

*Petr Fousek, Lori Lamel and Jean-Luc Gauvain*

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{fousek,lamel,gauvain}@limsi.fr

## Abstract

This paper describes incorporating discriminative features from a multi layer perceptron (MLP) into a state-of-the-art Arabic broadcast data transcription system based on cepstral features. The MLP features are based on a recently proposed Bottle-Neck architecture with long-term warped LP-TRAP speech representation at the input. It is shown that the previously reported improvements on a development Arabic transcription system carry through to a full system at a state-of-the-art level. SAT, CMLLR and MLLR adaptation techniques are shown to be useful for both MLP and combined features, though to a lesser degree than for PLPs. Without adaptation, MLP features obtain superior performance to cepstral features in all test conditions, and with adaptation both feature sets give comparable results. Combining the features, either by feature concatenation or system hypotheses, gives significant gains. Gains from MMI model training seem to be additive to the gain coming from discriminative MLP features.

**Index Terms**: MLP, LP-TRAP, broadcast transcription, bottleneck, discriminative training

## 1. Introduction

One of the recent trends in speech-to-text systems is using discriminative techniques with large corpora for more accurate acoustic modeling. Maximum likelihood training of Gaussian mixture HMMs is often being replaced by Maximum Mutual Information (MMI) or Minimum Phone Error (MPE) criteria and features are being modified by discriminatively trained transforms such as feature-level MPE [1]. An area of growing interest is incorporating the discriminative property in the feature extraction by using discriminative classifiers such as MLPs. By covering a wide temporal context MLP features can potentially capture different speech properties than the widely used cepstral features. In addition, MLPs can be trained to deliver estimates of class posteriors which can be used as features for Gaussian mixture acoustic model. Over the years, ICSI, SRI, UW and other groups have developed mature techniques for extracting probabilistic MLP features such as TRAPs, and have experience incorporating these MLP features in speech-to-text (STT) systems [2, 3]. One of the important properties of MLP features is their complementarity to cepstral features, it is thus desirable to know how to best include both feature types in a system.

This paper presents recent research into how MLP features can be efficiently incorporated in a state-of-the-art transcription system for broadcast data utilizing cepstral features. In addition to the more widely used 9 frames of PLP based features, time-warped linear predictive TRAP features [4] are used. To the best of our knowledge these latter features have yet to be incorporated in a state-of-the-art system. Although it is difficult to improve upon a competitive system, doing so can certainly increase the uptake of such novel technologies in the community. Since the MLP features used here differ from the better known ones, it is of interest to explore suitable ways of combining them with cepstral features. A four-layer Bottle-Neck MLP architecture [5] is used to deliver two types of MLP features differing in their input speech representations. Extending our previous work on a development task [6], acoustic models are trained on both PLP and MLP features as well as a combination of the two. This paper incorporates MLP features in an Arabic STT system and examines how the MLP features compare to cepstral ones, how both features combine, how the system performance is dependent on the amount of training data, and how the acoustic models utilizing MLP features can benefit from discriminative training and from model adaptation. Experiments are carried out with large training of a full system, applying model adaptation techniques such as speaker adaptive training (SAT), Constrained Maximum Likelihood Linear Regression (CMLLR) and MLLR.

This work was in part supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022 and in part by OSEO under the Quaero program.

## 2. Task & System Overview

The speech recognizer was derived from the LIMSI Arabic speech-to-text component system used in the AGILE participation in the GALE'07 evaluation. The transcription system has two main parts, an audio partitioner and a word recognizer [7]. The audio partitioner is based on an audio stream mixture model, and serves to divide the continuous stream of acoustic data into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment. The recognizer makes use of continuous density HMMs for acoustic modeling and $n$-gram statistics for language modeling. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained.

Word recognition is performed in one or more passes, where each decoding pass generates a word lattice with cross-word, position-dependent, gender-independent acoustic models, followed by consensus decoding with 4-gram and pronunciation probabilities [7, 8]. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [9] techniques prior to the next decoding pass.

All of the available manually transcribed Arabic broadcast news and broadcast conversation data distributed by the Linguistic data consortium was used to train acoustic models. This corpus contains over 1380 hours of raw data, with roughly 730

| *ID* | *Raw features (#)* | *HMM features (#)* |
|---|---|---|
| PLP | PLP (13) | PLP+$\Delta$+$\Delta^2$ (39) |
| $MLP_{9xPLP}$ | 9x(PLP+$\Delta$+$\Delta^2$) (351) | MLP (39) |
| $MLP_{wLP}$ | wLP-TRAP (475) | MLP (39) |

Table 1: *Naming conventions for MLP features and how the raw input features relate to the features for HMM.*

| *MLP train set* | *bnat06 WER (%)* |
|---|---|
| 17 hrs | 24.7 |
| 63 hrs | 24.2 |
| 300 hrs | 23.4 |
| 1200 hrs | 22.2 |
| PLP baseline | 25.1 |

Table 2: *Word error rates on the bnat06 data set as a function of the amount of data use to train the $MLP_{9xPLP}$. All the HMMs are trained on 300 hours of speech. Single decoding pass with a 4-gram LM, no adaptation, no MLLT, no MMIE.*

hours of broadcast news and 550 hours of broadcast conversations. After removing non-speech portions (music, publicity) and portions that fail forced alignment, there are about 1250 hours of data used for HMM training. We refer to this corpus as the 1200 hour training set. These data were used to train the baseline gender-independent acoustic models, without maximum-likelihood linear transform (MLLT) or speaker-adaptive training (SAT). The models cover 44k contexts with 11.5k tied states, and have 32 Gaussians per state.

Various language models were trained on corpora comprised of 11 million words of audio transcriptions and 1 billion words of texts from a wide variety of sources. The recognition word list contains either 200k or 290k non-vocalized, normalized entries. The language models result from the interpolation of models trained on subsets of the available data, with the interpolation weights optimized on the combined GALE development data from 2006 and 2007. The largest coefficients are associated with the audio transcriptions, accounting for almost half the LM weight, even though these texts represent only about 1% of the available data. This highlights the importance of audio transcripts for language model training. Results are reported using word-based language models and language models estimated on morphologically decomposed texts [10]. For multipass decoding, lattices are rescored by a neural network LM [11] interpolated with a 4-gram backoff LM. The pronunciation lexicon is represented with 71 symbols, including 31 simple consonants, 30 geminate consonants, 3 long and 3 short vowels, a generic vowel plus 3 pseudo phones for non-linguistic events (breath, filler, silence).

Results are reported for several sets of test data used in the GALE community, where each set contains about 3 hours of broadcast news (bn) or broadcast conversation (bc) data. These test sets are referred to in the GALE community as bnad06, bnat06, bcad06, bcat06, eval06, dev07, eval07. The last three test sets contain both bn and bc data. The out-of-vocabulary rate with this word list is about 1%, and the devset perplexities with a 4-gram language model are about 790 for dev06 and 430 for dev07.

## 3. Training MLP Features

Neural network feature extraction consists of two steps. The first step is *raw feature extraction* which constitutes the input to the MLP. Typically this vector covers a wide temporal context (100–500 ms) and therefore is highly dimensional. Second, the raw features are processed by the MLP followed by a PCA transform to yield the HMM features.

| *MLP train set* | *MLP parameters* | *WER(%)* |
|---|---|---|
| 63 hrs | 1.4M | 24.2 |
| 63 hrs | 6.5M | 24.4 |
| 1200 hrs | 1.4M | 22.2 |
| 1200 hrs | 5.3M | 21.9 |

Table 3: *Influence of the MLP size on performance for two different quantities of data use to train the $MLP_{9PLP}$. Results on the bnat06 data. All the HMMs are trained on 300 hours of speech. Single decoding pass with a 4-gram LM, no adaptation, no MLLT, no MMIE.*

**Raw features:** Two different sets of raw features are used which cover different temporal contexts: 9 frames of PLPs (9xPLP) and time-warped linear predictive TRAP (wLP-TRAP) [4]. The 9xPLP set is based on the PLP features from the baseline system which are mean and variance normalized per speaker. The raw features are formed by 9 neighboring frames of PLPs (12 coefficients plus energy, with derivatives $\Delta$ and $\Delta^2$), centered at the current frame. The feature vector has $9 \times 39 = 351$ values and covers a 150 ms window. The wLP-TRAP raw features are obtained by warping the temporal axis in the LP-TRAP feature calculation. Linear prediction is used to model the Hilbert envelopes of 500 ms long energy trajectories in auditory-like frequency sub-bands [12]. 25 LPC coefficients in 19 frequency bands form the raw features, yielding $19 \times 25 = 475$ values which cover a 500 ms window. wLP-TRAPs are not derived from the same sub-band energies sampled at 100Hz rate as PLPs so they have a potential of producing more complementary features to PLPs than 9-PLPs or TRAPs, which is an advantage for feature combination. The adopted naming conventions are given in Table 1 along with how the raw features relate to the features for HMM.

**MLP architecture:** The MLP architecture is based on a four layer bottle-neck network with an input layer, two hidden layers and an output layer. The input layer distributes the raw features in the second layer, which is large in order to provide the necessary modeling power. The third layer is small, its size is equal to the required number of features, which in this work was fixed to 39 for easy comparison with PLP features. The output layer computes the estimates of the target class posteriors. The classes are context independent phone states obtained from a HMM automatic alignment which were shown to outperform phone targets [5]. There are 69 three-state phones and 3 units with all states merged (silence, filler, breath) resulting in 210 target classes. The outputs of the small hidden layer neurons (prior to a sigmoid function) are decorrelated by a PCA transform and used as final features. Note that this MLP structure allows the feature vector size to be arbitrarily chosen, independent of the number of MLP targets.

**MLP training data:** The MLP features were trained on what we call the 1200 hour train set. Since the MLP features make use of a temporal context up to 500ms, the frames coming from 250ms segment boundaries are not used for training (except for providing a context), thus the available MLP train data account for 1168 hours. It is known that more data and/or more parameters in the MLP help, but at certain point the gain is not worth the effort. Table 2 gives the word error rate as a function of the amount of MLP training data for a MLP with a fixed number (1.4 million) of parameters with 9xPLP raw features. The HMMs were always trained on the 300 hour training set

| *MLP train set* | *63 hrs* | *300 hrs* |
|---|---|---|
| $MLP_{9PLP}$ | 24.2 | 23.4 |
| $MLP_{wLP}$ | 25.8 | 23.5 |
| $PLP+MLP_{9PLP}$ | 22.7 | 22.5 |
| $PLP+MLP_{wLP}$ | 21.7 | 21.3 |

Table 4: *Performance on the bnat06 data set of two types of MLP features, stand-alone or concatenated with PLP as a function of the amount of data used to train the MLP. All the HMMs are trained on 300 hours of speech. Single decoding pass with a 4-gram LM, no adaptation, no MLLT, no MMIE.*

and evaluated on bnat06 dev data. The MLP performance is seen to improve with the additional data, and no saturation is observed. The WER of the baseline PLP system (single pass decoding with speaker-independent models, no SAT, no MLLT, no MMIE and no adaptation) trained on the 300 hour train set is 25.1%.

**Training process:** Training a MLP on over thousand hours of speech required two modifications to the training process performed by QuickNet software. First, the storage space requirements of the raw features were reduced by almost a factor of four by using linear quantization of 32 bit float values to 8 bits, with no impact on performance. Once the MLP is trained, output features are created using non-compressed raw features. Second, to reduce the computation time of MLP training, a simplified training scheme is adopted from [2]. Instead of iterating all training data 7 to 12 times through the MLP as determined by cross-validation performance, a fixed number of 6 epochs with fixed learning rates is used. In addition, the data are randomized and split in three non-overlapping subsets of 13%, 26%, and 52% of frames. First three epochs are trained on 13% of data, two subsequent epochs use 26% of the data, the last epoch uses 52% of the data, and the remaining data is used for monitoring the performance. This reduces the training time by a factor of 5.4 with a minor impact on performance (in fact, simulations on the 300 hrs set even improve from 24.4% to 24.2% WER on the bnat06 data for 9xPLP raw features, with unadapted models). All these modifications lead to about one week training time on the 1200 hour train set using one four-threaded computer, and the wLP-TRAP raw training features occupy 200GB of space.

**MLP size:** To get the most benefit from the larger amount of training data may require using a more complex model. An experiment was carried out by enlarging the first hidden layer in the MLP in order to raise the number of free parameters, as shown in Table 3. For the small 63 hour train set, the larger MLP degraded performance, while for the full 1168 hours it brought a 1.6% relative improvement. However, such a gain was not judged to be worth the almost 4 times longer MLP training time, so further experiments used the smaller MLP.

## 4. Using MLP Features

This section presents contrastive results starting with the baseline system, and going to more complex models and decoding strategies typically used in state-of-the-art systems.

Table 4 compares performances of the MLP features when used stand-alone and when concatenated with PLP features at the input to the HMM system as a function of the amount of data used to train the MLP. Note that the concatenated vector has 78 features, whereas the stand-alone vector has 39 features. HMMs were all trained on the 300 hour data set. For all feature sets there is a significant WER reduction when the MLP training data is increased from 63 to 300 hours. The results with the two types of MLP features stand-alone are comparable when 300 hours are used to train the MLPs. HMMs trained with both MLP features outperform the PLP baseline (25.1%). Concatenating the PLP features with the MLP ones gives the best performance (the last two entries), however the improvement from training the MLP on more data is less than for the systems using only MLP features (the top two table entries). The best results are obtained with the HMM trained on the $PLP+MLP_{wLP}$ features.

| bnat06 | *WER (%)* | | |
|---|---|---|---|
| *Features* | *300h* | *300h/1200h* | *1200h* |
| PLP | 22.7 | 21.8 | |
| $MLP_{9xPLP}$ | 21.8 | 21.3 | 20.3 |
| $MLP_{wLP}$ | 21.9 | 21.3 | 20.7 |
| $PLP + MLP_{9xPLP}$ | - | 20.4 | 19.9 |
| $PLP+MLP_{wLP}$ | 20.1 | 19.7 | 19.2 |

Table 5: *Performance of PLP and MLP features, and feature concatenation with a single decoding pass. The amount of data used to train the MLP/HMM are given in the column headers. Single decoding pass with an improved 290k 4-gram LM, improved pronunciation modeling, gender-dependent models, no adaptation, no MMIE, with MLLT for PLP.*

| bnat06 | *WER (%)* | | |
|---|---|---|---|
| *Features* | PLP | $MLP_{wLP}$ | $PLP + MLP_{wLP}$ |
| No adaptation | 21.8 | 20.7 | 19.2 |
| SAT+CMLLR+MLLR | 19.0 | 18.9 | 17.8 |

Table 6: *Performance on bnat06 with the improved 290k 4-gram LM for PLP and $MLP_{wLP}$ features, and feature concatenation without and with adaptation. gender-dependent models, no MMIE, and with MLLT for PLP. Both the MLP and HMM are trained on 1200 hours of data.*

The results presented in Table 5 use a system that has an improved 290k 4-gram LM, improved pronunciation modeling and gender-dependent models. The PLP-based system also has MLLT. The table further explores performance as a function of the amount of data used to train the MLP. In the first column both the MLP and HMM are trained on 300 hours. In the second column, the same MLP is used but the HMMs are trained on 1200 hours. Finally in the third column both the MLP and HMM are trained on 1200 hours. Again, the two MLP features are seen to provide comparable performance, with a slight advantage for the $MLP_{9xPLP}$ features with the larger HMM training. As already observed with HMMs trained on 300 hours of data (see Table 4), the best results are obtained with the concatenated features $PLP+MLP_{wLP}$. This feature set gives an absolute gain of 1.2-1.6% over any other of the features.

Table 6 compares three feature sets with the 290k 4-gram LM, the improved pronunciation modeling, and gender-dependent acoustic models. The first entry corresponds to a single unadapted decoding, and the second to a two-pass decoding using the standard techniques of SAT training, and CMLLR and MLLR adaptation. These results show that without adaptation the $MLP_{wLP}$ and concatenated $PLP+MLP_{wLP}$ features clearly outperform the PLP ones. However, with CMLLR and MLLR adaptation, only the concatenated features perform significantly better than the PLP.

The last set of experimental results were produced with a more complete system, including gender-dependent SAT, MMIE acoustic models (with MLLT for PLP) trained on 1200

| *AM* | *LM* | *bnat06* | *bnad06* | *bcat06* | *bcad06* | *eval06* | *dev07* | *eval07* |
|---|---|---|---|---|---|---|---|---|
| PLP | word | 16.7 | 15.5 | 22.8 | 20.4 | 19.3 | 12.4 | 13.7 |
| $MLP_{wLP}$ | word | 16.8 | 15.7 | 22.7 | 20.5 | 20.1 | 12.7 | 14.3 |
| PLP+$MLP_{wLP}$ | word | 15.4 | 14.3 | 21.1 | 18.6 | 18.4 | 11.6 | 13.0 |
| PLP $\oplus$ PLP+$MLP_{wLP}$ | word | 15.0 | 13.8 | 20.7 | 18.3 | 17.7 | 11.2 | 12.4 |
| PLP | morph. | 16.7 | 15.3 | 23.2 | 20.6 | 19.4 | 12.2 | 13.8 |
| PLP+$MLP_{wLP}$ | morph. | 15.7 | 14.3 | 21.9 | 19.2 | 18.6 | 11.6 | 12.9 |
| 4-way rover | both | 14.5 | 13.2 | 20.2 | 17.9 | 17.1 | 10.6 | 11.9 |

Table 7: *WER on various GALE data sets with broadcast news (bn) or broadcast conversation (bc) data. The eval06, dev07, eval07 sets contain both bn and bc data. The acoustic models are gender-dependent SA, MMI trained PLP and MLP models (also with MLLT for PLP) trained on 1200h of manually transcribed data, with word duration models. Multiple pass decoding with CMLLR and MLLR adaptation, a 290k 4-gram NN LM, and improved pronunciation models. Results in line 4 and 7 are obtained with 2-way and 4-way* ROVER *combinations.*

hours of manually transcribed data, with word duration models. It uses a multiple pass decoding strategy with CMLLR and MLLR adaptation, a word- or morph-based 290k 4-gram neural network (NN) language model, and improved pronunciation models. What we refer to as the NN LM results from the interpolation of a connectionist language model with a standard 4-gram backoff LM. Table 7 gives the word error rates for three acoustic models (PLP, $MLP_{wLP}$ and PLP+$MLP_{wLP}$ for seven GALE test sets, with two NN LMs (word based and with morphological decomposition), as well as some combinations using ROVER [13]. It can be seen that the PLP and $MLP_{wLP}$ based systems give comparable results, with small differences across test sets. Based on the combination experiments reported in [6], we selected a 2-way ROVER combining the PLP and the PLP+$MLP_{wLP}$ based systems, which gives an average gain of almost 0.5%. The results with the morphologically decomposed LM [10] are seen to be comparable to those with the word-based LM. A 4-way ROVER combination gives an additional 0.4% gain over the 2-way ROVER.

Although the performance of the PLP system has been improved from the baseline of 25.1% to 16.7% on the bnat06 data set (a relative WER reduction of 33%), the combined PLP+$MLP_{wLP}$ based system obtains a lower WER for all test sets, with an average gain of 1.2% absolute. There is a gain (over 4% absolute) with ROVER combination of the PLP and PLP+$MLP_{wLP}$ based systems, even though the PLP features are in there twice. We attribute this to the observation that unsupervised model adaptation is more effective for the PLP-based system than the MLP-based one.

## 5. Summary

This paper has explored incorporating novel MLP features, derived using the bottle-neck MLP architecture, in a state-of-the-art Arabic broadcast data transcription system. In particular the influence on performance of the amount of data used to train the MLP, the number of free parameters used when training the MLP, and the amount of data used for HMM training was assessed. Experiments were carried out on the Gale Arabic broadcast news task using multiple development data sets. When used without adaption, the MLP features have better performance than standard PLP features. However, once SAT training and CMLLR/MLLR adaptation are used, both feature types have comparable performance. Feature concatenation appears to be the most efficient combination method, providing the best gain at the lowest decoding cost. In general, it seems best to combine features based on different time spans as they provide high complementarity. Since the PLP based system improves more than the MLP based with unsupervised adaptation, an additional gain is obtained by combining a PLP based system with one based on the concatenated features and with ROVER combination using different language models. It also seems that gains from MMI model training are additive to the gain coming from discriminative MLP features.

## 6. Acknowledgment

The authors acknowledge the contribution of Abdel Messaoudi who was instrumental in developing the baseline PLP system.

## 7. References

[1] J. Zheng and al., "Combining dicsriminative feature, transform, and model training for large vocabulary speech recognition," in *ICASSP 2007*. Hononulu: IEEE, April 2007.

[2] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *INTERSPEECH 2005*, 2005, pp. 2141–2144.

[3] A. Stolcke and al., "Recent innovations in speech-to-text transcription at sri-icsi-uw," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, September 2006.

[4] P. Fousek, *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*. Prague: PhD thesis, Czech Technical Univ., Faculty Electrical Engineering, March 2007.

[5] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *ICASSP'08*, Las Vegas, 2008.

[6] P. Fousek, L. Lamel, and J. Gauvain, "On the Use of MLP Features for Broadcast News Transcription," in *TSD 2008*, September 2008.

[7] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[8] L. Lamel, A. Messaoudi, and L. G. J, "Improved Acoustic Modeling for Transcribing Arabic Broadcast Data," in *InterSpeech'07*, Antwerp, August 2007.

[9] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[10] L. Lamel, A. Messaoudi, and J. Gauvain, " Investigating Morphological Decomposition for Transcription of Arabic Broadcast News and Broadcast Conversation Data," in *Interspeech 2008*, September 2008.

[11] H. Schwenk, "Continuous space language models," *Computer Speech and Language*, vol. 21, pp. 492–518, 2007.

[12] M. Athineos, H. Hermansky, and D. P. Ellis, "LP-TRAP: Linear predictive temporal patterns," in *International Conference on Spoken Language Processing (ICSLP)*, 2004.

[13] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," Santa Barbara, pp. 347–352, 1997.