Combining MLP and PLP Features for Speech Transcription *

Petr Fousek, Lori Lamel and Jean-Luc Gauvain
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
p.fousek@gmail.com, {lamel,gauvain}@limsi.fr

Abstract

This contribution reports on work carried out in part under the GALE program to incorporate discriminative features from a multi layer perceptron (MLP) into an optimized Arabic broadcast data transcription system based on cepstral features. The recently proposed 4-layer Bottle-Neck MLP architecture (Grézl and Fousek, 2008) is explored and used to produce three types of MLP features differing in their input speech representations. Initial experiments carried out with a development transcription system (300 hour acoustic training) demonstrated that standard techniques used in state-of-the-art systems with PLP features (SAT, CMLLR, MLLR, MMI) could be successfully used with MLP features alone and in combination with PLP ones. Further studies extend the model training to the full set of available audio data (over 1380 hours). Experimental results are reported on GALE data to illustrate the influence of the different MLP features, the amount of data used to train the MLP and the HMMs, and the different means of combining the PLP and MLP features on the system performance.. The improvements obtained with MLP features have been confirmed on other tasks and languages.

1 Introduction

One of the recent trends in speech recognition is using discriminative techniques with large corpora for more accurate acoustic modeling. Maximum likelihood training of Gaussian mixture HMMs is often replaced by Maximum Mutual Information (MMI), Minimum Classification Error (MCE), or Minimum Phone Error (MPE) criteria and features are being modified by discriminatively trained transforms such as feature-level MPE (Povey et al., 2005). There has been growing interest in incorporating some discriminative estimation in the feature extraction by using discriminative classifiers such as multi layer perceptrons. Since MLP features cover a wide temporal context, they can potentially capture different speech properties than are captured by the widely used cepstral features. In addition, MLPs can be trained to deliver estimates of class posteriors which can be directly used as emission probabilities in Hidden Markov Models. Over the years, ICSI, SRI, UW and other groups have developed mature techniques for extracting probabilistic MLP features such as TRAPs (TempoRAl Patterns), and have substantial experience incorporating these MLP features in speech-to-text (STT) systems (Zhu et al., 2005; Stolcke et al., 2006). While probabilistic features have never been shown to consistently outperform cepstral features in LVCSR, they have been shown to improve performance when used in conjunction with them. The MLP and PLP features being complementary, an important consideration is determining the best manner to incorporate both in an STT system. Different means of multi-stream combination have been successfully used for this purpose, four of which were studied in (Fousek et al., 2008a).

This contribution summarizes the most important results obtained on incorporating MLP features in a

*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.

transcription system for broadcast data. Two types of raw features are used: 9 frames of PLP based features and time-warped linear predictive TRAP features (Fousek, 2007). To the best of our knowledge this was the first time that the latter features were incorporated in a state-of-the-art system. Since the MLP topology, as well as the speech representations at the MLP input used here differ from the better known ones, it is of interest to explore the properties of these features and suitable ways of combining them with cepstral ones.

The experiments reported here use a 4-layer Bottle-Neck MLP architecture (Grézl and Fousek, 2008) to deliver two types of MLP features differing in the speech representation used at the MLP input. Extending previous work with smaller amounts of audio data (Fousek et al., 2008a), acoustic models are trained with PLP and MLP features as well as their combination. Using a full state-of-the-art Arabic STT system trained on over 1380 hours of raw data with model adaptation techniques such as speaker adaptive training (SAT), Constrained Maximum Likelihood Linear Regression (MLLR) and MLLR, experiments were carried out to examine how the MLP features compare to cepstral ones, how both features combine, how the system performance is dependent on the amount of training data (for the MLP and the HMM), and how the acoustic models utilizing MLP features can benefit from discriminative training and from model adaptation.

2 Task & System Overview

The speech recognizer was derived from the LIMSI Arabic STT component system used in the AGILE participations in the GALE evaluations. The transcription system has two main parts, an audio partitioner and a word recognizer (Gauvain et al., 2002). 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. Word recognition is performed in one or more passes, where each decoding pass generates a word lattice with cross-word acoustic models, followed by consensus decoding with 4-gram language model (LM) and pronunciation probabilities (Gauvain et al., 2002; Lamel et al., 2007). Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR (Leggetter and Woodland, 1995) techniques between decoding passes.

The manually transcribed Arabic broadcast news (bn) and broadcast conversation (bv) data distributed by the Linguistic Data Consortium were used to train the acoustic models. There are over 1380 hours of raw data, with roughly 730 hours bn and 550 hours of bc. After removing non-speech portions (music, publicity) and portions that fail forced alignment, about 1250 hours of data remain for HMM training. This is referred to as the 1200 hour training set, used to train the baseline gender-independent acoustic models, covering 44k word position-dependent contexts with 11.5k tied states (32 Gaussians/state) (Fousek et al., 2008b). Some results are reported for a development system, using 300 hours of data for acoustic model training.

Various language models were trained on corpora comprised of 11 M words of audio transcriptions and 1 B words of texts from a wide variety of sources. 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 coefficients associated with the audio transcriptions are assigned almost half the LM weight, even though these texts represent only about 1% of the available data. Language models were estimated on the normalized texts and morphologically decomposed texts (Lamel et al., 2008). For multipass decoding, lattices are rescored by a neural network LM (Schwenk, 2007) interpolated with a 4-gram backoff LM. Results are reported for several GALE data sets, where each set contains about 3 hours of broadcast data.

3 Training MLP Features

The neural network feature extraction has 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.

ID	*Raw features (#)*	*HMM features (#)*
PLP	PLP (13)	PLP+Δ+Δ^2 (39)
$\text{MLP}_{\text{9xPLP}}$	9x(PLP+Δ+Δ^2) (351)	MLP (39)
MLP_{wLP}	wLP-TRAP (475)	MLP (39)
MLP_{comb}	9x(PLP+Δ+Δ^2) + wLP-TRAP (826)	MLP (39)

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

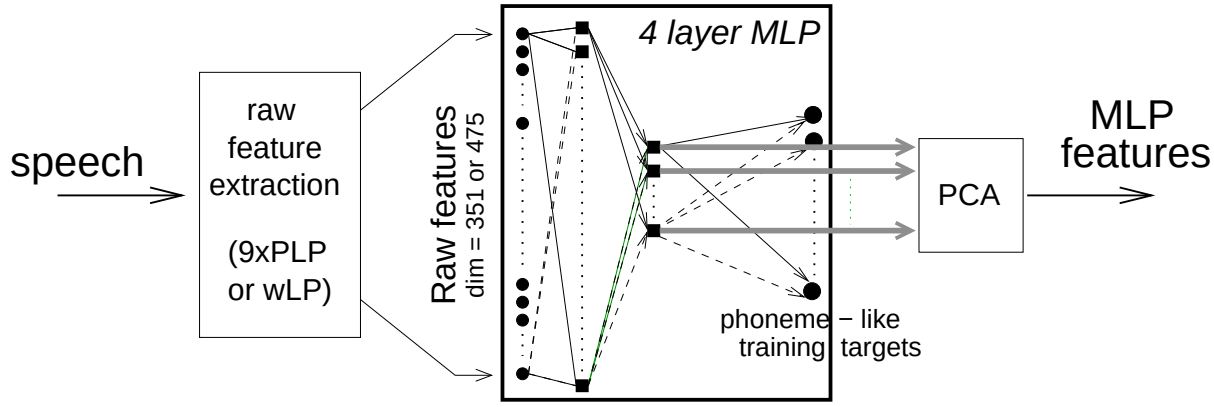

Figure 1: MLP feature computation with a four layer bottle-neck network, using phoneme-like training targets.

3.1 Raw features

Two sets of raw features are used which cover different temporal contexts: 9 frames of PLP (9xPLP) and *time-warped linear predictive* TRAP (wLP-TRAP) (Fousek, 2007). 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 PLP (12 coefficients plus energy, with derivatives Δ and Δ^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 (Athineos et al., 2004). 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 use a different bank of filters than are used for the PLP and they do not apply a short-term FFT, so they have a potential of producing more complementary features to PLP than 9-PLP or TRAPs, which should be an advantage for feature combination. The adopted naming conventions for the raw features and the HMM features derived from them are given in Table 1. The last feature set is obtained by combining the 9xPLP and the wLP-TRAP features at the input to the MLP and produces a 39-parameter feature vector.

3.2 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, as shown in Figure 1. 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.[1] 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 (Grézl and Fousek, 2008). There are 69 three-state phones and 3 units with all states merged (silence, filler, breath) resulting in 210 target

[1]The optimal number of features was explored in (Grézl and Fousek, 2008).

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 HMMs trained on 300 hours of speech. Single decoding pass with 200k 4-gram LM, no adaptation, no MLLT, no MMIE.*

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 MLP size on performance for two quantities of data use to train the MLP_{9PLP}. Results on bnat06 data. All HMMs trained on 300 hours of speech. Single decoding pass with 200k 4-gram LM, no adaptation, no MLLT, no MMIE.*

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 architecture allows the feature vector size to be arbitrarily chosen, independently of the number of MLP targets.

3.3 MLP training data

The MLP features were trained on the 1200 hour train set. Since the MLP features make use of a temporal context up to 500ms, the frames from first and last 250ms of each segment are not used for training (except for providing a context), thus the data available to train the MLP is 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 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 Maximum Likelihood Linear Transform (MLLT), no Maximum Mutual Information Estimation (MMIE) and no adaptation) trained on the 300 hour train set is 25.1%.

3.4 Training process

Training a MLP on over thousand hours of speech required two modifications to the training process performed by QuickNet [2] 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 was adopted after (Zhu et al., 2005). Instead of iterating on all training data about 10 times through the MLP as determined by cross-validation performance, a fixed number of 6 epochs with fixed learning rates is used with subsets of the data. The data are randomized and split in three non-overlapping subsets of 13%, 26%, and 52% of the frames. First, three epochs are trained on 13% of the data, then two subsequent epochs use 26% of the data, the last epoch uses 52% of the data, with the remaining data used to monitor the performance. This reduced the training time by a factor of 5.4 with a negligible impact on performance (in fact, tests with the 300 hour set even improved from 24.4% to 24.2% WER on the bnat06 data for the 9xPLP raw features, with unadapted models). All these modifications reduced the training time to about one week on the 1200 hour train set using one four-threaded computer, and the wLP-TRAP raw training features occupy 200GB of disk space.

3.5 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

[2] http://www.icsi.berkeley.edu/Speech/qn.html

#	bnat06 *Features*	*WER (%)* *1-pass*	*2-pass*
1	PLP	25.1	22.5
2	$\text{MLP}_{\text{9xPLP}}$	24.2	22.7
3	MLP_{wLP}	25.8	23.1
4	MLP_{comb}	23.8	21.9
5	PLP + $\text{MLP}_{\text{9xPLP}}$	22.7	21.2
6	PLP + MLP_{wLP}	21.7	20.4
7	$\text{MLP}_{\text{9xPLP}}$ + MLP_{wLP}	22.2	21.0

Table 4: *Performance of PLP and MLP features, MLP combined features and feature concatenation, without and with unsupervised acoustic model adaptation on the bnat06 data and 200k 4-gram ML. MLPs trained on 63 hours of data. HMMs trained on 300 hours of speech.*

bnat06 *Combined systems*	*WER (%)* *1-pass*	*2-pass*
$3 \rightarrow 1$	25.8	21.5
$1 \rightarrow 3$	25.1	22.0
$7 \rightarrow 1$	22.2	20.7
$1 \rightarrow 7$	25.1	21.2
$1 \oplus 2 \oplus 3$	22.3	20.6
$1 \oplus 3$	23.3	21.0
$5 \oplus 6$	21.2	19.9
$1 \oplus 6 \oplus 7$	21.0	19.7

Table 5: *Comparing cross-adaptation and* ROVER *for combining multiple systems on bnat06 data.*

the number of free parameters. The results are given 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 computation cost (almost 4 times longer to train the MLP), so further experiments used the smaller MLP with 1.4 million parameters.

4 Using MLP Features

This section presents contrastive results starting with the baseline system, and going to more complex models and decoding strategies.

4.1 Experiments with a small system

A first series of experiments were carried out to compare the four basic features from Table 1 without and with unsupervised acoustic model adaptation, as shown in Table 4. The 1-pass results use a 4-gram LM, no adaptation, no MLLT, no MMIE. The MLPs were all trained on 63 hours of data, and all HMMs were trained on 300 hours of speech. The baseline performance of the standard PLP features without adaptation is 25.1% and with adaptation is 22.5%. Without adaptation, the $\text{MLP}_{\text{9xPLP}}$ features are seen to perform a little better (about 4% relative) than PLP, but with adaptation both $\text{MLP}_{\text{9xPLP}}$ and MLP_{wLP} are slightly worse than PLP. This leads to the conclusion that MLLR adaptation is less effective for MLP features than for PLP features.

Different means of fusing the information coming from the cepstral and the MLP features were investigated. The MLP_{comb} results are for combination at the input to the MLP. These models give the best results with 39 parameters. A simple approach is to concatenate together the features at the input to the HMM system (this doubles the size of the feature vector, $2 \times 39 = 78$ features) and to train an acoustic model. Three possible pairwise feature concatenations were evaluated and the results are given in the lower part of Table 4. These concatenated features all substantially outperform the PLP baseline, by up to 9% relative, showing that feature concatenation is a very effective approach. Given the significantly better performance of the PLP+MLP_{wLP} features over the PLP+$\text{MLP}_{\text{9xPLP}}$ and $\text{MLP}_{\text{9xPLP}}$ + MLP_{wLP} features, the three-way concatenation was not tested as it was judged to be not worth the increased computational complexity needed to deal with the resulting feature vector size (3×39).

Two other more computationally expensive approaches were studied, cross model adaptation and ROVER (Fiscus, 1997). Table 5 gives some combination results using cross adaptation (top) and ROVER (bottom). The first entry is the result of adapting the PLP models with the hypotheses of the MLP_{wLP} system. The second entry corresponds to the reverse adaptation order, i.e. the MLP_{wLP} are adapted using the hypotheses of the PLP system. The next two entries use cross adaptation on top of feature concatenation. In the first 3 cases, cross adaptation reduces the WER (note that the 2nd pass er-

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 6: *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 HMMs trained on 300 hours of speech. Single decoding pass with 200k 4-gram LM, no adaptation, no MLLT, no MMIE.*

ror rates must be compared with those in Table 4). Larger gains are obtained when the PLP models are used in the second pass, supporting the earlier observation that MLLR adaptation is more effective for PLP features than for MLP features. This may be because the MLP already removes the variability due to the speaker or because other, perhaps non-linear, transformations are needed to adapt MLP features. The WERs in the bottom part of the table result from ROVER combination of the first or second pass hypotheses of the listed systems. ROVER combination of the three basic features performed better than the best pair-wise cross-adaptation amongst them ($3 \rightarrow 1$) however, neither combination outperformed the simple feature concatenation WER of 20.4% (entry 6 in Table 4). ROVER also helps when applied jointly with other combination methods (see the last two rows in Table 5), beating the baseline PLP system by up to 12% relative. This best ROVER result however requires 6 decoding passes.

It is interesting to observe that the PLP features are generally best combined with MLP_{wLP}, even though the MLP_{9xPLP} gives better score than MLP_{wLP}. This may be due on one side to the fact that the MLP_{9xPLP} features are derived from the PLP, and on the other side that there is a larger difference in time spans between the standard PLP and the wLP-TRAP features.

4.2 Amount of MLP training data

Table 6 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.

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 7: *Performance of PLP, MLP and concatenated features. 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.*

Note that the concatenated vector has 78 features, whereas the stand-alone vector has 39 features. The 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.

Table 7 gives results with an updated system: it has an improved 290k 4-gram LM, improved pronunciation modeling and gender-dependent models. The PLP-based system also has MLLT. The table summarizes further exploration of performance as a function of the amount of data used to train the MLP. In column *300h*, both the MLP and the HMM are trained on 300 hours. In the second column, the same MLP is used, but the HMMs are trained on 1200 hours. Finally last column both the MLP and HMM are trained on 1200 hours. Comparable performance is seen for both MLP features, 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 6), the best

bnat06	*WER (%)*	
Features	*No adapt.*	*SAT+CMLLR+MLLR*
PLP	21.8	19.0
MLP_{wLP}	20.7	18.9
PLP + MLP_{wLP}	19.2	17.8

Table 8: *Performance on bnat06 with 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. MLP and HMM both trained on 1200 hours of data.*

results are obtained with the concatenated features PLP+MLP_{wLP}. This feature set gives an absolute gain of 1.2-1.6% over all other features.

Table 8 compares three feature sets with the improved system. The first entry corresponds to a single pass 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 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 word duration models and MLLT for PLP, trained on 1200 hours of manually transcribed data. It uses a multiple pass decoding with CMLLR and MLLR adaptation, a word- or morph-based 290k 4-gram neural network (NN) language model, and improved pronunciation models. What is referred to as the NN LM results from the interpolation of a connectionist language model with a standard 4-gram back-off LM. Table 9 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 ROVER (Fiscus, 1997) combinations. 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 (Fousek et al., 2008a), 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 (Lamel et al., 2008) 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. 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 is seen to obtain a lower WER for all test sets, with an average gain of 1.2% absolute. ROVER combination of the PLP and PLP+MLP_{wLP} based systems gives a gain of over 4% absolute, even though the PLP features are in there twice.

5 Summary

This contribution 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 in the MLP, and the amount of data used for HMM training was assessed. Different means of combining MLP and cepstral features were also explored. Experiments were carried out on the GALE Arabic broadcast task using multiple 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. 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.

Recently speech recognizers have been trained using the combined PLP + wLP-TRAP features for broadcast data transcription in Dutch, Flemish,

AM	*LM*	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
PLP	word	16.7	15.5	22.8	20.4	19.3	12.1	13.5
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 9: *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 lines 4 and 7 are obtained with 2-way and 4-way* ROVER *combinations.*

French and Mandarin, and have observed comparable system behavoirs and performance gains. An attempt was made use an MLP developed for one language to produce features for another, but this was not successful, suggesting that the MLP features are capturing language-specific information. These features were also found to improve performance for the transcription of conversational telephone speech in Dutch and Flemish (Despres et al., 2008).

One of the difficulties in carrying out such experiments with a full system, is that generating the time-warped linear predictive TRAP features and training the MLP are quite computationally expensive. Some initial experiments have been carried out with other features, Multi-RASTA (MR) and TRAP-DCT (TD), that are much less costly to obtain (Hermansky and Fousek, 2005). The TRAP-DCT features are obtained from a 19-band Mel scale spectrogram, using a 30 ms window and a 10 ms frame step. A discrete cosine transform (DCT) is applied to each band, resulting in 475 raw features, which are fed to a 4-layer MLP (bottleneck architecture). Using the small training, the MLP_{TD} features give a word error rate of 24.4% alone and 21.4% when combined with PLP features on the bnat06 data set, which are comparable to results reported in Table 6 for the MLP_{wLP} features. With full training on 1200 hours, both feature sets obtain comparable results (13.7 PLP+MLP_{wLP} and 13.6 PLP+MLP_{TD}), outperforming standard PLP features (14.5%) on the dev08 data. This result is very interesting since the raw TD features are not much more costly to compute than the PLP ones.

Acknowledgments

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

References

Marios Athineos, Hynek Hermansky, and Daniel P.W. Ellis. 2004. LP-TRAP: Linear predictive temporal patterns. In *ICSLP'04*, pages 1154–1157, Jeju, KR.

Julien Despres, Petr Fousek, Jean-Luc Gauvain, Sandrine Gay, Yvan Josse, Lori Lamel, and Abdel Messaoudi. 2008. The LIMSI-Vecsys Research Systems for N-Best 2008. In *N-Best workshop*, Soesterberg, NL.

Jon Fiscus. 1997. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 347–352, Santa Barbara, CA.

Petr Fousek, Lori Lamel, and Jean-Luc Gauvain. 2008a. On the use of MLP features for broadcast news transcription. In *TSD'08*, volume 5246 of *Lecture Notes in Computer Science*, pages 303–310.

Petr Fousek, Lori Lamel, and Jean-Luc Gauvain. 2008b. Transcribing Broadcast Data Using MLP Features. In *Interspeech'08*, pages 1433–1436, Brisbane, AU.

Petr Fousek. 2007. *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*. Ph.D. thesis, Czech Technical University in Prague, Faculty of Electrical Engineering.

Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. 2002. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108.

František Grézl and Petr Fousek. 2008. Optimizing bottle-neck features for LVCSR. In *ICASSP'08*, pages 4729–4732.

Hynek Hermansky and Petr Fousek. 2005. Multi-resolution RASTA filtering for TANDEM-based ASR. In *Interspeech'05*, pages 361–364, Lisboa, PT.

Lori Lamel, Abdel. Messaoudi, and Jean-Luc Gauvain. 2007. Improved Acoustic Modeling for Transcribing Arabic Broadcast Data. In *Interspeech'07*, pages 2077–2080, Antwerp, BE.

Lori Lamel, Abdel. Messaoudi, and Jean-Luc Gauvain. 2008. Investigating Morphological Decomposition for Transcription of Arabic Broadcast News and Broadcast Conversation Data. In *Interspeech'08*, pages 1429–1432, Brisbane, Australia.

Chris Leggetter and Phil Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185, April.

Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau, and Geoffrey Zweig. 2005. fMPE: Discriminatively trained features for speech recognition. In *ICASSP '05*, volume 1, pages 961–964.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21:492–518.

A. et al. Stolcke, Barry Chen, H. Franco, Venkata, M. Graciarena, Mei-Yuh Hwang, K. Kirchhoff, A. Mandal, N. Morgan, Xin Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, Wen Wang, Jing Zheng, and Qifeng Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(5):1729–1744.

Qifeng Zhu, Andreas Stolcke, Barry Y. Chen, and Nelson Morgan. 2005. Using MLP features in SRI's conversational speech recognition system. In *Interspeech'05*, pages 2141–2144.