A Comparative Large Scale Study of MLP Features for Mandarin ASR

Fabio Valente[1], *Mathew Magimai Doss*[1], *Christian Plahl*[2], *Suman Ravuri*[3], *Wen Wang*[4]

[1]IDIAP Research Institute, CH-1920 Martigny, Switzerland
[2] Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany
[3] International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704
[4] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.

fabio.valente@idiap.ch,mathew@idiap.ch, plahl@cs.rwth-aachen.de
ravuri@icsi.berkeley.edu, wwang@speech.sri.com

Abstract

MLP based front-ends have shown significant complementary properties to conventional spectral features. As part of the DARPA GALE program, different MLP features were developed for Mandarin ASR. In this paper, all the proposed front-ends are compared in systematic manner and we extensively investigate the scalability of these features in terms of the amount of training data (from 100 hours to 1600 hours) and system complexity (maximum likelihood training, SAT, lattice level combination, and discriminative training). Results on 5 hours of evaluation data from the GALE project reveal that the MLP features consistently produce relative improvements in the range of $15\% - 23\%$ at the different steps of a multipass system when compared to the conventional short-term spectral based features like MFCC and PLP. The largest improvement is obtained using a hierarchical MLP approach.

Index Terms: TANDEM features, Multi-Layer Perceptron, Acoustic features, GALE project, LVCSR.

1. Introduction

Multi-Layer Perceptron (MLP) based front-ends originally proposed in [1], have evolved in different fashions including the use of long signal time span at the MLP input [2], the combination of multiple MLP outputs based on probabilistic rules (a.k.a. multi stream approaches) [3] and the use of more complex architectures as compared to the conventional three-layer MLP [2], [4]. Furthermore considerable improvements have been reported in LVCSR systems whenever they are used in concatenation with MFCC or PLP features (for a review see [5]). More recently, MLP features have been applied in a number of different systems and languages like Mandarin and Arabic, e.g., see [4],[6],[7],[8].

As part of the DARPA GALE[1] program, the development of different types of MLP features evolved along with the development of ASR systems. The first contribution of this work is to compare MLP front-ends used in a Mandarin ASR system in a systematic manner, i.e., using the same phoneme set, same speech-silence segmentation system, amount of training data, and number of free parameters. Towards this end, section 2 briefly describes the different MLP features and benchmarks them using a system trained on 100 hours of data. In contrast to [9], the comparison covers all the MLP front-ends integrated in the latest GALE Mandarin evaluation system. As an outcome, two competitive feature sets are obtained and referred to as MLP1 and MLP2, which provide a relative reduction of 17% Character Error Rate compared to standard spectral features.

The second part of the paper, i.e., section 3, investigates how the previously described improvements scale up with the amount of training data (from 100 hours to 1600 hours of training data) and with more complex LVCSR systems that include Speaker Adaptative Training (SAT), lattice level combination and discriminative training. The study is done using the RWTH Mandarin LVCSR system [8]. Contrary to our previous related works, the contrastive experiments are obtained using a *full* multipass LVCSR system trained with and without the MLP features on the entire 1600 hours of the training set. The results are then summarized and discussed in section 4 which concludes the paper.

2. Small scale experiments

These studies are based on a simplified version of the large vocabulary ASR system for transcription of the Mandarin language described in [7], developed by SRI/UW/ICSI for the GALE project. Recognition is performed using the SRI Decipher recognizer and results are reported in terms of Character Error Rate (CER). The training is done using approximatively 100 hours of broadcast news and conversational data manually transcribed including speaker labels. Results are reported on the DARPA GALE evaluation 06 data (eval06). The baseline system uses 13 standard MFCC plus a smoothed log-pitch estimate, as described in [10], augmented with first and second order temporal derivatives resulting in an acoustic vector of dimension 42. Vocal Tract Length Normalization (VTLN) and speaker level mean-variance normalizations are applied. The training consists of conventional Maximum Likelihood training. Details on acoustic modeling and language modeling can be found in [7]. The decoding phase consists of two decoding passes, a maximum likelihood speaker independent (si) decoding followed by a speaker adapted (sa) decoding. Speaker adaptation is done using 3-class constrained Maximum Likelihood Linear Regression (CMLLR). Performance of this baseline system on eval06 data is 27.8% CER for the speaker independent (si) and 25.8% CER for the speaker adapted (sa) models.

In the following section, we experiment with different MLP features obtained equalizing the total number of parameters (unless the contrary is stated) in order to obtain a fair comparison. The training is done using the same toneme set composed of 71 tones and using the same alignment. After PCA, a dimensionality reduction accounting for 95% of the total variability is applied. Let us briefly detail the different features.

[1] http://www.darpa.mil/ipto/programs/gale/gale.asp

Table 1: Summary of feature performances on eval06 data. Results are reported using MLP features only and in concatenation with MFCC+f0 for the speaker independent/adapted system. In brackets the relative improvement w.r.t. the baseline is reported. The two front-ends that produce the largest improvements are based on the multi-stream approach (MLP1) and on the augmented hierarchical MRASTA (MLP2).

Features	MLP w/o MFCC+f0		MLP with MFCC+f0	
	CER (si)	CER (sa)	CER (si)	CER (sa)
TANDEM-9framesPLP	27.6 (+0%)	25.5 (+1%)	23.4 (+15%)	22.2 (+13%)
HATS	30.5 (-9%)	29.1 (-12%)	23.8 (+14%)	22.7 (+12%)
Multi-stream (MLP1)	**24.6 (+11%)**	**23.1 (+10%)**	**22.8 (+18%)**	**21.7 (+16%)**
MRASTA	32.4 (-16%)	30.7 (-19%)	24.4 (+12%)	23.1 (+10%)
Hier	27.8 (+0%)	26.5 (-2%)	22.9 (+17%)	21.9 (+15%)
A-Hier (MLP2)	**26.4 (+5%)**	**24.1 (+6%)**	**22.3 (+20%)**	**21.2 (+17%)**

TANDEM- 9 frames PLP features: The input to the three-layer MLP consists of 9 consecutive frames of PLP features obtained after VTL normalization augmented with first and second order temporal derivatives. Furthermore this representation is augmented with 9 consecutive frames of the log pitch estimate (f0) with its temporal derivatives. Speaker level mean and variance normalization are performed. This produce a 42 × 9 dimensional input feature vector. Their performance is reported in Table 1.

Hidden Activation TRAPS (HATS) features: The HATS feature extraction [2] aims at including information from relatively long signal time spans. At first the 19 critical band auditory spectrum of the speech signal is extracted. In a first stage, a separate MLP is trained for each critical band in order to classify phonetic targets. The input for each MLP is represented by 51 consecutive log critical band energy vectors corresponding to 500 ms of speech. In a second stage a merger MLP is trained with the hidden activations obtained (for the training data) from the 19 MLPs in the first stage as input feature. This process produces a single posterior stream out of the 19 different estimates obtained at the previous stage. The phoneme probability estimates obtained at the output of the second stage MLP are then used for TANDEM features. Their performance is reported in Table 1.

Multi-stream features: MLP outputs represent phoneme posterior probabilities and they can be combined according to probabilistic rules. This approach is known as Multi-stream ASR [3]. Phoneme posterior estimates obtained using TANDEM-PLP and HATS that correspond respectively to short and long temporal context are combined using the Dempster-Shafer (DS) method [11][2]. Their performance is reported in Table 1.

Multiple RASTA (MRASTA) features is an extension of RASTA filtering introduced in [12]. Similarly to HATS, it uses the 19 critical bands auditory spectrum. A 600 ms long temporal trajectory in each critical band is filtered with a bank of Gaussian derivative filters aiming at dividing the available modulation frequency range into its individual sub-bands. Identical filters are used for all critical bands. After MRASTA filtering, frequency derivatives across three consecutive critical bands are introduced. The total number of features at the MLP input is 432. Similarly to HATS, MRASTA aims at using long signal time spans as input to the MLP. Their performance is reported in Table 1.

Hierarchical MRASTA (Hier) features: Previous studies on English and Mandarin data [13],[9] showed that significant gains can be obtained combining classifiers trained on separate ranges of modulation frequencies in hierarchical fashion. This represents the main motivation for the Hierarchical MRASTA processing in which the filter-banks are split in two separate filter banks that filter respectively fast and slow modulation frequencies. The cutoff frequency for both filter-banks is approximatively 10 Hz. The output is then processed according to a hierarchy of two MLPs progressively moving from high to low modulation frequencies. Details can be found in [13]. Their performance is reported in Table 1.

The MRASTA filtering can be also augmented with the value of the critical band energy and the smoothed log-pitch estimates. We refer to this set of features as Augmented Hierarchy (A-Hier). Their performance is reported in Table 1.

Results of the different MLP features are summarized in Table 1 both as stand-alone features and in concatenation with MFCC. Only the multi-stream approach uses a number of parameters doubled compared to other architectures. Results reveal that:

1) Most of the MLP features do not outperform the MFCC baseline when used as stand-alone front end. Only complex MLP front-ends significantly outperform the MFCC baseline i.e., the multi-stream approach (row 7) and the augmented hierarchical approach (row 10). In the rest of the paper, we will refer to those as MLP1 and MLP2 respectively. The performance of MLP features that use long signal time spans, e.g., HATS and MRASTA is particularly poor as stand alone front-end.

2) On the other hand, even when their individual performance is poor, MLP features in concatenation with MFCC always produce considerable improvements in the range of 12-20% relative in the speaker independent system and in the range of 10-17% in the speaker adapted system. The largest improvement after adaptation is obtained by the MLP2 feature set (+17% relative).

3) The relative improvements after speaker adaptation are generally reduced by 2% relative respect to the speaker independent system. This is consistent with what was already observed in [14] on English ASR experiments.

4) MLP1 features (multi-stream approach that combines TANDEM-PLP and HATS) produce the lowest CER as stand-alone feature set while the MLP2 produces the lowest CER in concatenation with MFCC features.

The following section investigates if these findings on features MLP1 and MLP2 hold also on larger amounts of training data (1600 hours of speech) and in a more complex multipass ASR system.

[2] DS combination has replaced the inverse entropy combination after experiments performed in [7]

Figure 1: RWTH evaluation system composed of two subsystems trained on MFCC and PLP features. The two subsystems consist of ML training followed by SAT/CMLLR training. The subsystems lattice outputs are finally combined together.

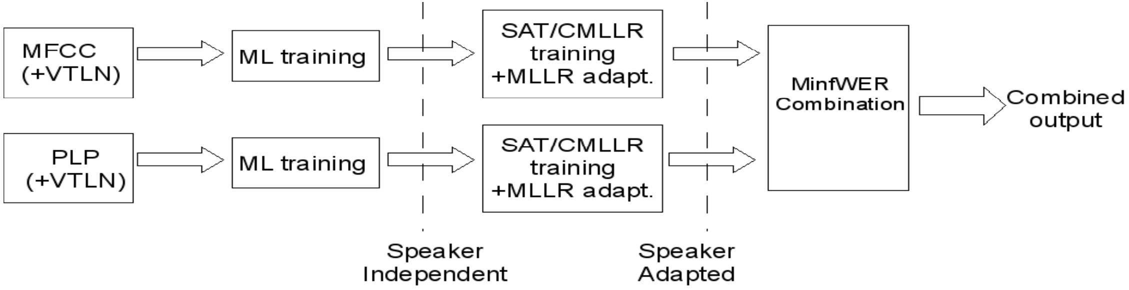

Table 2: CER for MFCC, PLP and MLPs features for the speaker independent system. In brackets relative improvements respect to the spectral features (MFCC or PLP) only are reported.

Feature	GALE-dev07	GALE-dev08	GALE-eval07
MFCC	15.7	14.0	15.8
MLP1	13.1 (+20%)	12.4 (+11%)	13.8 (+12%)
MLP2	13.3 (+18%)	13.1 (+6%)	13.9 (+12%)
MFCC+MLP1	12.3 (+21%)	11.5 (+17%)	13.1 (+17%)
MFCC+MLP2	11.6 (+26%)	11.3 (+19%)	12.8 (+19%)
PLP	16.4	14.9	16.2
MLP1	13.1 (+20%)	12.4 (+16%)	13.8 (+14%)
MLP2	13.3 (+18%)	13.1 (+12%)	13.9 (+14%)
PLP+MLP1	12.3 (+25%)	11.3 (+24%)	12.9 (+20%)
PLP+MLP2	11.7 (+28%)	11.2 (+24%)	12.7 (+21%)

Table 3: CER for MFCC, PLP and MLPs features for the speaker adapted system. In brackets relative improvements respect to the spectral features (MFCC or PLP) only are reported.

Feature	GALE-dev07	GALE-dev08	GALE-eval07
MFCC	14.0	12.5	14.5
MLP1	12.4 (+14%)	11.4 (+8%)	13.4 (+7%)
MLP2	12.3 (+14%)	11.6 (+7%)	13.2 (+8%)
MFCC+MLP1	11.3 (+19%)	10.2 (+18%)	12.2 (+15%)
MFCC+MLP2	10.6 (+24%)	10.1 (+18%)	11.8 (+19%)
PLP	14.4	13.4	14.5
MLP1	12.4 (+14%)	11.4 (+14%)	13.4 (+7%)
MLP2	12.3 (+14%)	11.6 (+13%)	13.2 (+9%)
PLP+MLP1	11.4 (+20%)	10.4 (+22%)	12.2 (+16%)
PLP+MLP2	11.1 (+22%)	10.3 (+23%)	12.1 (+16%)

3. Large scale experiments

In order to study how the previous results generalize on more complex LVCSR systems and large amounts of training data, the experiments are extended using a highly accurate automatic speech recognizer for continuous Mandarin speech trained on 1600 hours of data collected by LDC. The data are used for training of the HMM/GMM systems as well as the MLP front-ends. The evaluation is done on the GALE 2007 development corpus (dev07), used for hyper-parameters tuning, the GALE 2008 development and the sequestered data of the GALE 2007 evaluation (eval07-seq) for a total amount of 5 hours of data.

The evaluation system is composed of two subsystems trained using MFCC and PLP augmented with log pitch estimates as base features. More details on feature normalizations, acoustic models, and language models can be found in [8].

Figure 1 shows the RWTH evaluation system. The first pass consists of simple Maximum Likelihood training, referred as the speaker independent system (SI). In the second pass, speaker variations are compensated for using Speaker Adaptive Training (SAT/CMLLR). During recognition, Maximum Likelihood Linear Regression (MLLR) is applied to the means of the acoustic models. We will refer to it as the speaker adaptative (SA) system. The recognition is performed using a 4-gram language model. Finally, the outputs of the different subsystems are combined at the lattice level using the min.fWER combination method described in [15]; min.fWER has been shown to outperform other lattice combination methods as ROVER or Confusion Network Combination (CNC) [15], and also within the GALE project it has shown to yield competitive systems [8]. This system is referred to as system combination (SC).

Furthermore, we also study the effect of discriminative training both on individual sub-systems and on the lattice combination using a modified Minimum Phone Error criteria [16].

Speaker Independent - Adapted system Table 2 reports the performance of the SI system trained on MFCC and PLP features as well as the two MLP front-ends. Furthermore results obtained concatenating spectral features with MLP front-ends are also reported. The values in brackets represent the relative improvements w.r.t. systems trained on spectral features only. Table 3 reports the same performance for the Speaker Adapted system.

The results show similar trends as for the 100 hours. In other words, the MLP feature performance scales with the amount of training data. In particular:

1) The MLP1 and MLP2 front-ends outperform the spectral features and produce a relative improvement in the range of 15%-25% when used in concatenation with MFCC or PLP. The improvements are verified on all the three data sets.

2) The relative improvements after Speaker Adaptative Training (SAT) are generally reduced respect to the speaker independent system.

3) After SAT, the MLP2 features (based on a hierarchical approach) yield the best performance in concatenation with both MFCC and PLP.

System combination The results of MFCC and PLP subsystems combination are reported in Table 4 (first row). For investigation purposes, corresponding sub-systems trained using MLP1 and MLP2 front-ends are combined in the same way and their performance is reported in Table 4 (second row). Their performance is superior to the MFCC-PLP system by $9-14\%$ relative.

In order to increase the complementarity of the subsystems, features MLP2 and MLP1 were concatenated with MFCC and PLP, respectively. The performance of the lattice

Table 4: Lattice combination (designated with $\oplus$) of MFCC or PLP subsystems in concatenation with MLP features

Features	GALE-dev07	GALE-dev08	GALE-eval07
MFCC $\oplus$ PLP	12.9	11.9	13.5
MLP1 $\oplus$ MLP2	11.1 (+14%)	10.7 (+10%)	12.3 (+9%)
MFCC+MLP2 $\oplus$ PLP+MLP1	9.9 (+23%)	9.4 (+21%)	11.0 (+18%)

Table 5: Effect of discriminative Training on different subsystems and their combination (designated with $\oplus$).

Features	GALE-dev07	GALE-dev08	GALE-eval07
PLP+MLP1	9.9 (+13%)	9.3 (+10%)	11.0 (+9%)
MFCC+MLP2	9.6 (+9%)	9.2 (+9%)	11.0 (+7%)
MFCC+MLP2 $\oplus$ PLP+MLP1	8.8 (+11%)	8.5 (+10%)	10.4 (+6%)

level combination of those two sub-systems is reported in Table 4 (third row). The results show that using MLP features in concatenation with MFCC/PLP features produces an additional relative improvement in the range of $18 - 23\%$ in system combination.

Discriminative Training Table 5 reports CER obtained after discriminative training. Results are reported for the PLP+MLP1 system, the MFCC+MLP2 system and their lattice level combination. In all the three cases, discriminative training is reducing the CER in the range 6-13% relative, showing that it is effective also when used together with different MLP front-ends. For computational reasons, fully contrastive results with and without discriminative training are not available on the 1600 hours system. However the relative improvements reported in Table 5 are comparable to those obtained without MLP features [17], showing that improvements obtained from the two techniques can be additive.

4. Summary and discussion

This paper first investigates several MLP front-ends proposed during the GALE project on a small scale experimental setup. Results reveal that most of the MLP features do not outperform the MFCC baseline when used as stand-alone front end. Only complex front-ends like, Multi-stream feature (MLP1) and augmented Hierarchical features (MLP2) outperform spectral features. They outperform the MFCC when used as a stand alone feature (10% relative improvement in CER for MLP1) and a considerable improvement is obtained in concatenation with spectral features (up to 17% relative for MLP2).

In the second part of the paper, we investigate these two features with large amount of training data as well as on a state-of-the-art multipass system. The findings from the small scale study hold for large amount of training data on speaker independent, speaker adapted systems and after lattice level combination. This is verified both in concatenation with MFCC and PLP features. The hierarchical MLP approach (MLP2) holds the largest reduction in CER. The final gain after lattice combination is in the range of $18 - 23\%$ relative for the different evaluation data sets.

In the future we intend to extend these studies to other recently introduced MLP front-ends, such as bottleneck features [4] which has been mainly tested for Arabic language.

5. Acknowledgments

This work was partly supported by the the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and by Swiss National Science Foundation through the NCCR IM2. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA. The authors would like to thank colleagues involved in the GALE project at IDIAP, ICSI, RWTH and SRI especially Arlo Faria and Andreas Stolcke for their help and discussions.

6. References

[1] Hermansky H., Ellis D., and Sharma S., “Connectionist feature extraction for conventional hmm systems.,” *Proceedings of ICASSP*, 2000.

[2] Chen B., Chang S., and Sivadas S., “Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers,” in *Proceedings of Eurospeech*, 2003.

[3] Hermansky H., Tibrewala S., and Pavel M., “Towards ASR on partially corrupted speech,” *Proc. ICSLP*, 1996.

[4] Fousek P., Lamel L., and Gauvain J.L., “Transcribing Broadcast Data Using MLP Features.,” *Procedings of Interspeech 2008*.

[5] Morgan N. et al., “Pushing the envelope - aside,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, 2005.

[6] Vergyri D. at al., “Development of the SRI/Nightingale Arabic ASR system,” *Procedings of Interspeech 2008*.

[7] Mei-Yuh Hwang et al., “Building a highly accurate Mandarin speech recognizer,” *Proc of ASRU.*, 2007.

[8] Plahl C. et al., “Development of the GALE 2008 Mandarin LVCSR system,” in *Proceedings of Interspeech*, Brighton, U.K., Sept. 2009, pp. 2107–2110.

[9] Valente F., Magimai.-Doss M., Plahl C., and Ravuri S., “Hierarchical modulation spectrum for the GALE project,” in *Proceedings of Interpseech*, 2009.

[10] Lei X. et al., “Improved Tone Modeling for Mandarin Broadcast News Speech Recognition .,” *Proceedings of Interspeech*, 2006.

[11] Valente F. and Hermansky H., “Combination of acoustic classifiers based on Dempster-Shafer theory of evidence,” *Proc. ICASSP*, 2007.

[12] Hermansky H. and Fousek P., “Multi-resolution RASTA filtering for TANDEM-based ASR.,” in *Proceedings of Interspeech*, 2005.

[13] Valente F. and Hermansky H., “Hierarchical and parallel processing of modulation spectrum for ASR applications,” in *Proceedings of ICASSP*, 2008.

[14] Zhu Q. et al., “On using MLP features in LVCSR,” *Proceedings of ICSLP 2004*.

[15] Hoffmeister B. et al., “Frame based system combination and a comparison with weighted ROVER and CNC,” in *Proceedings of Interspeech*, Pittsburgh, PA, USA, Sept. 2006, pp. 537–540.

[16] Heigold G. et al, “Modified MPE/MMI in a transducer-based framework,” in *Proceedings of ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3749–3752.

[17] Heigold G. et al., “Margin-based discriminative training for string recognition.,” *Journal of Selected Topics in Signal Processing - Statistical Learning Methods for Speech and Language Processing*, to appear December 2010.