Analysis and Comparison of Recent MLP Features for LVCSR Systems

Fabio Valente[1], *Mathew Magimai Doss*[1] *and Wen Wang*[2]

[1]IDIAP Research Institute, CH-1920 Martigny, Switzerland
[2] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.
{fabio.valente,mathew}@idiap.ch, wwang@speech.sri.com

Abstract

MLP based front-ends have evolved in different ways in recent years beyond the seminal TANDEM-PLP features. This paper aims at providing a fair comparison of these recent progresses including the use of different long/short temporal inputs (PLP,MRASTA,wLP-TRAPS,DCT-TRAPS) and the use of complex architectures (bottleneck, hierarchy, multistream) that go beyond the conventional three layer MLP. Furthermore, the paper identifies which of these actually provide advantages over the conventional TANDEM-PLP . The investigation is carried on an LVCSR task for recognition of Mandarin Broadcast speech and results are analyzed in terms of Character Error Rate and phonetic confusions. Results reveal that as stand alone features, multistream front-ends can outperform by 10% conventional MFCC while TANDEM-PLP only improve by 1% . On the other hand, when used in concatenation with MFCC features, hierarchical/bottleneck front-ends reduce the character error rate by $+18\%$ relative compared to $+14\%$ relative from TANDEM-PLP. The various input long-term representations recently developed provide comparable performances.

Index Terms: TANDEM features, Multilayer Perceptron, Acoustic features, GALE project, LVCSR.

1. Introduction

Since the original work of Hermansky and colleagues [1], a large number of Multilayer Perceptron (MLP) front-ends have been proposed for Automatic Speech Recognition (ASR). The first MLP based front-end [1] consisted of a three-layer MLP trained on nine consecutive frames of PLP features as input. The MLP outputs represent phonetic posterior probabilities, which, after a Log/KLT transform, are used as conventional features in HMM/GMM recognition systems. In recent times, MLP front-ends have significantly progressed along two main directions: *1- the use of different input representations to the MLP* and *2- the use of complex MLP architectures beyond the conventional three-layer perceptron.* The first direction includes speech representations that aim at using information from long speech temporal trajectories which could capture phenomena such as co-articulation and provide complementarity to MFCC or PLP features [2]. Because of the large dimension of these time windows, a number of techniques for efficiently encoding the information have been proposed like MRASTA [3], DCT-TRAPS [4], and wLP-TRAPS [5]. The second direction includes a number of heterogeneous techniques that aim at overcoming the pitfalls of the three-layer MLP classifier, including bottleneck architectures [6], hierarchical architectures [7], and multi-stream approaches [8].

In our previous related work [7], we investigated a subset of these techniques, namely, the MRASTA processing and its hierarchical version in a Mandarin broadcast LVCSR system developed in the framework of the GALE project[1]. This paper aims at complementing that study including other MLP input features (DCT-TRAPS and wLP-TRAPS) as well as Bottleneck architectures in order to cover all the front-ends that have been proposed and integrated into LVCSR systems. Furthermore, the paper investigates which of these techniques actually improve over the conventional TANDEM-PLP.

The study is carried on the same Mandarin Broadcast system described in [7] and we examined the MLP feature performances as stand-alone front ends and in concatenation with spectral features (MFCC). The remainder of this work is organized as follows. Section 2 describes the baseline system and the experimental setup. Section 3 experiments with long temporal input in a three-layer MLP architecture. Section 4 experiments with long temporal input in more complex architectures such as bottleneck and hierarchies and the results are analyzed in terms of phonetic confusions. The results are then summarized and discussed in Section 5.

2. Experiments setup

The following studies are based on a simplified version of the large vocabulary ASR system for transcription of Mandarin broadcast described in [9], 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 conversation data manually transcribed including speaker labels. Results are reported on the DARPA GALE 2006 evaluation test set (eval06). The baseline system uses 13 standard MFCC plus smoothed log-pitch estimate as described in [10] as Mandarin is a tonal language. Furthermore, they are augmented with first and second order temporal derivatives resulting in a feature 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. The decoding phase consists of two decoding passes, speaker independent (si) decoding followed by a speaker adapted (sa) decoding. The performance of this baseline system on eval06 data is reported in Table 1.

Let us first examine the TANDEM-PLP features performances, where the input to the MLP is 9 consecutive frames of mean-variance speaker normalized PLP features. Furthermore, this representation is augmented with 9 consecutive frames of the log pitch estimate [10] with its temporal derivatives, producing a 42×9 dimensional input feature vector. The training is done on a toneme set composed of 71 tonemes. The total number of parameters in the MLP is equal to one million. After PCA, a dimensionality reduction accounting for 95% of the

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

total variability is applied, resulting in a feature vector of dimensions 35. TANDEM-PLP feature performance is reported in Table 1. While comparable to the MFCC baseline as stand alone features, the MLP front-end produces an improvement of 14% relative when concatenated with spectral features. Next, we investigate the use of different input features while keeping constant the total number of parameters in the MLP to one million.

Table 1: Performances of the MFCC baseline system, TANDEM-9frames PLP and their concatenation. The relative improvement w.r.t. the baseline is reported in the parentheses.

	MFCC	TANDEM	MFCC+TANDEM
CER	25.8	25.5 (+1%)	22.2 (+14%)

3. Long Temporal Inputs

We replaced the 9frames-PLP input to the MLP with a Temporal Pattern or TRAPS [11], i.e., a long-time span of speech signal. Given the high dimensionality of the TRAPS, several techniques for efficiently extracting information have been proposed.

The **Multiple RASTA (MRASTA)** filtering [3] is an extension of RASTA filtering consistent with human perception of modulation frequencies modeled using a bank of filters equally spaced on a logarithmic scale. This bank of filters subdivides the available modulation frequency range into separate channels with a decreasing resolution moving from slow to fast modulations. The feature extraction is composed of the following parts: 19 critical band auditory spectrum is extracted from Short Time Fourier Transform of a signal every 10 ms. A 600 ms long temporal trajectory in each critical band is filtered with a bank of band-pass filters. These filters represent first derivatives and second derivatives of Gaussian functions with variance σ_i varying in the range 8-60 ms. After MRASTA filtering, frequency derivatives across three consecutive critical bands are introduced. The total number of features used as input for a three-layer MLP is 432.

The **DCT-TRAPS** aims at reducing the dimension of the trajectories using a Discrete Cosine Transform (DCT) [4]. The critical band auditory spectrum is extracted from Short Time Fourier Transform of a signal every 10 ms. Then 500 ms long energy trajectories are extracted for each of the 19 critical bands that compose the spectrogram. Those are projected on the first 16 coefficients of a DCT transform resulting in a vector of size $19 \times 16 = 304$, which is then used as input to the MLP. In contrary to the MRASTA, they do not emulate any sensitivity of the hearing properties to the different modulation frequencies.

wLP-TRAPS [5] represents a third alternative which does not use the short term spectrum. These features are obtained by warping the temporal axis after LP-TRAP features calculation [12]. The feature extraction is composed of the following steps: first, linear prediction is used to model the Hilbert envelops of pre-warped 500ms long energy trajectories in auditory-like frequency sub-bands. The warping ensures that more emphasis is given to the center of the trajectories compared to the borders [5], thus emulating again human perception. 25 LPC coefficients in 19 frequency bands are then used as input to the MLP, producing a feature vector of dimension $19 \times 25 = 475$.

As Mandarin is a tonal language, those representations can be augmented with the smoothed log-pitch estimate [10] and with the value of the critical band energy (19 features per frame). In the following, we will refer to these as Augmented features (A-MRASTA, A-DCT-TRAP, A-wLP-TRAPS).

Table 2: CER for MLP features making use of long time spans of the signal as stand alone features and in concatenation with MFCC. The relative improvement w.r.t. the baseline is reported in parentheses.

	MLP	MFCC+MLP
MRASTA	30.7 (-19%)	23.1 (+10%)
DCT-TRAPS	31.7 (-23%)	23.2 (+10%)
wLP-TRAPS	28.2 (-9%)	23.0 (+11%)
A-MRASTA	**26.6 (-3%)**	**22.2 (+14%)**
A-DCTTRAPS	28.9 (-12%)	22.5 (+13%)
A-wLPTRAPS	27.3 (-6%)	**22.2 (+14%)**

Table 2 reports the performances of MLP features obtained from training on those long temporal inputs. They perform quite poorly as stand alone features but they still provide improvements around 10% relative in concatenation with the MFCC. As stand-alone front-end, the wLP-TRAPS outperforms the other two (DCT-TRAPS and MRASTA). While in concatenation with spectral features and after adaptation, the three representations are comparable. Table 2 also reports performances of augmented features. Also in this case the three representations have comparable performances in concatenation with MFCC. In summary, as stand alone features and in concatenation with MFCC, long temporal window inputs do not outperform conventional TANDEM-PLP whenever a three-layer MLP is used.

In order to understand the differences between the various MLP front-ends, we analyzed the errors they produce in terms of broad phonetic classes (Vowels, Stops, Fricatives, Affricatives, Approximants, Nasals). Figure 1 plots the per-class accuracy in case of MLP trained using 9frames-PLP and DCT-TRAPS inputs. The overall performance of the former is superior to the overall performance of the latter. However, the DCT-TRAPS outperforms the TANDEM-PLP on almost all the stop consonants 'p', 't', 'k', 'b', 'd' and the affricative 'ch'. Stop consonants are short sounds known to be prone to strong co-articulation from the following vowel and their recognition can be largely improved considering information from the following vowel. Vowels and other consonants are still better recognized from the short term features. These facts are verified also on MLPs trained on MRASTA and wLP-TRAPS. In summary, training MLPs using short-term spectral input outperforms training using long term temporal input on most of the phonetic classes apart plosives and affricatives. After augmentation with pitch and energy, the performances of long and short temporal inputs are comparable.

4. MLP architectures

The other direction in which MLP front-ends have evolved is the use of more complex architectures beyond the three layer MLP. The main alternatives to the three-layer architectures include the following.

Bottleneck features are recently introduced MLP non-probabilistic features [13]. The conventional three-layer MLP is replaced with a four- or five-layer MLP where the first layer is the input features and the last layer is the phonetic targets. In the five-layer case, the size of the second layer is large to provide enough modeling power, the size of the third layer is small, typically equal to the desired feature dimension, while the size of the fourth one is approximatively half of the second layer [13]. Instead of using the output of the MLP, features

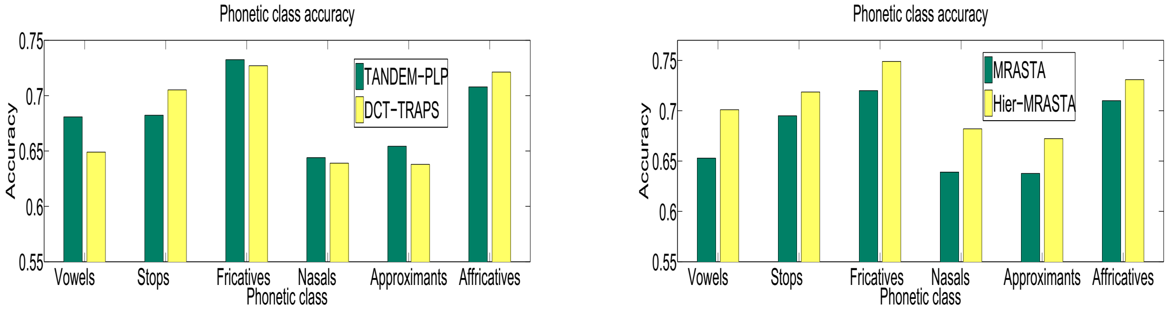

Figure 1: (Left plot) Phonetic-class accuracy obtained by the TANDEM-9framesPLP and DCT-TRAPS input. The former outperforms the latter on most of the classes apart from stops and affricatives. (Right plot) Phonetic-class accuracy obtained by the MRASTA and the Hierarchical MRASTA.

Table 3: CER for MLP features making use of bottleneck architectures as stand alone features and in concatenation with MFCC. The relative improvement w.r.t. the baseline is reported in parentheses.

	MLP	MFCC+MLP
bottleneck-MRASTA	25.9 (+0%)	21.5 (+17%)
bottleneck-DCTTRAPS	25.7 (+0%)	22.0 (+15%)
bottleneck-wLPTRAPS	24.9 (+3%)	21.5 (+17%)
A-bottleneck-MRASTA	24.0 (+6%)	**21.2 (+18%)**
A-bottleneck-DCTTRAPS	24.9 (+3%)	21.5 (+17%)
A-bottleneck-wLPTRAPS	24.1 (+6%)	**21.2 (+18%)**

are obtained from the linear activation of the third layer. Bottleneck features do not require a dimensionality reduction, as the desired dimension can be obtained fixing the size of the bottleneck layer. Furthermore, the linear activations are already Gaussian distributed thus they do not require any Log transform. Performances obtained replacing the three-layer MLP with a bottleneck architecture are reported in Table 3. While keeping constant the number of total parameters, the bottleneck produces a reduction in the errors both with and without MFCC. Furthermore, in their augmented version, the long temporal inputs coupled with the bottleneck architectures outperform the conventional TANDEM-PLP with and without spectral feature concatenation.

Beside increasing the number of layers in the MLP, hierarchies of classifiers have also been proposed in literature as alternative. In the **Hierarchical MRASTA features**, the gaussian filter-banks are split in two separate filter banks that filter respectively fast and slow modulation frequencies, or equivalently the filters with short and long temporal support. The cutoff frequency for both filter-banks is approximatively 10Hz. The output of the MRASTA filtering is then processed according to a hierarchy of MLPs progressively moving from high to low modulation frequencies or equivalently from short to long temporal context [7]. The effect of this sequential processing is that the first MLP trained on short temporal context is effective on most of the phonetic classes apart stops and affricatives. Those estimates are then corrected from the second MLP using the information from longer temporal context. Figure 1 plots the phonetic class accuracy obtained by the three-layer MLP trained using the MRASTA input and the hierarchical approach. It is noticeable that the second outperforms the first on all the targets. Performances of Hierarchical MRASTA and its augmented version (A-Hier) are reported in Table 4. Again the total number of parameters in the architecture is kept constant to one million. Results reveal that hierarchical processing considerably improves the performances obtained from training on long temporal inputs. Furthermore, after augmentation, the approach appears superior to the TANDEM-PLP. Although based on two different rationales, hierarchical and bottleneck architectures provide comparable performances.

Table 4: CER for MLP features making use of architectures beyond the three-layer models as stand alone features and in concatenation with MFCC. The relative improvement w.r.t. the baseline is reported in parentheses.

	MFCC	MFCC+MLP
Hier	26.5 (-3%)	21.9 (+15%)
A-Hier	24.1 (+6%)	**21.2 (+18%)**
Multi-stream	**23.1 (+10%)**	21.7 (+16%)

A third alternative architecture is the **Multi-stream model** [8]. The MLP outputs are phonetic target posterior probabilities that can be combined into a single estimate using probabilistic rules. The rationale behind it consists in the fact that MLPs trained using different input representations, e.g., short and long temporal windows, will perform differently in multiple conditions. Dynamically weighting the posterior streams should take advantage of both representations. Thus posteriors obtained from MLPs trained on spectral features (9frames-PLP) and long signal time spans (MRASTA) are combined using the Dempster-Shafer method [14] and used as features after a Log/PCA transform. Multi-stream comes at the obvious cost of doubling the total number of parameters in the system. Results reported in table 4 reveals that multi-stream reduces the CER by 10% relative when used as stand alone and by 16% relative when used in concatenation with MFCC.

Another interesting finding is the fact that as stand-alone features, the multi-stream approach has the largest CER improvement (10% relative over MFCC), while in concatenation with MFCC, the hierarchical or bottleneck architectures produce the largest CER reduction . This suggests that the best MLP features may not be the most complementary. The effect can be explained by the fact that the multi-stream approach makes use of spectral information (through the 9frame PLP) while the hierarchical/bottleneck architectures do not. This information produces a large improvement whenever MLP features alone are used but does not appear complementary to the MFCC features as they both represent spectral information. On the other hand, the hierarchical/bottleneck architectures which do not use any spectral information, appears more complementary when used in concatenation with the MFCC.

5. Summary and Discussion

Following the original work of Hermansky and colleagues [1], a large number of different input representations have been proposed in the context of MLP based feature extraction. They are often coupled with architectures that go beyond the simple

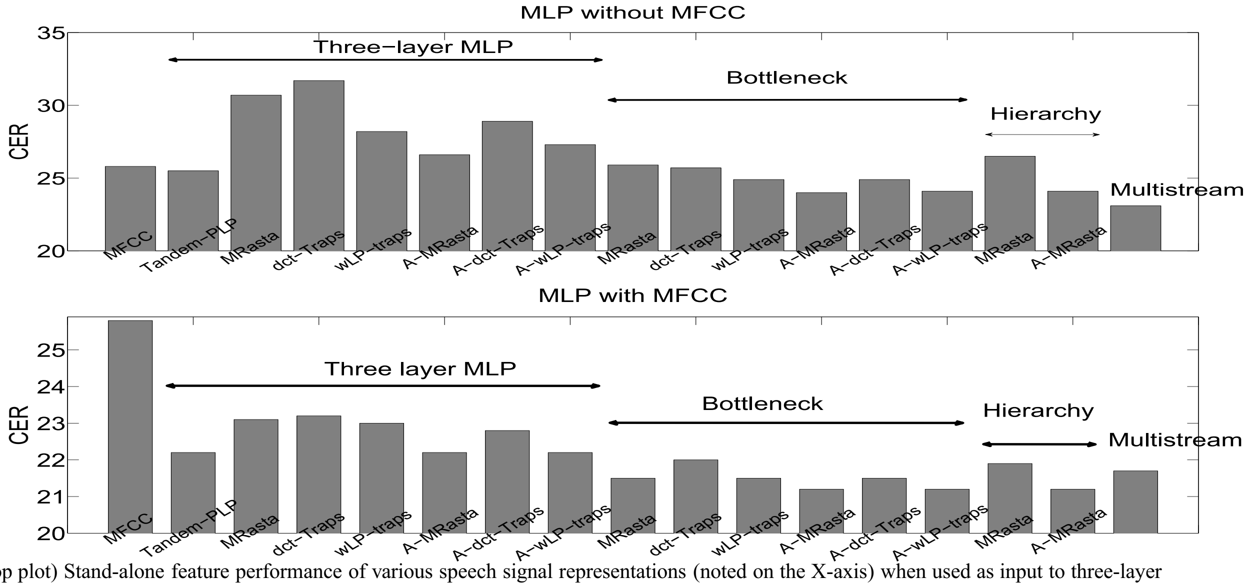

Figure 2: (Top plot) Stand-alone feature performance of various speech signal representations (noted on the X-axis) when used as input to three-layer MLP, bottleneck, hierarchical and multi-stream architectures. The down plot reports the feature performances when used in concatenation with MFCC.

three-layer perceptron. The performances of the various MLP front-ends are summarized in Figures 2 as stand-alone features (top plot) and in concatenation with MFCC (down plot).

Figure 2 (top plot) reveals that, when a three-layer MLP is used, none of the long temporal inputs (MRASTA, DCT-TRAPS, wLP-TRAPS, and their augmented versions) outperform the conventional TANDEM-PLP nor the MFCC baseline. On the other hand, replacing the three-layer MLP with a bottleneck or hierarchical architecture (while keeping constant the total number of parameters) considerably reduces the error, achieving a CER lower than the MFCC baseline. The lowest CER is obtained by the multi-stream architecture which combines outputs of MLPs trained on long and short temporal contexts improving by 10% relative over the MFCC baseline.

Figures 2 (down plot) reports CER obtained in concatenation with MFCC and reveals that, even when their performances are poor as stand-alone front-end, three-layer MLP features based on long temporal spans always appear to provide complementary information to the MFCC with improvements in the range of 10-14% relative. When the three-layer MLP is replaced with bottleneck or hierarchical architectures, the improvements are increased to the range of 16-18%. The various methods for encoding the information (DCT-TRAPS, MRASTA, wLP-TRAPS) perform equally well when augmented with pitch and energy. It is interesting to notice that, in concatenation with MFCC, the lowest CER is obtained by the bottleneck/hierarchical architectures rather then the multi-stream features (see previous section for explanation).

Table 5 summarizes the improvements that modifications to the three-layer MLP can produce with respect to the original TANDEM-PLP features. As stand-alone front-end, the lowest CER is produced by multi-stream features (+10% relative over the MFCC baseline, compared to +1% obtained by TANDEM-PLP); in concatenation with MFCC, the lowest CER is produced by bottleneck/hierarchical architectures (+18% relative, compared to +14% obtained by TANDEM-PLP, over the MFCC baseline)[2].

Table 5: Summary Table of CER and improvements.

	TANDEM	Multistream
MLP	25.5 (+1%)	23.1 (+10%)
	TANDEM	Hier/Bottleneck
MLP+MFCC	22.2 (+14%)	21.2 (+18%)

6. References

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

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

[3] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based asr.," in *Proceedings of Interspeech 2005*, 2005.

[4] Schwarz P., Matejka P., and Cernocky J., "Extraction of features for automatic recognition of speech based on spectral dynamics," in *Proceedings of TSD04, Brno, Czech Republic*, September 2004, pp. 465 – 472.

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

[6] Grezl F., Karafiat M., Kontar S., and Cernocky J., "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proceedings of ICASSP07,Hononulu*, 2007.

[7] Valente F., Magimai-Doss M., Plahl C., Ravuri S., and Wang W., "A Comparative Study of MLP Front-ends for Mandarin ASR ," in *Proceedings of Interpseech*, 2010.

[8] Hermansky H. and Tibrewala S., "Towards ASR on partially corrupted speech," *Proc. ICSLP*, 1996.

[9] Hwang M.-Y., Gang P., Ostendorf M., Wang W., Faria A., and Heidel A., "Building a highly accurate mandarin speech recognizer with language-independent technologies and language-dependent modules," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 17, no. 7, 2009.

[10] Lei X., Siu S., Hwang M.-Y., Ostendorf M., and Lee T., "Improved Tone Modeling for Mandarin Broadcast News Speech Recognition .," *Proceedings of Interspeech, 2006.*

[11] Hermansky H. and Sharma S., "Temporal Patterns (TRAPS) in ASR of Noisy Speech," in *Proceedings of ICASSP'99, Phoenix, Arizona, USA*, 1999.

[12] Marios Athineos, Hynek Hermansky, and Daniel P. W. Ellis, "Lp-trap: Linear predictive temporal patterns," in *Proc. ICSLP*, 2004, pp. 1154–1157.

[13] Grezl F. and Fousek P., "Optimizing bottleneck features for lvcsr," in *Proceedings of ICASSP08,Las Vegas*, 2008.

[14] Valente F. and Hermansky H., "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence," *Proc. ICASSP*, 2007.

[2] This work was supported by the the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and by the Swiss National Science Fundation through IM2 grant. Authors would like to thanks colleagues involved in the GALE project at IDIAP, ICSI, RWTH and SRI as well as Dr. Petr Fousek.