A study on speaker normalized MLP features in LVCSR

Zoltán Tüske[1,2], *Christian Plahl*[1], *Ralf Schlüter*[1]

[1]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany
[2]IDIAP Research Institute, CH-1920 Martigny, Switzerland
{tuske, plahl, schluter}@cs.rwth-aachen.de

Abstract

Different normalization methods are applied in recent Large Vocabulary Continuous Speech Recognition Systems (LVCSR) to reduce the influence of speaker variability on the acoustic models. In this paper we investigate the use of Vocal Tract Length Normalization (VTLN) and Speaker Adaptive Training (SAT) in Multi Layer Perceptron (MLP) feature extraction on an English task. We achieve significant improvements by each normalization method and we gain further by stacking the normalizations. Studying features transformed by Constrained Maximum Likelihood Linear Regression (CMLLR) based SAT as possible input for MLP, further experiments show that MLP could not consistently take advantage of SAT as it does in case of VTLN.

Index Terms: GMM-HMM, VTLN, SAT, CMLLR, LVCSR, MLP, Dempster-Shafer

1. Introduction

In order to decrease the performance gap between speaker independent and speaker dependent modeling resulting from speaker variability, the following methods are widely applied in state-of-the-art automatic speech recognition (ASR) systems.

Vocal Tract Length Normalization: The differing vocal tract sizes of different speakers lead to shift of the formants in frequency spectrum. To compensate these shifts, the cepstral features are calculated by use of a warped frequency scale (e.g by a piecewise linear warp). Using a generic speech model to estimate the warping factor for test data, a multi pass recognition could be avoided [1]. Moreover, in [2] it has been shown that VTLN could be performed as a linear transformation of the cepstral features.

Speaker Adaptive Training (SAT): To reduce the influence of irredundant variability on acoustic models —e.g. the speaker variability arising from the different physical attributes, accent, etc.— SAT is applied on the acoustic front-end during the construction of the ASR system [3, 4]. Using a preliminary acoustic model (AM) being trained on the original features, the features are linearly transformed with respect to the speaker by applying the CMLLR approach. Finally, the speaker adapted transformed features are used to train the speaker adapted AM. For recognition, the adaptation is based on the statistic collected from the previous pass (multi pass recognition).

Multi Layer Perceptron (MLP) features: In state-of-the-art Hidden Markov Model (HMM) based ASR systems the MLP based features have become a crucial component [5, 6] of the front-end techniques. Its strong gender and speaker normalization ability is well studied in the literature [7, 8]. To better fit to Gaussian Mixture Model (GMM) based AM, the MLP outputs are decorrelated after applying a logarithm (TANDEM scheme) [9]. Providing complementary information, concatenation of the MLP based features and standard cepstral based features —like Mel Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction Coeff. (PLP)— results in better performance [4, 10]. As the MLPs are usually trained so that their outputs represent phoneme posterior probabilities, multiple MLP outputs could be combined by use of probabilistic rules [11, 12, 13].

According to our knowledge, no investigation on the contribution of all these three components (MLP, SAT, VTLN) in LVCSR task has been done yet. Moreover, the application of VTLN and SAT on MLP features based ASR is not consistent in the literature. Usually either the MLP features are not normalized or adaptation results are not presented. Therefore, we systematically investigate the effect of VTLN and SAT on cepstral features, MLP features, and on combined cepstral and MLP features. We also examine whether MLP could benefit from SAT transformed input features —like from a VTLN transformed input. The latter question has also not been studied in the literature yet.

The paper is organized as follows: Following the overview of related work in Section 2, Section 3 describes the training and testing corpora. Section 4 gives the details of the feature extraction methods used in the experiments. Section 5 reports the experimental setup followed by the results (Section 6). The paper closes with conclusions.

2. Related Work

In state-of-the-art GMM/HMM ASR systems the acoustic models are trained on cepstral features augmented by combined (multi-stream) MLPs features, where the MLPs are typically trained on short ($\sim$100 ms) and long ($\sim$1 sec) time series of critical band energies [8, 13, 14]. In [13] the Dempster-Shafer theorem based combination rule is introduced for ASR which is one of the most efficient methods to fuse MLP outputs.

As VTLN on cepstral features could be expressed as linear transformation [2], applying both VTLN and SAT does not result in accumulation of their contribution. Although the MLP features are proven to have good gender normalization ability the use of VTLN on the input is still suggested according to the results of [7].

In [4] the adapted augmented cepstral and MLP features outperform the cepstral features, although the relative improvement is less than without the adaptation. Previous related work [4, 10, 14, 15], did not exhaustively investigate the use of speaker normalized MLP features.

3. Corpus description

Within the European project Technology and Corpora for Speech to Speech Translation (TC-STAR) about 95 hours of manually transcribed English speech data of the European Parliament Plenary Session (EPPS) are collected. All data are

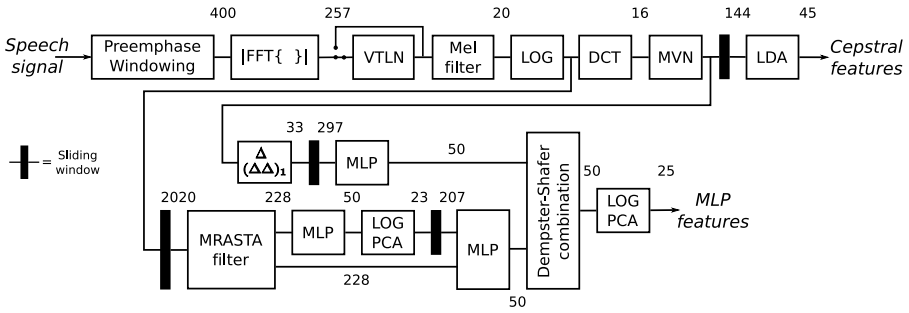

Figure 1: *The straightforward cepstral and hierarchical multi-stream MLP feature extraction with feature dimension corresponding to each processing step*

recorded with professional equipment and sampled with 16 kHz. The training of the AM as well as the training of the MLP is performed on 88h of acoustic data.

Table 1: *EPPS training and testing corpus statistics*

EPPS	train	dev07	eval07
total data	88h	3.2h	2.9h
# running words	761k	40k	37k

The performance of the final systems has been evaluated on the EPPS development and testing data of 2007. Each corpus contains 3h of audio data and the development corpus has been used for tuning. Table 1 shows the corpus statistics of the training and testing data.

4. Feature extraction

In [3] the training of the MLP features of System3 is based on critical band energies (CRBE) of long time span only and no VTLN or SAT transformations are included in the MLP training process. We reconsider to update the feature extraction scheme based on [13].

First, the cepstral features are extracted from the audio file. The pre-emphasized power spectrum is computed every 10 ms over a window of 25 ms. After integration of the power spectrum —20 triangular filters are used, equally spaced on Mel-scale— the features are logarithmized. Finally, we compute the 16 MFCCs from the logarithmic CRBE and apply mean and variance normalization. Features within a sliding window of length 9 are projected by linear discriminant analysis (LDA) to a 45 dimensional subspace.

Now, we have trained two MLPs in parallel and have combined the phoneme posterior estimates by Dempster-Shafer [4]. The final posteriors are further transformed by logarithm and Principal Component Analysis (PCA). All MLPs are trained using cross-entropy criterion and approximate phoneme class posterior probabilities. All activations of the nodes within the output layer are transformed by the softmax function —all outputs sum up to 1—, whereas the sigmoid transfer function is applied in all other layers. 9 consecutive MFCC frames and its first and second derivatives are fed to the first MLP. The second posterior estimates are derived from a hierarchical processing of two MLPs. The input of the first MLP in this hierarchy is based on the fast modulation frequencies of the multi-resolution rasta filtering (MRASTA) [16] which are based on a temporal context of one second. Instead of the proposed PLP spectrum we use the Mel spectrum based CRBEs because former experiments have shown that CRBEs have performed slightly better. The second MLP within the hierarchy contains the slow modulation frequencies and the posterior estimates of the first MLP. The overall feature extraction is shown in Figure 1. Finally, the GHMM/HMM system is trained on the cepstral, MLP or on the concatenated features.

The feature extraction is changed slightly, when speaker adaptation is applied to the MLP training. The MFCC and MLP features are transformed by speaker adaptive matrices and are fed into a final MLP, see Figure 2.

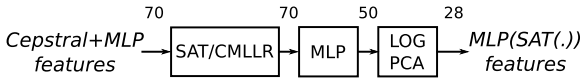

Figure 2: *MLP features extracted from SAT transformed input features, here from concatenated cepstral and MLP features*

5. Experimental Setup

The acoustic models for all systems are based on triphones with cross-word context, modeled by a 6-state left-to-right HMM. A decision tree based state tying is applied resulting in a total of 4500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix.

Instead of training of AMs from scratch an initial alignment created by the VTLN MFCC baseline model from [3] was used in estimation of the state-tying and in the first iterations.

The LDA matrix estimation of the cepstral features are performed in an iterative procedure. First, a phonetic decision tree is estimated on initial alignment followed by the LDA matrix estimation. Next, we repeat the decision tree and LDA estimation.

In the ML training first single Gaussian densities are estimated. Afterwards, depending on the results on the development set 7 or 8 splits are performed ending up with 500k or 1M Gaussians. We denote this training method as Speaker Independent (SI).

The filterbank underlying the CRBE extraction undergoes VTLN. The warping factor classifier is trained with cepstral features beforehand on the complete training corpus, where the warping parameter is estimated by a grid search in the range of 0.8 - 1.2 and a step size of 0.02 [17].

In order to compensate for speaker variations, we apply constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR) in the second pass using the simple target approach. The speaker adapted AM are trained on the CMLLR transformed features similarly to SI. In recognition

the SAT matrices are estimated on the results of the previous recognition pass. This second pass training and testing is referred in the followings as Speaker Adapted (SA).

For training the MLPs, the number of nodes in the hidden layer is fixed to 5000. The final posterior estimates are logarithmized and are transformed by PCA, according to 95% of the variability.

In the experiments where MLPs are trained on SAT transformed features, the same two pass training and recognition process are run, in fact, as 3rd and 4th pass.

A 4-gram language model (LM) is used in recognition. The LM has been trained on the final text editions and verbatim transcriptions of the European Parliament Plenary Sessions.

6. Results

The different acoustic features are investigated with and without VTLN, and with or without SAT/CMLLR.

6.1. Single Features System

In the first experiment the performance of the cepstral features —denoted as MFCC— and the MLP features —as shown in Figure 1— is compared. Since the MLP features are combined posterior estimates by Dempster-Shafer, these features are denoted as DS. In the case where VTLN is applied in the feature extraction, the notations change to $\text{MFCC}_{\text{VTLN}}$ and DS_{VTLN}. The absolute performance in Word Error Rate (WER) and the effect of VTLN expressed in relative improvement can be seen in Table 2.

Table 2: *Performance in WER [%] of the single features without or with VTLN, in brackets the relative improvement [%] compared to the corresponding non-VTLN features*

Features	dev07 SI	dev07 SA	eval07 SI	eval07 SA	dim/split
MFCC	17.3	14.3	16.2	13.1	45/8
DS	15.0	13.3	13.5	11.8	25/8
$\text{MFCC}_{\text{VTLN}}$	16.5 (4.6)	14.3 (0.0)	15.9 (1.9)	13.1 (0.0)	45/8
DS_{VTLN}	14.2 (5.3)	12.9 (3.0)	13.2 (2.2)	11.6 (1.7)	25/8

The Dempster-Shafer combined MLP outputs clearly outperform the MFCC in all cases. In the case of non-VTLN features, the absolute improvement of 2.3% corresponds to 13.3% relative improvement. In contrast, the MLP features based system in [3] does not achieve better performance than the MFCC based system. After SAT the improvement get smaller (rel. 7%), which corresponds to the results reported in the literature [4]. We also measured the efficiency of single (before Dempster-Shafer combination) MLP features yielding in slightly worse recognition performance compared to MFCC.

After SAT we achieve the same performance w.r.t WER for the $\text{MFCC}_{\text{VTLN}}$ and the MFCC features, which corresponds to the conclusion of [2]. Although the VTLN properties of MLP are known, it is still worth to use VTLN on the input features of the MLP. The relative improvement related to VTLN is greater with the MLP features than with cepstral features on both the development and evaluation corpora in the first path. Even after the SAT, the improvement is still observed with MLP features, even though the gain is less than for the speaker independent model.

6.2. Concatenated Features System

In the second experiment the performance of the concatenated cepstral features and MLP features is measured. The two systems are denoted as MFCC+DS and $\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$ with respect to the application of VTLN.

Results for the concatenation of the two features with or without VTLN are shown in Table 3. In brackets the relative improvements produced by the VTLN are reported.

Table 3: *Performance in WER [%] of the concatenated features, in brackets the relative improvement [%] compared to the features without VTLN*

Features	dev07 SI	dev07 SA	eval07 SI	eval07 SA	dim/split
MFCC + DS	14.0	11.7	12.7	10.4	70/7
$\text{MFCC}_{\text{VTLN}}$ + DS_{VTLN}	13.1 (6.4)	11.6 (0.8)	12.1 (4.7)	10.2 (1.9)	70/7

The comparison of the single and the concatenated features is shown in Table 4. Providing complementary information, although both of the MFCC and MLP features are derived from the same critical band energies, the concatenation of the features without VTLN shows more than 6% relative improvement on the development set and more than 5% on the evaluation set in the first pass, compared to the best single features system (DS). In the case of VTLN features, the gain increases further up to 7% and 8% relative improvement. Furthermore, after the second pass the gap grows as well, between the best single features and the concatenated features. The application of the VTLN could mitigate the effect of SAT on the development set. Considering the MFCC features compared to the concatenated features, the relative improvements are more than 18% in all cases. According to [4, 8] the less improvement after the second pass is observed. However, the relative improvement slightly increases with the use of VTLN.

Table 4: *Absolute improvement [%] of the concatenated cepstral and MLP features over the single features systems, in brackets expressed in relative improvement [%]*

Features	dev07 SI	dev07 SA	eval07 SI	eval07 SA
MFCC $\leftrightarrow$ MFCC+DS	3.3 (19.1)	2.6 (18.2)	3.5 (21.6)	2.7 (20.6)
$\text{MFCC}_{\text{VTLN}}$ $\leftrightarrow$ $\text{MFCC}_{\text{VTLN}}$+ DS_{VTLN}	3.4 (20.6)	2.7 (18.9)	3.8 (23.9)	2.9 (22.1)
DS $\leftrightarrow$ MFCC+DS	1.0 (6.7)	1.6 (12.0)	0.8 (5.9)	1.4 (11.9)
DS_{VTLN} $\leftrightarrow$ $\text{MFCC}_{\text{VTLN}}$+ DS_{VTLN}	1.1 (7.7)	1.3 (10.1)	1.1 (8.3)	1.4 (12.0)

Nevertheless, achieving 11.6% WER on development and 10.2% on evaluation data set after the second pass, the $\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$ system hardly outperforms the system without VTLN (MFCC+DS), which eventually challenges the application of VTLN.

In [3] presented VTLN normalized cepstral features are concatenated with the voicedness feature providing a better $\text{MFCC}_{\text{VTLN}}$ system. Nonetheless, our concatenated features achieve even after the application of SAT significantly, rel. 17.1% better results (11.6%) than in [3] reported one (14.0%)

on the development set and clearly outperform the best reported recognition accuracy on eval07.

6.3. SAT/CMLLR transformed features as MLP input

For the investigation of the SAT/CMLLR transformed features as input to an MLP, the two best systems namely the SAT transformed MFCC+DS and the $\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$ are used. The MLP features trained on SAT transformed features are denoted as MLP(SAT(MFCC+DS) and MLP(SAT($\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$)), where SAT(.) means the CMLLR transformed features. The recognition results can be seen in Table 5.

Table 5: *Performance in WER [%] of the MLP transformed SAT features without or with VTLN, compared with the corresponding best system*

Features	dev07		eval07		dim/
	SI	SA	SI	SA	split
MFCC+DS	14.0	11.7	12.7	10.4	70/7
MLP(SAT(MFCC+DS))	12.9	12.6	11.5	11.4	27/8
$\text{MFCC}_{\text{VTLN}}$ +DS_{VTLN}	13.1	11.6	12.1	10.2	70/7
MLP(SAT($\text{MFCC}_{\text{VTLN}}$+ DS_{VTLN}))	12.6	12.6	11.3	11.2	28/8

On SAT transformed features trained MLPs are not able to achieve the performance of the previously reported best systems, although the feature space dimension is reduced to less than half. The achieved results are comparable to the accuracy showed by single DS_{VTLN} features after the second pass. Concatenation of the new features to different features —like DS to MFCC— does not show much gain over the best system, as well. The results are presented in Table 6 were computed using only MLP(SAT($\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$)) features. Although the relative 2.9% gain on the evaluation set is not indicated by the results on development data, the SAT($\text{MFCC}_{\text{VTLN}}$)+MLP(SAT($\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$)) system performs as well as the discriminatively trained and then combined systems in [3].

Table 6: *Performance in WER [%] of the MLP transformed SAT features with VTLN in concatenation with other features, compared with the previously best VTLN features system*

Features	dev07		eval07		dim/
	SI	SA	SI	SA	split
$\text{MFCC}_{\text{VTLN}}$ +DS_{VTLN}	13.1	11.6	12.1	10.2	70/7
$\text{MFCC}_{\text{VTLN}}$+ MLP(SAT($\text{MFCC}_{\text{VTLN}}$ +DS_{VTLN}))	12.1	11.6	10.6	10.2	73/7
SAT($\text{MFCC}_{\text{VTLN}}$)+ MLP(SAT($\text{MFCC}_{\text{VTLN}}$ +DS_{VTLN}))	11.8	11.6	10.1	9.9	73/7
SAT($\text{MFCC}_{\text{VTLN}}$+$\text{DS}_{\text{VTLN}}$)+ MLP(SAT($\text{MFCC}_{\text{VTLN}}$ +DS_{VTLN}))	11.9	11.5	10.3	10.3	98/7

7. Conclusions

This paper explored the application of SAT and VTLN on MLP features based systems on large vocabulary parliamentary speech recognition task. Recently, in state-of-the-art LVCSR system used MLP features were introduced and investigated on the English EPPS corpus using several speaker normalization methods. The new MLP features —in concatenation of cepstral features— significantly outperformed (rel. 17.1%) the previously reported results on this corpus.

Systematically investigating the effect of VTLN and SAT on MLP features based ASR system, our conclusions are:

- the use of VTLN and SAT is non-redundant on MLP features, but VTLN is redundant with SAT on cepstral features
- to achieve the best recognition performance (concatenated MLP and cepstral features) the use of VTLN is questionable with SAT
- the MLPs trained on SAT/CMLLR transformed features could not consistently contribute to better ASR performance.

Our future plans include the investigation of the effect of SAT and VTLN on other MLP front-end based ASR systems. We also intend to extend our study to MLPs trained on different domains.

8. Acknowledgement

This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. [213850]. 11, Speech Communication with Adaptive Learning - SCALE.

9. References

[1] S. Wegmann *et al.*, "Speaker normalization on conversational telephone speech," in *ICASSP*, vol. 1, 1996, pp. 339–341.

[2] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," in *Eurospeech*, 2003, pp. 1445–1448.

[3] J. Lööf *et al.*, "The RWTH 2007 TC-STAR evaluation system for European English and Spanish," in *Interspeech*, 2007, pp. 2145–2148.

[4] F. Valente *et al.*, "A comparative large scale study of MLP features for Mandarin ASR," in *Interspeech*, 2010, pp. 2630–2633.

[5] C. Plahl *et al.*, "Recent improvements of the RWTH GALE Mandarin LVCSR system," in *Interspeech*, 2008, pp. 2426–2429.

[6] M. Nußbaum-Thom *et al.*, "The RWTH 2009 QUAERO ASR evaluation system for English and German," in *Interspeech*, 2010, pp. 1517–1520.

[7] T. Schaaf and F. Metze, "Analysis of Gender Normalization Using MLP and VTLN Features," in *Interspeech*, 2010, pp. 306–309.

[8] Q. Zhu *et al.*, "On using MLP features in LVCSR," in *Interspeech*, 2004, pp. 921–924.

[9] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, vol. 3, 2000, pp. 1635–1638.

[10] F. Metze *et al.*, "The 2010 CMU GALE Speech-to-Text System," in *Interspeech*, 2010, pp. 1501–1504.

[11] J. Kittler *et al.*, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[12] H. Misra *et al.*, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *ICASSP*, vol. 2, 2003, pp. 741–744.

[13] F. Valente, "Multi-stream speech recognition based on Dempster-Shafer combination rule," *Speech Communication*, vol. 52, no. 3, pp. 213–222, Mar. 2010.

[14] P. Fousek *et al.*, "Transcribing broadcast data using MLP features," in *Interspeech*, 2008, pp. 1433–1436.

[15] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *ICASSP*, 2008, pp. 4729–4732.

[16] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Interspeech*, 2005, pp. 361–364.

[17] L. Welling *et al.*, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sep. 2002.