Using VTLN for Broadcast News Transcription

D.Y. Kim, S. Umesh, M.J.F. Gales, T. Hain and P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK.
{dyk21,su216,mjfg,th223,pcw}@eng.cam.ac.uk

Abstract

Vocal tract length normalisation (VTLN) is a commonly used speaker normalisation approach. It is attractive compared to many normalisation schemes as it is typically dependent on only a single parameter, allowing the *warp factors* to be robustly calculated on little data. However, the scheme normally requires explicitly coding the data at multiple warp factors. Furthermore, it is only possible to approximate the *Jacobian* associated with the VTLN transformation. A new, simple, linear approximation to VTLN is described in this paper. This linear approximation allows the *Jacobian* to be exactly computed. It can also be highly efficient in terms of warp factor estimation and application of the warp factors. Both the linear and standard CUED VTLN schemes were evaluated in the 2003 BNE evaluation framework and found to yield similar performance. When used in system combination both VTLN schemes yielded slight gains over the baseline system.

1. Introduction

Broadcast News (BN) transcription is one of the most challenging and interesting tasks in large vocabulary continuous speech recognition. Despite significant progress being made in reducing word-error rates, this task is still difficult because of the widely varying acoustic environment such as different speakers, speaking styles (read/spontaneous), background noise and/or music and different audio transmission channels (wideband/telephoneband). To overcome these problems there has been much interest in the use of normalisation [1, 2, 3] and adaptation [4] techniques to take into account this highly non-homogeneous data.

Among the widely used normalisation techniques, vocal tract length normalisation (VTLN) [1, 2, 3] is one of the most popular methods to reduce inter-speaker variability. VTLN is motivated by a desire to reduce inter-speaker variability that arises due to physiological differences in the vocal-tracts. VTLN is usually performed by warping the frequency-axis of the spectra of speakers/clusters by appropriate warp factor prior to the extraction of cepstral features. The warp factors are estimated by performing a maximum likelihood search with respect to a model and a transcription (and therefore may not have a relation with the physiological differences). The resulting speech features are usually less sensitive to inter-speaker variations.

Although many successful implementation and experimental results have been reported using VTLN, particularly for conversation telephone speech (CTS) recognition, there are only a few that have been reported for the BN task. VTLN was used in the 1998 CUED BN English (BNE) transcription system [5], but not in the 2003 system [6] where a more complex recognition framework was used. This paper describes experiments using VTLN within the 2003 BNE transcription framework and presents a simple linear VTLN approximation.

One of the problems with VTLN is that, since it is a transformation of the features, it is necessary to compute the *Jacobian* of the warping transformation. This is then used in the calculation of the likelihood to select the appropriate warp factors. However, as the VTLN transformation is typically non-linear, exact calculation of the Jacobian is highly complex and is normally approximated. This has lead many research groups [2, 3, 7] to explore the possibility of substituting the frequency-warping operation by a linear transformation in the cepstral domain. This allows the Jacobian to be exactly calculated. For the case of the bilinear transform [2] or piece-wise linear [3] VTLN warping can be expressed as a linear transformation in the linear (no mel-warping) cepstral domain. In [8] the problem of finding the transformation between VTLN-warped MFCC and the un-warped MFCC (where the MFCC is directly computed from the power-spectrum rather than using a filter-bank) is addressed. In [7], a linear-transformation between warped and un-warped cepstra is estimated such that the approximation error is minimised. In this paper a modified scheme for estimating a linear transformation to represent VTLN on any data is described.

An issue with applying VTLN to BN transcription is that, unlike CTS, there are no distinct, homogeneous, speaker-sides. Hence, it is important to have an automatic segmentation and clustering scheme to get clusters corresponding to similar speakers in similar acoustic conditions. There is also a practical issue associated with this. If standard, non-linear, VTLN schemes are used it is necessary to either explicitly cut the complete shows into segments (which may impact the calculation of dynamic coefficients), or code complete shows at all possible warp factors. Using a linear VTLN approximation overcomes this problem as it is simple to dynamically switch the linear transformation on demand.

This paper is organised as follows. First, an overview of conventional VTLN is presented and a description of the proposed linear VTLN scheme. In section 3, the CUED BNE transcription system is briefly reviewed. Section 4 gives detailed experimental results. Conclusions are then given.

2. Vocal Tract Length Normalisation

VTLN is usually performed by warping the frequency-axis of the spectra of the speaker by an appropriate warp factor prior to the extraction of cepstral features. The warp factors are estimated by maximising the likelihood of the warped utterances

This work was supported by DARPA grant MDA972-02-1-0013 under the EARS program. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors would like to thank all of the member of the HTK STT team. T. Hain is now at Sheffield University.

with respect to a model and some transcription. This model is usually either be based on a Gaussian mixture model, or hidden Markov model. The transcription is either known in training or obtained by an initial, non-VTLN, decoding pass for test data. If $\mathbf{x}$ and $\mathbf{x}^{\alpha}$ are the original and transformed feature vectors respectively then the log-likelihood is given by

$$\log(p(\mathbf{x};\alpha,\boldsymbol{\lambda})) = \log(J^{\alpha}) + \log(p(\mathbf{x}^{\alpha};\boldsymbol{\lambda})) \quad (1)$$

where J^{α} is the *Jacobian* of the VTLN transformation for warp factor α and $\boldsymbol{\lambda}$ is the set of model parameters. During estimation the aim is to find the value of the warp factor, α, that maximises this likelihood. The data is then warped using the warp factors and the VTLN models trained on this warped data.

The next section describes the standard implementation of VTLN at CUED. This is followed by a description of the modified linear VTLN scheme examined in this paper.

2.1. Conventional VTLN

The standard form of VTLN used in this paper is similar to that described in [5]. The warping is applied using the standard HTK scheme. Here a linear frequency warping operation is implemented by inversely scaling the centre frequencies and bandwidth of the filterbank prior to multiplication with DFT coefficients. The warp factors are assumed to lie in the range 0.8 to 1.2. The estimation of the appropriate warp factor for each homogeneous training segment is based on likelihoods from a HMM. Since the per-frame log-likelihood tends to be a parabolic function of the warp factor, warp factors were found by conducting a parabolic search (instead of a grid search) over data likelihoods versus warp factors. In order to overcome the issue of the Jacobian, cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN) is applied for each of the warp factors for each segment. By applying CVN the global dynamic range of each of the elements of the feature is scaled to be the same. This may be viewed as a crude way of normalising the data so that the Jacobian is roughly for the same for each warp factor. It may then be ignored. The following scheme is used to obtain the warp factors for the training data.

1. $\boldsymbol{\lambda}^{0}$ is set to an appropriate non-VTLN model set, $k = 0$.
2. Estimate the warp factor for each segment. For each hypothesised warp factor the data is re-aligned using this initial model and the transcription.

$$\alpha^{k+1} = \arg\max_{\alpha}\left(p(\tilde{\mathbf{X}}^{\alpha};\boldsymbol{\lambda}^{k})\right) \quad (2)$$

   where $\tilde{\mathbf{X}}^{\alpha} = \{\tilde{\mathbf{x}}_1^{\alpha},\ldots,\tilde{\mathbf{x}}_T^{\alpha}\}$ is the warped data (with CMN and CVN) associated with the segment.
3. Warp the training data. A new model set, $\boldsymbol{\lambda}^{k+1}$, is then trained using single pass retraining and standard Baum-Welch estimation.
4. $k = k + 1$. Goto (2) until warp factors have stabilised.

The final set of warp factors is then quantised to a resolution of 0.01, yielding a total of 41 possible values. Using this final set of VTLN warp factors a new decision tree is estimated. This is then used in the standard training schemes. For this work about four iterations were required for the warp factors to stabilise.

2.2. Linear VTLN

In linear VTLN (LVTN) a linear transformation is used to approximate the complex, non-linear, warping of the frequency axis. In this implementation, rather than transforming the data to the VTLN domain, the linear transform is trained to effectively "un-warp" the observed data. Thus

$$\mathbf{x}^{\overline{\alpha}} = \mathbf{A}^{\alpha}\mathbf{x} + \mathbf{b}^{\alpha} = \mathbf{W}^{\alpha}\boldsymbol{\zeta}^{\alpha} \quad (3)$$

where $\boldsymbol{\zeta}$ is the extended feature vector and $\overline{\alpha}$ is used to indicate the inverse warping to standard VTLN. With this form of linear transformation, the Jacobian has a simple closed form solution $J^{\alpha} = |\mathbf{A}^{\alpha}|$. The following scheme is used to obtain the linear transformations and warp factors.

1. $\boldsymbol{\lambda}^{0}$ is set to an appropriate non-VTLN model set $k = 0$.
2. Randomly select a subset of training data. For each warp factor α compute the set of warped feature vectors $\tilde{\mathbf{X}}^{\alpha}$.
3. For each α compute the linear transform, $\mathbf{W}^{\alpha}$,

$$\mathbf{W}^{\alpha} = \arg\max_{\mathbf{W}}\left(p(\tilde{\mathbf{X}}^{\alpha};\boldsymbol{\lambda}^{k},\mathbf{W})\right) \quad (4)$$

   This can be estimated using the standard constrained MLLR transform formulae [4].
4. For each segment of data estimate the warp factor

$$\alpha^{k+1} = \arg\max_{\alpha}\left(p(\mathbf{X};\boldsymbol{\lambda}^{k},\mathbf{W}^{\alpha})\right) \quad (5)$$

5. Linearly warp the training data. A new model set, $\boldsymbol{\lambda}^{k+1}$, is then trained on using single pass retraining and standard Baum-Welch estimation.
6. $k = k + 1$. Goto (3) until warp factors have stabilised.

As the transformation matrices undoes the effects of the VTLN warping the warp factors that are estimated at stage (4) will be the inverse of the warp factor estimated using conventional VTLN. In order to be able to directly relate the LVTN to VTLN the inverse of the LVTN warp factors were estimated. The set of possible warp factors was set to be the same as that of the conventional VTLN.

The warp factor estimation for LVTN may be performed in the same fashion as conventional VTLN[1]. However a more efficient approach is possible using the auxiliary function that is used to estimate the transform. The auxiliary function is

$$\mathcal{Q}(\hat{\alpha},\alpha) = \beta\log(|\mathbf{A}^{\alpha}|) - \frac{1}{2}\sum_{i=1}^{d}\left(\mathbf{w}_i^{\alpha}\mathbf{G}^{(i)}\mathbf{w}_i^{\alpha\prime} - 2\mathbf{w}_i^{\alpha}\mathbf{k}^{(i)\prime}\right)$$

where $\beta = \sum_{m,\tau}\gamma_m^{\hat{\alpha}}(\tau)$, $\mathbf{G}^{(i)}$ and $\mathbf{k}^{(i)}$ are defined in [4] and $\gamma_m^{\hat{\alpha}}(\tau)$ is the component posterior given the current estimate of the warp factor, $\hat{\alpha}$. Thus using the sufficient statistics the auxiliary function can be used to estimate the warp factor in a single pass. This can then be refined by obtaining new component posteriors with this estimate.

3. BNE 10xRT Framework

The system used for the experiments was developed for the March 2003 Rich Transcription (RT03) evaluation, and employs a structure of multi-branch, multi-pass and system-combination for improved accuracy. Full details of the system structure and the models involved are given in [6]. PLP coefficients with first, second and third derivatives projected down to 39

[1]For the implementation used for the experiments an exhaustive search for LVTN was used. This yielded no difference in performance to the conventional VTLN search.

dimensions using HLDA are used as the acoustic features. The cross-word triphone HMMs which contained about 7000 states each with 16 Gaussians were estimated using the English BN data released by the LDC in 1997 and 1998. Since some of BN data, for example telephone interviews, is transmitted over bandwidth-limited channels, both wideband and narrowband spectral analysis variants of each model set were trained. All model sets were trained using MPE and gender-dependent versions were derived using MPE-MAP. A number of broadcast and newswire text corpora were used to train a word 4-gram language model and a class trigram model. The systems were evaluated on two 3 hour test sets the 2003 development, `dev03`, and evaluation, `eval03`, test sets. The system structure is shown in Figure 1.

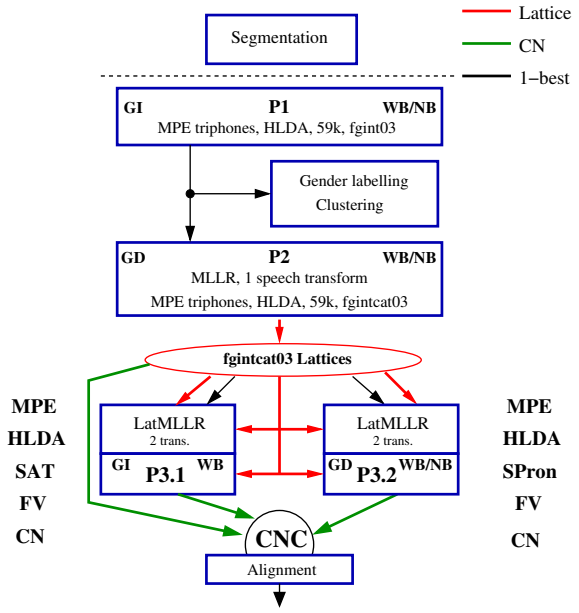

Figure 1: CUED BN-E system structure.

P1 initial transcription: The purpose of the P1 pass is to provide an initial word-level transcription as the supervision for gender determination and clustering for adaptation of the P2 models. The adaptation uses global least squares regression mean transforms and MLLR variance transforms.

P2 lattice generation: Word lattices are generated using the adapted acoustic models and the 4-gram word LM. The associated 1-best hypotheses are used in the estimation of up to two speech MLLR transforms.

P3 lattice rescoring: Two separate model sets are used to rescore the P2 lattices. The P3.1 system was built using Speaker Adaptive Training (SAT) employing global constrained MLLR transforms. The P3.2 system was trained in the normal speaker-independent fashion but employed a special single pronunciation (SPRON) dictionary. Both P3 model sets were adapted using lattice MLLR and a global full-variance transform, then used to rescore the word lattices from P2.

The final system output was derived by combining the confusion networks generated by the P2, P3.1 and P3.2 passes using Confusion Network Combination (CNC). Finally, a forced alignment of the final word-level output was used to obtain accurate word times before scoring. The full system ran in 9.1×RT on the 2003 evaluation set.

4. Experimental results

4.1. Estimated Warp Factors

It is interesting to see how closely related the two sets of estimated warp factors are for conventional VTLN and linear VTLN. A set of warp factors was estimated using each of the conventional VTLN and LVTN with a block-diagonal transformation matrix (static, delta, delta-deltas and third derivatives where used) and no bias. Note for the LVTN scheme the inverse of the estimated warp factor is used for consistency with conventional VTLN.

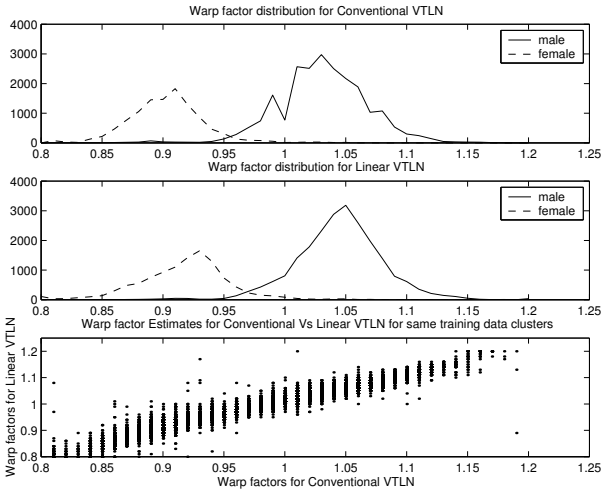

Figure 2: Training data warp factors for VTLN (top) and LVTN (middle). A scatter plot of VTLN and LVTN is shown at the bottom.

Figure 2 shows the two sets of warp factors. The top diagram shows the warp factors estimated using conventional VTLN. The standard bimodal distribution can be seen, one peak is associated with the male speakers the other with the female speakers. A similar distribution can be seen for the the LVTN approach. The bottom plot of figure 2 shows a plot for each of the segments of the VTLN against the LVTN warp factor. The two estimates are highly correlated with one another with a correlation coefficient of 0.9812.

4.2. Unadapted Results

Initial comparisons of the system did not make use of the full BNE 10x framework. Instead experiments using unadapted models and a trigram language model were conducted to allow a simple comparison of the two schemes without further adaptation. The first issue to be addressed for VTLN normalisation is to determine the appropriate amount of data to compute the warp factors. In initial experiments it was found that a reasonable value was to use the segments obtained from the standard segmenter and then clustered together to yield a minimum occupancy of 500. This clustering was used for all VTLN and LVTN experiments.

Table 1 shows the performance of segment level CMN (CMN), used in the baseline BNE configuration and cluster level CMN and CVN (CMVN), with a minimum occupancy

Configuration		Front-end			
		CMN	CMVN	VTLN	LVTN
dev03	MLE	19.7	19.1	18.4	18.1
	+HLDA	17.9	17.9	16.9	17.1
	+MPE	15.2	15.3	14.6	14.6
	+MAP	14.9	—	14.5	14.4
eval03	MLE	17.8	17.1	16.5	16.4
	+HLDA	15.9	15.9	14.9	14.9
	+MPE	13.7	13.7	13.2	13.0
	+MAP	13.4	—	13.0	12.7

Table 1: %WER of dev03 and eval03 with CMN, CMN+CVN (CMVN), conventional VTLN and linear VTLN (LVTN) acoustic models.

	dev03	eval03
P1	15.9	14.6
P2	12.7	11.6
P3.0 (CMN)	12.3	11.2
P3.1 (SAT)	12.3	11.0
P3.2 (SPRON)	12.0	11.1
P3.3 (VTLN)	11.9	10.8
P3.4 (LVTN)	12.1	10.9
P2+P3.1+P3.2	11.6	10.6
P2+P3.1+P3.2+P3.3	11.4	10.4
P2+P3.1+P3.2+P3.4	11.4	10.3

Table 2: %WER for the 10xRT system.

count of 500. The performance gain of CMVN varies according to the complexity of the system. For the standard MLE 16 component system (MLE) the CMVN system outperformed the baseline CMN scheme. However after training an HLDA transform the performance was about the same. Similarly after using MPE training the error rates were about the same. Thus for this task there is no advantage in using CMVN over the standard CMN for the more complex models.

Table 1 also shows the performance of the VTLN and LVTN. Comparing the VTLN performance to the CMN and CMVN schemes shows that using VTLN on top of the mean and variance normalisation yields consistent gains over all the systems. Comparing the performance of VTLN and LVTN shows that the two schemes yield about the same performance. For MPE training the two schemes show a gain of 0.6% absolute over the baseline CMN approach for dev03 and about 0.5% for eval03. Many BNE systems use gender dependent (GD) models. For table 1 GD models were generated using MPE-MAP. Even with GD models, which take into account the most significant speaker difference, VTLN and LVTN both show gains over the baseline CMN front-end.

4.3. 10xRT system results

The use of VTLN and LVTN was then examined in the BNE 10x framework. The models used for these schemes were more complicated than those presented in table 1 to be consistent with the standard CMN system. Prior to MPE training *varmix*, a scheme for more distributing the number of components per state according to the amount of data associated with a state, was used. In addition *lattice regeneration* was used[2]. Here, after performing initial MPE training, the lattices are regenerated. New models are then trained on a combination of these new and original lattices. These two additional techniques reduced the error rated by about 0.3-0.4% absolute.

Table 2 shows the results of using the VTLN and LVTN systems in the 10x BNE framework. The two approaches are added as additional paths in the third stage, labelled P3.3 (VTLN) and P3.4 (LVTN). Within this more complex framework the individual VTLN (P3.3) and LVTN (P3.4) systems outperform the standard CMN system (P3.0). The performance of the two VTLN systems at the P3 stage are similar. Though in contrast to the unadapted results VTLN shows small gains over LVTN.

Using either of the VTLN approaches as an additional system for combination reduced the final error rate by about 0.2% absolute. This was consistent over both dev03 and eval03. The best performance, a gain of 0.3% absolute was obtained using LVTN as an additional stage for system combination.

[2]The SAT system presented does not use lattice regeneration.

5. Conclusions

In this work the use of VTLN and a linear VTLN method for broadcast news transcription was investigated. A linear VTLN scheme was described that can be simply applied to any data. Experimental results showed that both VTLN models consistently outperformed non-VTLN models. The proposed linear VTLN showed comparable performance with conventional VTLN in terms of WERs while reducing computational time for re-coding whole training data.

6. References

[1] L. Lee and R.C.Rose, "Speaker Normalisation Using Efficient Frequency Warping Procedures," *Proc. ICASSP96*, Atlanta, GA, 1996.

[2] J. McDonough, W. Byrne and X. Luo, "Speaker Normalization with All-pass Transforms," *Proc. ICSLP98*, Sydney, Australia, 1998.

[3] M. Pitz and H. Ney, "Vocal Tract Normalization as Linear Transformation of MFCC," *Proc. EuroSpeech2003*, Geneva, Switzerland, 2003.

[4] M.J.F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," *Computer Speech & Language*, 12:75–98, 1998.

[5] T. Hain, P.C. Woodland, T.R. Niesler and E.W.D. Whittaker, "The 1998 HTK System for Transcription of Conversational Telephone Speech," *Proc. ICASSP99*, Phoenix, AZ, 1999.

[6] D.Y. Kim, G. Evermann, T. Hain, D. Mrva, S. Tranter, L. Wang and P.C. Woodland, "Recent Advances in Broadcast News Transcription," *Proc. ASRU2003*, St. Thomas, U.S. Virgin Islands, 2003.

[7] L.F. Uebel and P.C. Woodland, "An Investigation into Vocal Tract Length Normalisation," *Proc. EuroSpeech99*, Budapest, Hungary, 1999.

[8] M. Pitz, S. Molau, R. Schlter, H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *Proc. EuroSpeech2001*, Aalborg, Denmark, 2001.