AN INVESTIGATION INTO VOCAL TRACT LENGTH NORMALISATION

L.F. Uebel & P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
Email: {lfu20,pcw}@eng.cam.ac.uk

10.21437/Eurospeech.1999-553

ABSTRACT

This paper investigates several different methods for performing vocal tract length normalisation (VTLN) which are either completely linear or piece-wise linear. Furthermore the combination of VTLN with either standard unconstrained maximum likelihood linear regression (MLLR) or constrained MLLR is considered. Results on the Switchboard corpus show that there is little difference in performance between the different forms of VTLN, and that as previously reported that the effects of VTLN and unconstrained MLLR are largely additive. However it was found that if multiple iterations of constrained MLLR is used there is no additional advantage to also using VTLN.

1. INTRODUCTION

Vocal Tract Length Normalisation (VTLN) is now a fairly widely used speaker normalisation technique (e.g. [8]). However, there are still a number of issues concerning VTLN that are not well understood. This paper discusses some of these issues and also performs an experimental comparison of several VTLN-type techniques: a direct frequency domain warping; the bilinear transform; and a square matrix linear transform approximation to the first two techniques. One aim of this comparison is to determine whether the non-linear aspects of frequency-scaling based VTLN offer any advantage.

Often in VTLN the appropriate warping factor is selected by maximising the likelihood of the data: however this likelihood should be computed while accounting properly for the effect of the transformation matrix. Here the use of variance normalisation was evaluated which approximately performs the appropriate normalisation for likelihood comparisons.

Previously it has been noted that the combination of VTLN with Maximum Likelihood Linear Regression (MLLR) has yielded near additive improvements. The paper investigates this claim further in the case of both constrained MLLR [3] and unconstrained MLLR [5, 2].

The paper first describes the different VTLN approaches and then reviews constrained and unconstrained MLLR adaptation. Experimental results using these techniques are presented using the MiniTrain training and test subsets of the Switchboard corpus.

2. VOCAL TRACT NORMALISATION

Vocal Tract Length Normalisation (VTLN) has the goal of *normalise* the speech spoken by speakers by effecting a frequency warping to account for the differing vocal tract lengths. There have been a number of different approaches investigated to this problem for both the form and implementation of the warping function as well as the method of selecting the warping parameter (or scale factor) for a particular utterance.

Here three different procedures for VTLN have been implemented. In each case a range of frequency warping values are selected by calculating the most likely warped speech by performing a grid search over differing warp factors. The likelihood calculations in each case used a forced alignment procedure with a set of HMMs using the recognised output from an initial recognition pass.

The first VTLN technique implemented applies a direct frequency domain warping. There is a linear frequency warping from the origin to a user-defined breakpoint frequency and then a second stage so that the maximum frequency in the warped and unwarped spectrum correspond.

The second uses the bilinear transform approach [7]. This can be implemented using a linear transformation of the cepstral coefficients, but it requires the retention of higher order unwarped cepstra to those that are needed after warping.

Finally, to determine the effects of using a linear transformation applied to just the cepstra used in recognition, a matrix approximation technique was used to learn the mapping (for a particular warp factor) between the unwarped and warped cepstra by doing a least squares estimation of the transformation matrix.

2.1. Direct Frequency Domain Warping

One often used method for VTLN is to directly scale the frequency axis so that $f_{OUT} = \alpha\, f_{IN}$: however care needs to be taken at the limits of the frequency range. In the implementation used here, the frequency axis is scaled until a particular point and then the scaling is changed so that the frequency scales match at a particular cut-

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

off (F_{HIGH}). Figure 1 illustrates this Direct Frequency Domain Warping (DF) technique.

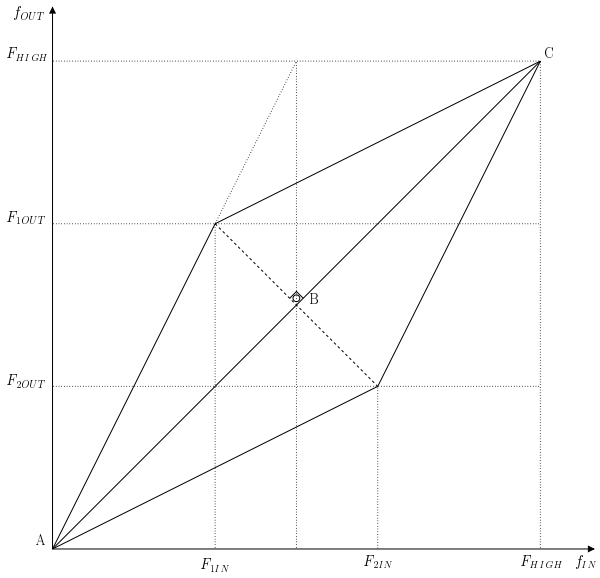

Figure 1: Graphic Description of Direct Frequency Domain Warping.

In DF warping basic trigonometric relationships are used to calculate the points F_{1IN}, F_{2IN}, F_{1OUT} and F_{2OUT} (input and output breakpoint frequencies), as well as the required scaling to match the frequency scales at the limits. To avoid numerical problems, the amplitude at f_{OUT} is calculated over a grid of points using a linear regression of the corresponding f_{IN} amplitudes of the original speech. The FFT size is chosen to give the required frequency resolution for accurate warping.

In the experiments reported here, a fixed breakpoint at 90% of the frequency scale was used and it was found to produce similar results to a more flexible approach so that there was just one free parameter in the warping process.

2.2. Bilinear Transformation

The bilinear transformation approach to VTLN [1, 7] implements a bilinear filter ($Q_\alpha(z) = (z+\alpha)/(\alpha z+1)$) for each warping factor α and can be implemented as a matrix transformation [7] directly in cepstral domain i.e. the transformation is linear.

There exists a closed form equation to compute each element of the warping matrix. The warped cepstral vectors are found using the warping matrix multiplied by unwarped cepstral vector:

$$\begin{bmatrix} s_{11} & \dots & s_{1n} \\ \vdots & \vdots & \vdots \\ s_{m1} & \dots & s_{mn} \end{bmatrix} \times \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \tilde{c}_1 \\ \vdots \\ \tilde{c}_m \end{bmatrix}$$

where c_n's are unwarped cepstral coefficients, s_{mn}'s are warping matrix elements, $\tilde{c}_m$'s are warped cepstral coefficients, n is the number of original cepstral coefficients and m is the number of warped coefficients. Note that extra unwarped cepstral coefficients are required to avoid loss of frequency resolution and here a warping matrix 12×24 elements was used. The transformation matrix is approximately band diagonal with seven significant elements in the band. Advantages of the method include the fact that it operates directly on cepstra and that an appropriate warping matrix for any value of the warping parameter is easy to calculate.

2.3. Linear Cepstral Approximation

To study the effect of non-linearity on the transform and to operate on a fixed set of 12 cepstra a linear approximation technique to the DF and bilinear transform methods was used. The matrix was estimated for particular warping factors so that the approximation error between the warped and unwarped cepstra is minimised. The warped cepstra are found by

$$\begin{bmatrix} c_{11} & \dots & c_{1n} \\ \vdots & \vdots & \vdots \\ c_{m1} & \dots & c_{mn} \end{bmatrix} \times \begin{bmatrix} k_{11} & \dots & k_{1n} \\ \vdots & \vdots & \vdots \\ k_{n1} & \dots & k_{nn} \end{bmatrix} = \begin{bmatrix} \tilde{c}_{11} & \dots & \tilde{c}_{1n} \\ \vdots & \vdots & \vdots \\ \tilde{c}_{m1} & \dots & \tilde{c}_{mn} \end{bmatrix}$$

where c_{nm}'s are unwarped cepstral coefficients, k_{nm} are warping matrix elements, $\tilde{c}_{mn}$ are warped cepstral coefficients, n is number of cepstral parameters and m is number of speech frames.

The transformation matrix can be calculated using a Moore-Penrose pseudoinverse approach, $K = C^\dagger\ \tilde{C}$. Unfortunately this is computationally expensive to use directly but can be computed using $C^\dagger = (C^T C)^{-1} C^T$ [6] and transformation matrix is given by

$$K = (C^T C)^{-1} C^T \tilde{C}$$

2.4. Variance Normalisation

In each case the optimal warping factor is selected by comparing the likelihood of the warped data for a number of candidate warping factors. However, when a transformation is applied in the data the likelihood of data should be computed as [9]

$$P'(\hat{x}) = J(\alpha) * P(\hat{x})$$

where $P(\hat{x})$ is likelihood of warped data $\hat{x}$ and $J(\alpha)$ is the Jacobian of the transformation (whether it is linear or non-linear). In the case of a linear transform, the Jacobian is given by the determinant of the transformation matrix.

Since the appropriate normalisation can only be directly applied to a linear transformation, normalisation of the variance of warped cepstra (as in [4]) was used before likelihood computation to obtain a similar effect as using the Jacobian.

3. MAXIMUM LIKELIHOOD LINEAR REGRESSION

MLLR estimates a linear transform which maximises the likelihood of adaptation data. Two different approaches can be used to perform MLLR. Unconstrained MLLR [5, 2] implements an unrelated transform of means

and variances whereas constrained MLLR [3] maximises the likelihood by transforming the observation data (or equivalently using a related transform for the means and variances).

3.1. Unconstrained MLLR

Unconstrained MLLR transforms the HMM mean vectors (μ) and diagonal covariance matrix Σ using separate unrelated transforms and is a flexible general method. The transformation is given by

$$\hat{\mu} = A\mu + b$$
$$\hat{\Sigma} = H\Sigma H^T$$

where A is a square transformation matrix and b is an offset vector and H is the transformation for the variance. The estimation of the transformation matrices uses an expectation-maximisation (EM) approach that allows a closed form solution of the maximisation of the associated auxilary function.

3.2. Constrained MLLR

Constrained MLLR requires that the transformation applied to the variance corresponds to that applied to the means. This is achieved by applying a transformation to the observation vectors so that for a particular Gaussian which uses a transformation A and offset b then the observation vector used by that Gaussian $\hat{o}(\tau)$ is given by

$$\hat{o}(\tau) = Ao(\tau) + b$$

where $o(\tau)$ is the observation vector produced by the front-end processing. The optimisation procedure for constrained MLLR again uses an EM approach, but in this case an iterative method of optimisation of the auxiliary function is required.

4. EXPERIMENTAL SETUP

The various techniques for VTLN along with the use of both constrained and unconstrained MLLR were evaluated using training and test subsets of the the Switchboard-I corpus. The training subset (referred to as MiniTrain) covers 398 sides containing 17.8 hours of speech and is approximately gender balanced.

Each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 (including c_0) MF-PLP cepstral parameters, computed from a filterbank spanning 125-3800Hz, and their first and second differentials. The cepstra used per conversation side mean and variance normalisation.

For testing a gender balanced half-hour set (MTtest) containing Switchboard-I data was used. All recognition experiments were conducted using a 2 million word Switchboard trigram language model that used a 22k word vocabulary and a pronunciation dictionary based on the 1993 LIMSI pronunciation dictionary and rescored lattices generated by a gender independent system.

The set of HMMs used in these experiments were those produced for the front-end experiments in [4]. The basic gender independent state-clustered cross-word triphone HMMs used 2945 speech states and 12 Gaussians per state. These models did not use any form of VTLN in training and are denoted (GIU). Further corresponding sets of gender independent model trained on VTLN data (GIW) as well as gender dependent sets trained on both unwarped data (GDU) and warped data (GDW) were also used.

5. VTLN AND UNCONSTRAINED MLLR RESULTS

Results using the different HMM sets with various types of warping are shown in Table 1 which include the direct frequency domain warping (DF), the bilinear transformation (BL) and also linear (square matrix) approximations to bilinear transformation (ABL), and linear approximation to DF warping (ADF). The results of applying unsupervised mean or mean and variance MLLR using block-diagonal mean transformations and diagonal variance transforms are also shown. A separate transform was used for silence models and a single global transform used for speech states.

			Warped			
	MLLR	U	BL	DF	ABL	ADF
GIU	—	44.59	43.11	43.19	43.18	43.43
	Mean	41.45	40.86	40.68	41.04	41.25
	M+Var	41.09	40.17	40.43	40.48	40.50
GIW	—	47.50	42.63	41.97	42.69	41.75
	Mean	41.31	39.54	39.49	39.77	39.73
	M+Var	41.17	39.41	39.21	39.63	39.73
GDU	—	43.21	41.62	41.69	41.70	41.65
	Mean	40.23	39.71	39.59	39.90	40.28
	M+Var	40.10	39.29	39.45	39.43	39.81
GDW	—	45.28	41.17	40.71	41.53	40.51
	Mean	40.64	38.79	38.88	38.74	39.02
	M+Var	40.59	38.72	38.24	38.76	39.07

Table 1: % word error rates for unwarped (U) data and warped data using VTLN and Unconstrained MLLR.

The various types of VTLN all perform similarly as do the square matrix linear approximation to either BL or the DF warping. The best error rates were obtained using DF warping with mean and variance unconstrained MLLR adaptation and gender dependent HMM models.

For the gender independent case, unconstrained MLLR reduced the word error rate (WER) over the unwarped recognition results by 3.5% absolute. Note that GIW and GDW represent a mismatch between warped training data and testing data that is not warped, but unconstrained MLLR can largely compensate for this.

The joint improvement from using both VTLN and MLLR is, in many cases, largely additive. For example, with gender independent models use of BL warping gives a 1.96% absolute improvement (GIU unwarped to GIW BL warped) and mean and variance MLLR (without warping) 3.50%. The result of using both gives an improvement of 5.18% (GIU unwarped, no MLLR to GIW BL warped with mean and variance MLLR).

6. CONSTRAINED MLLR RESULTS

The aim of the constrained MLLR experiments is to see if in this case the largely additive improvements would again be apparent. Since the various types of VTLN gave similar error rates, only BL warping was used here. Furthermore, the results enable constrained and unconstrained MLLR to be compared. To try and closely compare the effects of VTLN and constrained MLLR either a single (for speech and silence) or separate speech and silence transforms were used.

Since it was found that using multiple iterations of constrained MLLR (using the same phone-level transcriptions but changing the state-frame alignments and re-estimating the transform matrices) gave reduced word error rates these results are also included. Interestingly this improvement due to multiple iterations was not observed with unconstrained MLLR.

	Its	Unwarped		BL Warped	
	MLLR	1 Tran	2 Tran	1 Tran	2 Tran
GIU	—	44.59	44.59	43.11	43.11
	1	42.64	41.95	41.31	41.42
	2	42.23	41.72	41.44	41.39
	6	41.84	40.62	41.72	40.53
GIW	—	47.50	47.50	42.63	42.63
	1	42.45	41.94	40.54	40.42
	2	42.03	41.18	40.90	40.12
	6	40.70	39.95	40.79	39.99
GDU	—	43.21	43.21	41.62	41.62
	1	41.45	40.76	40.40	39.93
	2	41.26	40.46	40.43	39.77
	6	40.60	39.92	40.09	39.18
GDW	—	45.28	45.28	41.17	41.17
	1	41.45	41.39	39.68	39.48
	2	40.62	40.53	39.40	39.01
	6	39.88	38.96	39.54	38.94

Table 2: %WER for constrained MLLR using 1 and 2 Transforms with unwarped or BL warped data for various model sets.

Table 2 shows the results for unwarped data in the first two columns and then warped data in the second two. In each case it can be seen (especially with unwarped data) that the use of multiple iterations (up to 6) of constrained MLLR is advantageous. The use of two transforms can also be seen to be advantageous in all cases.

After 6 iterations of constrained MLLR for all conditions (except gender dependent unwarped models) there is no significant advantage to using VTLN as well as constrained MLLR. This might be as expected since constrained MLLR is performing a very similar operation to a linear transform VTLN but is somewhat more powerful since it allows both a more general linear transform than the BL transform (although no higher cepstra are used) and estimates separate transforms for each of the static cepstra, the first and the second differential coefficients. Therefore it would be expected that estimating both sets of transforms would be no better than just constrained MLLR unless the effect of initialisation is of key importance. It appears that this is the case given that many iterations of constrained MLLR are required.

It may also be noted that constrained MLLR gives slighly smaller improvements than unconstrained MLLR even with multiple iterations.

7. CONCLUSIONS

This paper presented some results using various forms of VTLN, unconstrained MLLR and constrained MLLR. It was found that the different VTLN methods gave very little difference in performance.

As has been previously reported, the effects of VTLN and unconstrained MLLR are largely additive. However it was found that if multiple iterations of constrained MLLR are used then there was no advantage for also using VTLN.

Acknowledgements

Luis Felipe Uebel is funded by a CNPq (Brazilian Council of Research) scholarship, Brazil. BBN supplied the definitions for the MiniTrain training and test sets.

8. REFERENCES

1. A. Acero (1990) *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University.
2. M.J.F. Gales & P.C. Woodland (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech and Language*, Vol. 10, pp. 249-264.
3. M.J.F. Gales (1998) Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition. *Computer Speech and Language*, Vol. 12, pp. 75-98.
4. T. Hain, P.C. Woodland, T.R. Niesler & E.W.D. Whittaker (1999) The 1998 HTK System for Transcription of Conversational Telephone Speech. *Proc. ICASSP'99*, pp. 57-60, Phoenix.
5. C.J. Leggetter & P.C. Woodland (1995) Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. *Computer Speech and Language*, Vol. 9, pp. 171-185.
6. J.R. Magnus & H. Neudecker(1998) *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, New York.
7. J.W. McDonough, G. Zavaliagkos & H. Gish. An Approach to Speaker Adaptation based on Analytic Functions. *Proc. ICASSP'96*, pp. 721-724, Atlanta.
8. D. Pye & P.C. Woodland. Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition. *Proc. ICASSP'97*, pp. 1047-1050, Munich.
9. S. Ross (1988). *A First Course in Probability.* Macmillan, New York.