VTLN WARPING FACTOR ESTIMATION USING ACCUMULATION OF SUFFICIENT STATISTICS

Jonas Lööf and Hermann Ney

Lehrstuhl für Informatik VI, Comp. Sci. Dept.
RWTH Aachen Univ, 52056 Aachen, Germany
{loof, ney}@cs.rwth-aachen.de

Srinivasan Umesh

Department of Electrical Engineering
Indian Institute of Technology, Kanpur, India
sumesh@iitk.ac.in

ABSTRACT

In this paper we present an efficient and flexible approach to VTLN warping factor estimation. Due to the equivalence of frequency warping and linear transformation of cepstral coefficients, warping factors can be efficiently estimated by accumulating the sufficient statistics for linear transformation estimation, and searching the constrained space of transformations given by the explicit mapping between warping factors and linear transformation matrices. We show that the positive effect of using a properly normalized optimization criterion for warping factor estimation, which has been previously demonstrated for a signal analysis front-end without a filterbank, carries over to a MFCC front-end, resulting in a net improvement in word error rate.

1. INTRODUCTION

Vocal tract length normalization (VTLN) [1] is an important method to compensate for inter-speaker variation in speaker-independent automatic speech recognition (ASR). To achieve this, the frequency axis is warped using a parameterized invertible function, and the parameter, or *warping factor*, is optimized for each speaker.

The equivalence of VTLN and linear transformation for a general frequency warping was demonstrated in [2]. This work was later refined in [3] to explicitly take into account the frequency discrete nature of the ASR signal processing front-end. We briefly review the refined method.

As described in [3], if the spectrum is quefrency limited, samples of the warped spectrum can be exactly obtained from the unwarped spectrum. For plain cepstral coefficients (without filterbank smoothing and discrete cosine transform (DCT)) the cepstrum is computed from the spectrum using

$$C_k = \frac{1}{N}\sum_{q=0}^{N-1} \log|X[q]|^2 e^{+j\frac{2\pi}{N}qk}, \tag{1}$$

where C_k are the cepstral coefficients and $X[q]$ is the spectrum. Since C_k and $\log|X[q]|^2$ form a discrete Fourier transform pair we can recover the second from the first. $X[q]$ cannot be recovered though, because of the magnitude operation. We are interested only in the warped log-magnitude spectrum $\log|\tilde{X}[q]|^2$ though, and $\log|\tilde{X}[q]|^2$ can be exactly reconstructed from $\log|X[q]|^2$ if C_k is quefrency limited and unaliased. If this condition holds the warped spectrum can be computed directly, by

$$\begin{aligned}\log|\tilde{X}[l]|^2 &= \log|\tilde{X}(\tilde{\omega}_l)|^2 = \log|X(g^{-1}(\omega_l))|^2 \\ &= \sum_{k=0}^{N-1} C_k e^{-j\frac{2\pi}{N}g^{-1}(\omega_l)k}.\end{aligned} \tag{2}$$

By substituting (2) into (1) a linear transformation between C_k and the warped cepstral coefficients $\tilde{C}_n$ is reached, i.e.

$$\begin{aligned}\tilde{C}_n &= \sum_{k=0}^{N-1} C_k \frac{1}{N}\sum_{l=0}^{N-1} e^{-\frac{2\pi}{N}g^{-1}(\omega_l)k} e^{+j\frac{2\pi}{N}ln} \\ &= \sum_{k=0}^{N-1} W_{nk} C_k.\end{aligned} \tag{3}$$

To use the above relation with a typical ASR system using DCT based cepstral coefficients derived from filter-bank smoothed spectra, a relation between plain- and DCT cepstra can be derived. The DCT based cepstral coefficients are given by

$$d_k = \sum_{q=0}^{M-1} \log|X_{FB}[q]|^2 \cos\frac{k\pi(q+1/2)}{N}. \tag{4}$$

Similarly the plain cepstra of the filter bank output is given by

$$C_k = \frac{1}{2(M-1)}\sum_{q=0}^{M-1} b_q \log|X_{FB}[q]|^2 \cos\frac{qk\pi}{M-1}, \tag{5}$$

where $b_q = 1$ for $q = 0$ or $q = (M-1)$ and 2 otherwise. From these two equations a linear transformation relation between DCT- and plain cepstra can be derived. Furthermore, combining this transformation and its inverse with (3), a linear transform between warped and unwarped DCT cepstra is reached. It should be noted that the above holds for any invertible warping function.

This work was partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org)

Using these results, warping factor estimation is seen as a constrained linear transform estimation, where the constraint is given by the mapping between warping factors and transformation matrices as given above. It is sufficient to accumulate the sufficient statistics needed for estimating linear transformations for each speaker, and perform the constrained optimization off-line.

As presented in [4] the auxiliary function for estimating a linear feature transform using the expectation maximization (EM) algorithm is given by

$$Q(M, \hat{M}) = \beta \log \left|W\right|^2 - \frac{1}{2} \sum_{d=1}^{D} \left(w_d G^{(d)} w_d^T - 2 w_d k^{(d)T} \right), \tag{6}$$

where terms constant with respect to the transform W has been omitted. The sufficient statistics, G and k, are given by

$$G^{(d)} = \sum_{s=1}^{S} \frac{1}{\sigma_d^{(s)2}} \sum_{t=1}^{T} \gamma_s(t) x_t x_t^T \tag{7}$$

$$k^{(d)} = \sum_{s=1}^{S} \frac{1}{\sigma_d^{(s)2}} \mu_d^{(s)} \sum_{t=1}^{T} \gamma_s(t) x_t^T, \tag{8}$$

and β is $\gamma_s(t)$ accumulated over s and t.

VTLN warping factors are usually estimated using grid search by directly evaluating the acoustic scores when aligning a reference transcription with the speaker independent (SI) acoustic model. As has been pointed out earlier [2], this fails to take into account the Jacobian determinant of the warping transformation, and thus fails to properly normalize the model distributions. Although previous studies [2] showed only small performance gains in using the Jacobian, the closer resemblance of the current method to standard MFCC analysis motivates us to repeat this experiment.

2. ACCUMULATOR BASED WARPING FACTOR ESTIMATION

In this section we describe our approach to VTLN warping factor estimation. The starting point is the signal analysis front-end as presented in [3], a MFCC based front-end with certain modifications to ensure that the resulting cepstrum is quefrency limited, ensuring the equivalence between frequency warping and linear transformation. These changes are briefly described below.

Instead of integrating the Mel-warping into the filter-bank as is usually done, a uniformly spaced (in Hz) filter-bank is used, and the Mel-warping is included in the warping transform. In order to still get the same amount of smoothing as for normal MFCC, the filter width is constant in Mels (the same as for MFCC). To further ensure quefrency limitedness the number of cepstral coefficients, and hence the number of filters had to be increased. In total 129 filters were used, making sure that the cepstral coefficients decay to zero. For the output side of the transform only the first 16 cepstral coefficients (the same number as in our baseline system) were used, making the warping transform a projecting transform.

2.1. Interaction with Subsequent Signal Analysis Steps

To be able to accumulate the sufficient statistics for a linear transform, we require the transform to be the last step before calculating the likelihood. In our system, the warping transform is followed by cepstral mean normalization and dynamic feature generation (derivatives or LDA), which we wish to either combine with the warp transform, or move before it.

When doing cepstral mean normalization, the mean (computed over a window) of each cepstral coefficient is subtracted from the coefficient. For the warped and normalized cepstral coefficients a simple calculation, $c_w - \overline{c}_w = Wc - W\overline{c} = W(c - \overline{c})$, shows that it is possible to change the order of cepstral mean normalization and linear transformation.

If one wishes to use an LDA based system for warping factor estimation, the following method can be used. The LDA step consists of splicing of (in our case five) consecutive acoustic frames, followed by a projecting linear transform down to a lower number of output dimensions (in our case 45). The splicing can be moved before the warp transform; the warp transform will then be block diagonal, repeatedly containing the original warping matrix. Statistics are accumulated to optimize the transform from the spliced unwarped cepstra to the warped LDA transformed ones. For each warping factor considered in the optimization a block diagonal warping matrix is combined with the previously computed LDA transform and is evaluated using the accumulated statistics.

When using a system with derivatives (regression features) it is possible to exchange the order of the warp transformation and the application of the derivatives, since a discrete derivative of any order is commutative with matrix multiplication. After the exchange, the warping transform will be block diagonal, with one repeated copy for each order of dynamic features used.

2.2. Optimization Criteria

The maximum likelihood (ML) estimation of linear transformations, as described in [4], requires computing the Jacobian of the transform. The warping transforms we are considering are projections, making the plain ML estimation method unsuitable in unmodified form. One possibility is to simply ignore the Jacobian term in the ML calculation, using only the distance. This is numerically equivalent to the standard method of warping factor estimation; it will be called the *naive* criterion. Another possibility would be to use the heteroscedastic discriminant analysis (HDA) [5] criterion, which extends the transform to be non-projecting. The application of HDA to the current problem has not been studied in this work.

Another possibility would be to use a standard discriminative criterion such as maximum mutual information (MMI), but for unconstrained transformation estimation results show that this requires interpolation with an ML estimated matrix to be useful [6]. Although this is not likely to be a problem for warping factor estimation, since only one parameter is optimized, a simpler criterion was desired. One criterion that

proved to be useful for optimizing parameters in the signal processing front-end [7] is a likelihood ratio criterion motivated in [7] as a simplification of the MMI criterion.

Starting with the MMI criterion, the competing model is exchanged with a single full covariance Gaussian model that is optimized on the same data as the transformation. Explicitly solving for the mean and covariance of the Gaussian and inserting into the equation leads to

$$g_{\text{MMI}'} = T \log \Sigma' - P(\xi_1^T | M, w_1^N), \tag{9}$$

where ξ is the features as given by the front-end, and Σ' is the full covariance matrix of ξ_1^T.

The resulting criterion is called the MMI$'$ criterion. In [7] this criterion was used in a direct optimization framework, using multiple passes over the training data to compute the objective function and its derivative. On the other hand, the close formal similarity to the standard ML criterion makes it possible to use the EM algorithm by defining an auxiliary function, in exact correspondence to the ML case. The auxiliary function is given by

$$Q(M, \hat{M}) = \sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_s(t) \log \Sigma' - P(\xi_t | s, M_s). \tag{10}$$

For the specific case of linear transform estimation the auxiliary function is given by

$$Q'(M, \hat{M}) =$$

$$\beta \log \left| W \Sigma W^T \right| - \frac{1}{2} \sum_{d=1}^{D} \left(w_d G^{(d)} w_d^T - 2 w_d k^{(d)^T} \right), \tag{11}$$

where terms constant with respect to the transform W have been omitted. Σ is the full covariance of the untransformed adaptation data, with G, k, and β defined as in section 1. Using these equations, EM optimization can be carried out in exactly the same way as for non-projecting linear transforms by iteratively optimizing $\gamma_s(t)$ using the forward backward algorithm, and W by accumulating the sufficient statistics and optimizing Q.

2.3. Implementation Considerations

Since the warping matrices are large it is important to implement the accumulation in an efficient way. Using global accumulation of G and k (equations (7) and (8)) require $O(D^2 \tilde{D})$ time per frame for accumulation (D and $\tilde{D}$ are untransformed and transformed feature dimension). With one accumulator per covariance the time complexity decreases to $O(D^2)$ at the cost of increasing memory complexity from $O(D^2 \tilde{D})$ to $O(CD^2)$ (C is the number of covariances). For the system used here this is not a problem, since only one globally pooled covariance was used. Even for a system without covariance tying the storage requirements should not be a problem, since warping factor estimation is typically done using a single Gaussian acoustic model.

Since the accumulator based approach differs significantly from standard VTLN warping factor estimation, the speed advantage is likely to be largely system dependent. In our system the accumulator based implementation uses 1/3rd of the time required for the standard warping factor estimation, being a 21 point grid search. A further advantage of the approach is the possibility to further refine the search precision without large increases in run-time.

3. EXPERIMENTAL RESULTS

All recognition experiments were performed on the TC-Star project EPPS corpus as used in the 2005 evaluation [8]. The training material includes 41 hours of manually transcribed recordings. The tests were performed on the 4 hour development set. The system was based on a MFCC front-end, and the models used consisted of roughly 200k Gaussians sharing a single globally pooled covariance. For VTLN a piecewise linear warping function was used, and warping factor estimation in training was performed as a grid search over the interval 0.8 to 1.2. In recognition the so called *fast VTLN* [9] method was used, where a Gaussian mixture model is trained for each warping factor (on the training data), and the speech segments in recognition are assigned a warping factor by the models.

To demonstrate the usefulness of our approach we compared the results of the accumulator based linear transformation implementation with the standard VTLN system. Since we had to use an increased number of filters in order to achieve quefrency limitedness, we also performed experiments on a modified standard VTLN system using the same number of filters (129) as for the linear transform case to be able to do a fair comparison. Since we wanted to demonstrate the feasibility of the linear transformation approach itself, we used the naive optimization criterion for warping factor estimation, and we used the same search grid size for all experiments.

As seen in table 1, the increase in the number of filters result in no performance gain, while the linear transformation (LT) approach outperforms the baseline.

Table 1. Recognition performance

System	MFCC	VTLN
Baseline	14.9	14.4
Extended	14.9	14.4
LT (Naive)	14.7	14.3

To investigate the effect of the optimization criterion, experiments were performed comparing MMI$'$ with the naive criterion. Since accumulator based warping factor estimation was used, a more refined search method could have been used. However, to facilitate the use of fast VTLN in recognition, grid search was still used. The grid search resolution was increased though, since this only has a small effect on the run-time in the accumulation case.

Table 2. Optimization criteria

Criterion	WER
Naive	14.3
MMI′	14.2

As can be seen in table 2 the MMI′ criterion performs slightly better. A probability of improvement of 76% for MMI′ over the naive criterion was estimated using a bootstrap estimate. Similar improvements were reported in [2], but a front-end without a filter-bank was used. The current result is the first showing an improvement over an optimized baseline of a VTLN – MFCC system when using a criterion that takes normalization into account.

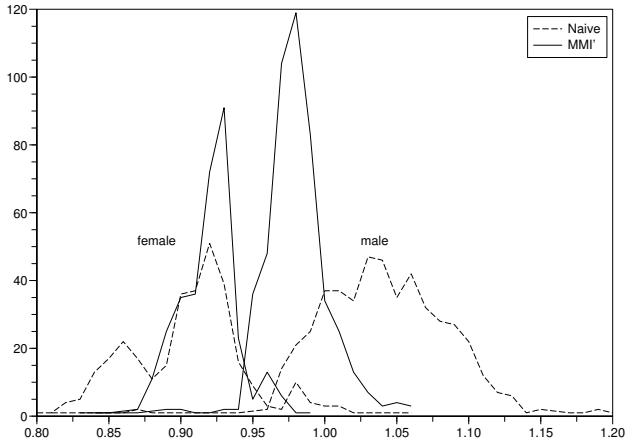

Fig. 1. Histogram of estimated warping factors

In order to further analyze the effect of properly normalized training criteria, figure 1 shows histograms of the number of speakers assigned to each warping factor in the grid search. Comparing the histograms for the naive and the MMI′ case, we observe that the histogram is more narrow in the MMI′ case. A similar effect was previously observed in [2] using a non-MFCC front-end. We also observe that the MMI′ histogram is not centered around warping factor 1.0. One possible explanation for this effect could be that the Jacobian-like term in the MMI′ criterion is sensitive to the fact that all frames, including the non-speech frames, were used for the estimation. It could also indicate that the basic Mel warping as used in our system is sub optimal and that the MMI′ criterion identifies this.

4. SUMMARY

In this paper we have presented an efficient and flexible approach to VTLN warping factor estimation. We have shown that the positive effect of using a properly normalized optimization criterion for warping factor estimation, which has been previously demonstrated for non-standard signal processing front-ends, carries over to a standard MFCC setup. Even though the positive effect is small it is still of practical interest, since it improves on an optimized baseline. Future work will be conducted in two directions. Different optimization criteria will be investigated. In particular the HDA criterion [5] will be considered, since it performs well for unconstrained transforms. Preliminary experiments have shown that it outperforms the MMI′ criterion for unconstrained linear projection estimation. The other direction is to investigate further refinements to the warping transforms, in order to allow for using smaller matrices, which would speed up the method.

5. REFERENCES

[1] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, May 1996, vol. 1, pp. 346–349.

[2] M. Pitz, *Investigations on Linear Transformations for Speaker Adaptation and Normalization*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, Mar. 2005.

[3] S. Umesh, A. Zolnay, and H. Ney, "Implementing frequency-warping and VTLN through linear transformation of conventional MFCC," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sept. 2005, vol. 1, pp. 269–272.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, Apr. 1998.

[5] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, Dec. 1998.

[6] L. F. Uebel and P. C. Woodland, "Discriminative linear transforms for speaker adaptation," in *ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia Antipolis, France, Aug. 2001, pp. 61–64.

[7] K. Visweswariah and R. Gopinath, "Adaptation of front end parameters in a speech recognizer," in *Proc. Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004, pp. 21–24.

[8] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross domain automatic transcription on the TC-Star EPPS corpus," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Mar. 2005, vol. 1, pp. 825–828.

[9] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, AZ, Apr. 1999, vol. 2, pp. 761–764.