A METHOD FOR COMPENSATION OF JACOBIAN IN SPEAKER NORMALIZATION

Rohit Sinha and S. Umesh

Department of Electrical Engineering
Indian Institute of Technology
Kanpur - 208 016, INDIA

ABSTRACT

In the conventional maximum likelihood based speaker normalization approach, the optimal frequency warping factors are estimated by maximizing the likelihood of warped features in a grid search. The conventional method of likelihood computation for warped features does not account for the Jacobian of the transformation. This fact is pointed out by some researchers who have also shown that frequency warping is equivalent to the transformation in cepstral domain. As an approximation, variance normalization of cepstral features is used before likelihood computation to account for the Jacobian. In this paper, we suggest an alternate method to avoid the Jacobian problem. Our preliminary investigation shows that our proposed method provides improvement in normalization performance compared to the conventional method of warping factor estimation for a digit recognition task.

1. INTRODUCTION

In this paper, we have made some investigations into the conventional maximum likelihood (ML) warping factor estimation based speaker normalization method for digit recognition task. In the conventional approach [1, 2], the estimation of optimal frequency warping factor required for normalization of an utterance from any speaker is done by finding the likelihoods for a set of warped features and choosing the maximum. The warped features are generated by transforming the frequency axis by a suitable warping function before obtaining the cepstra. The models used in the computation of the likelihood of warped features are trained on unwarped features. Therefore the likelihood computation of the warped features with respect to models trained on unwarped features would not be correct unless the Jacobian of the transformation is also taken into account. This fact is pointed out by some researchers [3, 4]. Further they have also shown that frequency warping can be interpreted as linear transformation in cepstral domain and the Jacobian would correspond to the determinant of the transformation matrix. As an approximation, the normalization of the variance of the warped spectra is also used before likelihood computation to obtain a similar effect as using Jacobian [5].

In this paper we suggest an alternate method to avoid the Jacobian problem in optimal warping factor estimation. In our approach instead of warping the features we warp the models so that likelihood computation for estimating the optimal warping factor does not require Jacobian of transformation. The paper is organized as follows. In Section 2, we address the problem of Jacobian of the transformation in the computation of the likelihood of the warped features for the normalization and suggest an approach to avoid the problem. The recognition experiments to evaluate the performance of proposed warping factor estimation method are described in Section 3. In Section 4, we described our recently proposed non-uniform warping based speaker normalization and evaluate its performance in light of the proposed warping factor estimation method. We finally conclude by discussing our results.

2. PROBLEM OF JACOBIAN

In this section, we first briefly review the most widely used procedure of speaker normalization [1, 2]. In this method, we warp the spectra of the i^{th} speaker by different linear or non-linear warping factors and compute feature vector as $\mathbf{x}_{i,t}^{\alpha}$ and the warped representation of the utterance of the i^{th} speaker is represented as a sequence of the warped feature vector, i.e., $X_i^{\alpha} = \{\mathbf{x}_{i,1}^{\alpha}, \ldots, \mathbf{x}_{i,T}^{\alpha}\}$. Let W_i denote the transcription of the utterance X_i from speaker i. If λ denotes a set of given HMM models trained from a large population of speakers then the optimal warping factor, $\hat{\alpha}_i$, for the i^{th} speaker is obtained by maximizing the likelihood of the warped utterances with respect to the model and the transcription, i.e.,

$$\hat{\alpha}_i = \arg\max_{\alpha} Pr(X_i^{\alpha}|\lambda, W_i) \qquad (1)$$

The normalization can be performed both in training and testing.

As pointed out in [3, 6] that if $\mathbf{x}_i$ and $\mathbf{x}_i^{\alpha}$ are original and transformed feature vectors respectively for a speaker i, then the log-likelihood of the former is *actually* given by

$$\log Pr(\mathbf{x}_i) = \log J(\alpha) + \log Pr(\mathbf{x}_i^{\alpha}; \lambda) \qquad (2)$$

where $J(\alpha)$ is the *Jacobian* of the transformation taking $\mathbf{x}_i$ to $\mathbf{x}_i^{\alpha}$.

In conventional warping based speaker normalization method, the contribution of Jacobian of transformation is not taken into account so it may cause some systematic error in the estimation of the optimal warping factors.

In the following subsection we propose an alternate approach to avoid the Jacobian problem.

2.1. Proposed Approach to Avoid Jacobian Problem

The problem of Jacobian arise since the utterance is warped before the likelihood is computed with respect to the base model. Ideally, if we had models built using speakers from each warping factor class then the likelihood of the unwarped utterance with each model can be computed without the need for computing the Jacobian of transformation since the utterance is not warped. However, in practice, we do not have enough data for each warping factor class and so we propose to counter this problem by warping all of training data by each of the warping factors and training an HMM for each warping factor.

The steps in our proposed method for speaker normalization during training and testing are described as follows:

1. We have used 7 point grid search in the range from 0.88 to 1.12 so all utterances in training data are warped by each of these warping factor and a corresponding *warped* HMM λ_α is trained.

2. For estimation of the optimal warping factor for any speaker, the log-likelihoods of the unwarped utterance with respect to different HMMs corresponding to different warping factors are computed and if the maxima corresponds to the HMM obtained by warping the training data by α_m then optimal warping factor for that speaker is given by inverse of α_m.

3. For speaker normalization during training, the optimal warping factor for each training utterance is estimated as described in step 2 and a *normalized* HMM is obtained using optimally warped training utterances.

4. For speaker normalization during testing, the optimal warping factor for a testing utterance is estimated as described in step 2 and the optimally warped utterance is then decoded using the *normalized* HMM obtained in step 3.

In [2], a similar approach was suggested in context of *fast estimation* of warping factor for test data using GMMs. The GMMs are built after warping factors have been determined for all speakers during training. Here we would like to point out that to build the GMMs, the speaker clusters belonging to a particular warping factor was found using conventional approach of warping factor estimation which does not account for the Jacobian. Hence the estimated warping factors may have some error. Further, since the mixture based method does not take advantage of the temporal information, the performance was found to be inferior to conventional HMM based warping factor estimation method.

3. RECOGNITION EXPERIMENTS

The comparison of the proposed and conventional warping factor estimation method is done on telephone based connected digit recognition task using HTK speech recognition toolkit. The acoustic training data is drawn from Numbers v1.1 corpus of Oregon Graduate Institute. The training set contained 6078 utterances totalling 33420 digits from adult male and female speakers. Two different testing sets are used to assess the performance of the features. The first is a *matched* test set called “Adults”, consisting of 2169 utterances totalling 12347 digits from adult male and female speakers. The other is a *mismatched* test set (where the age of speakers in the test set is very different from training set) called “Children”. The Children test set consists of 2798 utterances totaling 9974 words from predominantly children between age group of six to eighteen years recorded over telephone.

The digits are modeled with a word model. Continuous density left-to-right HMMs with 16 states and 5 multivariate Gaussian mixtures/state with diagonal covariance matrices are used to model the digits. The silence is modeled separately with 3 states left-to-right CDHMMs with 6 mix./state along with a short pause model of single state which is tied with the middle state of silence model. 39-dimensional feature vectors are used: normalized energy, 12 MFCCs (without c_0) computed from filterbank spanning 200-3452Hz and their first and second order derivatives. Finally cepstral features are liftered and *cluster based cepstral mean subtraction* is also performed.

In the first experiment, we studied the effect of transcription constraints on the performance of the speaker normalization during testing. Table 1 shows the performance of uniform normalization with the use of different transcription constraints in the estimation of warping factors for normalization. The normalization performance for children showed some improvement with extensive search when compared to using transcription generated by base model, while for adults both showed similar performance. In this paper we have, therefore, used the extensive search through word network during testing in our following experiments. We would also like to point out that in our proposed method, we have created the normalized model after one iteration, i.e., warping factor estimation was done only once to build the nor-

Transcription Constraint used in Warping Factor Estimation	WER (%)	
	Adults	Children
Unnormalized	3.39	14.45
Correct transcription	2.57	8.88
Transcription generated by base model	2.73	9.32
Extensive search through word network	2.75	9.07

Table 1. *Word error rate (WER) for uniform normalization using the constraint of (i) correct transcription, (ii) transcription generated by base model and (iii) extensive search through word network for both adults and children.*

Warping Factor Estimation Method used in Uniform Normalization	WER (%)	
	Adults	Children
Unnormalized	3.39	14.45
Proposed warping factor estimation method using unwarped features on warped models	2.75	7.65
Conventional warping factor estimation method using warped features on normalized model	2.75	9.07
Conventional warping factor estimation method using warped features on base model	2.92	9.30

Table 2. *Word error rate (WER) for uniform normalization for adults and children where warping factor estimated with conventional and proposed methods.*

malized models. Therefore, in the conventional method also only single iteration of warping factor estimation was carried out to build the normalized models.

Table 2 shows the normalization performance of proposed method of warping factor estimation which avoids the Jacobian problem along with the conventional method of warping factor estimation which does not account for the Jacobian of transformation. In Table 2, on comparing second and third row, we notice that proposed method has provided 15.6% reduction in WER over conventional method for children while for adults both method gave same performance.

But this comparison is not fair as in conventional method, the warping factor for test data is estimated using normalized model while the proposed method uses different warped but unnormalized models. Therefore, to study the effect of Jacobian compensation only, we have repeated the experiment by computing the warping factor estimation for test data using the base model of the conventional method and its performance is shown in fourth row of Table 2. Now comparing it with that of the proposed method we can see that proposed method reduced WER by about 6% for adults and by 17.7% for children. For fair comparison, henceforth, we will use base model for warping factor estimation for test data in conventional method.

Warping Factor Estimation Method used in Uniform Normalization	WER (%)	
	Adults	Children
Unnormalized + VN	3.28	14.51
Proposed warping factor estimation method using unwarped features on warped models + VN	2.60	7.28
Conventional warping factor estimation method using warped features on base model + VN	2.67	8.05

Table 3. *Word error rate (WER) for uniform normalization where warping factor is estimated with conventional and proposed methods and variance normalization (VN).*

In [5], it was suggested that normalizing the variance of the features can account for the Jacobian of the transformation. We have also evaluated the performance of the proposed and conventional warp factor estimation method with variance of features normalized to unity and it is shown in Table 3.

Here we find that for both approaches variance normalization provides some additional improvement. Further, using variance normalization provides additional improvement in our proposed method of Jacobian compensation.

4. NON-UNIFORM SPEAKER NORMALIZATION

All our above experiments used uniform (linear) normalization model, i.e., the spectra of any two speakers are related by $S_A(\omega) = S_B(\alpha_{AB}\omega)$. Recently we have proposed a nonlinear warping based speaker normalization [7]. In our previous work, we have numerically computed a piece-wise approximation to non-linear frequency warping function such that the spectra of any two speakers are translated versions of one anther in warped domain [8, 9]. The resulting piece-wise warping function was found to be similar to mel-scale. Therefore, in [7], we used the standard formula for the mel-scale as closed-form approximation to non-linear warping function and performed non-linear speaker normalization by shifting the warped spectra which provided some improvement over linear warping.

We have now fitted a function of form $\log(1 + f/A)$ in least squares sense to piece-wise warping function obtained for Peterson & Barney (PnB) and Hillenbrand (Hil) databases. The parameter A was estimated to be 168 for PnB and 218 for Hil. We have taken A to be 200 as a compromise because the estimates of warping function for Hillenbrand data have lower standard deviation. Thus function $\log(1 + f/200)$ is taken as the closed-form approximation to piece-wise warping function for non-linear warping. For this choice of closed-form warping function $\nu = \log(1+f/A)$, the relationship between the formant frequen-

Warping Factor Estimation Method used in Nonuniform Normalization	WER (%)	
	Adults	Children
Unnormalized + VN	3.28	14.51
Proposed warping factor estimation method using unwarped features on warped models + VN	2.61	7.14
Conventional warping factor estimation method using warped features on base model + VN	2.61	7.77

Table 4. *Word error rate (WER) for non-uniform normalization for adults and children where warping factor estimated with conventional and proposed methods and with variance normalization (VN).*

cies in the spectral domain can be found as follows: Let f_1 and f_2 represent the same formant frequency for any two speakers such that they differ in the warped domain by a translation factor, μ_{12}. In other words $\nu_2 = \nu_1 + \mu_{12}$, where $\nu_1 = \log(1 + f_1/A)$ and $\nu_2 = \log(1 + f_2/A)$. After some manipulation we can show,

$$(f_2 + A) = e^{\mu_{12}}(f_1 + A) \tag{3}$$

Thus two formant frequencies shifted by a constant factor, A, are related by constant scale factor, $e^{\mu_{12}}$, which is *speaker dependent*. This interpretation helps us implement the non-linear warping function directly in MFCC feature computation similar to efficient linear warping procedure suggested by Lee and Rose [2]. In this method, instead of scaling of spectra, the center frequencies and bandwidth of filter-bank in the computation of MFCC feature are scaled by a constant scaling factor. If C_i represents the center frequency of i^{th} channel in filter-bank and α_j represents particular scaling factor in the search range, then the modified center frequency C_i' for linear warping by α_j is given by,

$$C_i' = C_i/\alpha_j \tag{4}$$

For computation of non-linearly warped features, we followed the same procedure as discussed above except the center frequencies of filter-bank are modified using Eq.(3) as follows,

$$C_i' = \frac{(C_i + 200)}{\alpha_j} - 200 \tag{5}$$

For non-linear normalization, we have also performed 7 point grid search in the range from 0.8858 to 1.1121 so that higher cut-off frequency of the filter-bank is same in both in linear and non-linear warping cases.

Table 4 shows the performance of non-uniform warping based speaker normalization for the proposed and conventional approach of warping factor estimation. We note that non-uniform normalization provide some additional improvement for children.

5. DISCUSSION

Our preliminary experiments indicates that using warped models to account for Jacobian provides improvement. We are further investigating on doing iterative estimation of warping factors to obtain more compact models. In future work, we would compute the Jacobian for different frequency warping transformations and explicitly account for its contribution in likelihood computation to study the real impact of Jacobian of transformation.

6. REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in Vocal Tract Normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] Li Lee and Richard Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, pp. 49–59, Jan. 1998.

[3] J. McDonough, W. Byrne, and X. Luo, "Speaker Normalization with All-pass Transforms," in *Proc. of ICSLP'98*, 1998, vol. 6, pp. 2307–2310.

[4] M. Pitz, S. Molau, R. Schluter, and H. Ney, "Vocal Tract Length Normalization Equals Linear Transformation in Cepstral Space," in *Proc. of EUROSPEECH'01*, September 2001, vol. 4, pp. 2653–2656.

[5] L. Uebel and P. Woodland, "An Investigation into Vocal Tract Length Normalization," in *Proc. of EUROSPEECH'99*, September 1999.

[6] Ananth Sankar and Chin-Hui Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4(3), May 1996.

[7] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," in *Proc. of IEEE ICASSP'02*, May 2002, vol. 1, pp. 589–592.

[8] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-Warping in Speech," in *Proc. of ICSLP'96*, Philadelphia,USA, 1996.

[9] S. Umesh, S.V.Bharath Kumar, M.K.Vinay, Rajesh Sharma, and Rohit Sinha, "Simple Approach to Non-Uniform Vowel Normalization," in *Proc. of IEEE ICASSP'02*, May 2002, vol. 1, pp. 517–520.