Abstract

In this paper, we present a linear transformation (LT) to obtain warped features from unwarped features during vocal-tract length normalisation (VTLN). This linear transformation between warped and unwarped features is obtained within the conventional MFCC framework without any modification in the signal processing steps involved during the feature extraction stage. Further using the proposed LT, we study the effect of the Jacobian on the VTLN performance and show that it provides additional improvement in the recognition performance. The Jacobian of the proposed LT is simply the determinant of the LT matrix. Jacobian compensation is not done in conventional VTLN as the relation between warped and unwarped features is not known. We also study the effect of cepstral variance normalisation (CVN), which is often used as an approximation for Jacobian compensation in conventional VTLN. We show that the proposed Jacobian compensation gives better or comparable performance when compared to CVN.

Index Terms: Automatic Speech Recognition, Speaker Normalisation, VTLN, Linear Transformation, Jacobian Compensation, Cepstral Variance Normalisation.

1. Introduction

Vocal tract length normalisation (VTLN) is a commonly used procedure for speaker normalisation in speaker-independent (SI) automatic speech recognition (ASR) systems. The normalisation is implemented by warping the frequency spectra of speakers. Since, in practice, there is no information about the reference or golden speaker, a maximum likelihood based grid search is followed to find the best frequency warp factor $\alpha$, i.e.

$$\hat{\alpha} = \arg \max_{\alpha} \Pr(C_{\alpha}^i | \lambda, W_i) |dC_{\alpha}^i / dC_i|$$  \hspace{1cm} (1)

where, $C_i$ are the feature vectors of the $i^{th}$ utterance, $\lambda$, the SI or previous iteration model and $W_i$ is the true transcription during training and first pass recognition during testing for the $i^{th}$ utterance. The term $|dC_{\alpha}^i / dC_i|$ is called the Jacobian of the transformation and it accounts for the mismatch in the likelihood calculation of the warped features $C_{\alpha}^i$ with respect to the SI or previous iteration model ($\lambda$). This term is usually ignored in conventional VTLN as the transformation between warped and unwarped cepstral features is not known.

In practical ASR systems, the search range of $\alpha$ is restricted to be between 0.80 and 1.20 based on physiological constraints of the vocal tract. Warped features for conventional VTLN are obtained by scaling the filter bank proportionally for each $\alpha$ in the search range. If a linear relation can be found between $C_i$ and $C_{\alpha}^i$, then the computational complexity involved during the feature generation can be significantly reduced. We require only unwarped features and the LT matrices for each $\alpha$ to obtain the warped features. This LT approach has added advantage that the Jacobian of the transformation in Eq.1 can also be accounted for, since it is simply the determinant of the LT matrix.

VTLN may be preferable over adaptation based approaches, in applications where there is very little adaptation data. This is because VTLN requires estimation of only a single parameter as opposed to the estimation of elements of a matrix in adaptation. Further in some applications, VTLN has been shown to provide additional performance gain when used in combination with adaptation. However, the implementation of VTLN requires that features be generated for each $\alpha$ in the search range and is computationally expensive. In the proposed LT approach to VTLN, the warp-parameter estimate reduces to finding which of the pre-computed warp matrices best fits the model.

There has been lot of interest in obtaining a LT for VTLN. Pitz-Ney[1] and McDonough [2] proposed methods for analytical computation of the LT in the continuous frequency domain. These methods have alleviated problems when used in discrete domain. Further, these transformation approaches modify some aspect of the signal processing involved for conventional VTLN and also do not use DCT-II. Umesh.et.al[3] have shown using the idea of bandlimited interpolation that LT can be obtained in discrete frequency domain using plain cepstra. Using the idea of [3], Panchapagesan[4] proposed the idea of incorporating the transformation into the DCT matrix. This approach assumes the linear warping relation in the Mel-warped domain. This is not the same as linear warping in the physical frequency (Hz) domain, which is the model assumed for conventional VTLN. Very few of the above works have reported the study of Jacobian compensation using LT matrices and none have shown any improvement.

In this paper, we first propose a LT approach to VTLN that does not modify the signal processing involved for conventional VTLN. This approach is an extension of our earlier work in [3], where we have discussed the more general case of both Mel- and VTLN-warping being implemented through a linear transformation of plain cepstra. The method proposed in this paper will be a special case of bandlimited interpolation proposed in [3]. This LT approach to VTLN also allows us to compensate for the Jacobian of the transformation during the estimation of optimal frequency warp factor $\alpha$ in Eq.1.

The paper is organised as follows: In Section 2, we propose a LT approach to conventional VTLN with out any modification in the signal processing involved. In Section 3, we describe the experimental setup used in our experiments and compare the performance of the proposed LT approach with conventional
VTLN. In Section 4, we study the effect of Jacobian compensation using the determinant of the proposed LT matrix. We also study the effect of CVN, which is often used for Jacobian compensation in conventional VTLN. Finally in Section 5, we present our conclusions.

2. Linear Transformation Approach to Conventional VTLN

We first review the procedure to obtain conventional VTLN warped cepstra. The procedure is illustrated in Fig. 1. Mel and VTLN warping are integrated into the filter-bank for efficient implementation [5]. Let \( \mathbf{P} \) represent the power or magnitude spectrum of the speech frame, \( \mathbf{F}_m \) the filter bank smoothing with integrated Mel warping and \( \mathbf{D} \) represent the DCT operation. Mel frequency cepstral coefficients (MFCC), represented by \( \mathbf{C} \) (the filter-bank does not contain VTLN warping at this point and for convenience we call it \( \mathbf{C}^{1.00} \), i.e. \( \alpha = 1.00 \)) are obtained by:

\[
\mathbf{C}^{1.00} = \mathbf{D} [\log(\mathbf{F}_m, \mathbf{P})]
\]

(2)

Conventional VTLN warped cepstral features \( \mathbf{C}^{\alpha} \) are obtained by appropriately scaling the filter-bank, denoted by \( \mathbf{F}_m^{\alpha} \) and are given by:

\[
\mathbf{C}^{\alpha} = \mathbf{D} [\log(\mathbf{F}_m^{\alpha}, \mathbf{P})]
\]

(3)

From the above equations, the relation between \( \mathbf{C}^{\alpha} \) and \( \mathbf{C}^{1.00} \) is given by:

\[
\mathbf{C}^{\alpha} = \mathbf{D} [\log(\mathbf{F}_m^{\alpha}, \mathbf{P})] = \mathbf{D} [\log(\mathbf{F}_m^{\alpha}, \mathbf{P})] = \mathbf{D} [\log(\mathbf{F}_m^{\alpha} \cdot \exp(\mathbf{D}^{-1} \cdot (\mathbf{C}^{1.00})))]
\]

(4)

A LT between \( \mathbf{C}^{\alpha} \) and \( \mathbf{C}^{1.00} \) can be derived if all the intermediate operations can be represented as linear operations. But, it is evident that \( \log \) is a non-linear operation and in practice \( \mathbf{F}_m^{-1} \) does not exist and \( \mathbf{P} \) can not be completely reconstructed from the filter-bank outputs because of the smoothing operation [6]. We need to obtain \( \mathbf{P} \) since conventional VTLN warping relations are always specified in the linear frequency (Hz) domain, usually through a mathematical relation of the type \( \tilde{f} = g_x(f) \), where \( \tilde{f} \) is the warped-frequency and \( g_x(f) \) is the frequency-warping function. Therefore, in this frame work, it is not possible to recover \( \mathbf{P} \) from the filter-bank output and hence a LT is not possible.

2.1. Realising a Linear Transformation

In this section, we show that separating the VTLN-warping from the Mel filter bank helps us obtain a LT between warped and unwarped cepstral features. Let \( \mathbf{L}_m = \log(\mathbf{F}_m, \mathbf{P}) \) be the conventional log-compressed Mel filter bank outputs. Similarly, let \( \mathbf{L}^{\alpha} = \log(\mathbf{F}_m^{\alpha}, \mathbf{P}) \) be the conventional log-compressed Mel and VTLN filter bank outputs. We propose a method to obtain the LT between \( \mathbf{L}_m \) and \( \mathbf{L}^{\alpha} \), i.e.,

\[
\mathbf{L}_m = \mathbf{T}^\alpha \mathbf{L}_m
\]

(5)

From Eq. 2 and Eq. 3, the relation between \( \mathbf{C} \) and \( \mathbf{C}^{\alpha} \) is now given by:

\[
\mathbf{C}^{\alpha} = \mathbf{D} [\mathbf{T}^\alpha \mathbf{D}^{-1} \cdot \mathbf{C}^{1.00}]
\]

(6)

By defining a LT between \( \mathbf{L}_m \) and \( \mathbf{L}^{\alpha} \), we are completely avoiding the inversion of filter bank for obtaining the raw magnitude spectrum \( \mathbf{P} \) and also bypassing the \( \log \) operation. We would like to remind the reader that VTLN-warping relation is usually specified in the linear frequency-Hz domain, and therefore, at this point it is not clear what the relation between \( \mathbf{L}_m \) and \( \mathbf{L}^{\alpha} \) should be. In the next section, we describe a method to obtain a LT using the idea of bandlimited interpolation.

2.2. Proposed Approach - Bandlimited Interpolation

For a bandlimited continuous-time signal, given uniformly spaced samples of the signal, we can reconstruct the original continuous-time signal, i.e. we can recover the values of the time signal at time-instants other than those at uniformly spaced samples. We use this idea to obtain LT, except that we consider que-frenzy limited signals. The basic idea is that, with NO-VTLN warping, the conventional log-compressed Mel filter bank \( \mathbf{L}_m \) outputs are uniformly spaced samples in the Mel-domain. During VTLN warping, the outputs of the log-compressed Mel filter bank \( \mathbf{L}^{\alpha}_m \) are non-uniformly spaced in the Mel-domain. The problem is one of estimating the non-uniformly \( \mathbf{L}^{\alpha}_m = \mathbf{L}^{\alpha}_m(\nu_s) \) spaced samples given the uniformly spaced samples \( \mathbf{L}_m(\nu_s) \). This can be exactly done through bandlimited interpolation as Mel filter bank outputs are quefrenzy limited. This is assured in the current frame work through filter bank smoothing. Fig. 2 illustrates the proposed LT approach. The bandlimited interpolation matrix is given by:

\[
\mathbf{T}^{\nu_s}_{[k,n]} = \frac{1}{2N} \sum_{l=0}^{2N-1} e^{-j \frac{2\pi}{2N} l \nu_s} e^{-j \frac{2\pi}{N} \frac{l}{2N}} (\nu_s + l)
\]

(7)

where \( \nu_s \) is the sampling frequency in Mel domain. Using the even-symmetry property, we obtain NxN interpolation matrix \( \mathbf{T}^{\nu_s} \), i.e. \( \mathbf{L}^{\nu_s} = \mathbf{T}^{\nu_s} \mathbf{L}_m \). The feature generation process using the proposed LT is illustrated in Fig. 3.

The bandlimited interpolation matrix \( \mathbf{T}^{\nu_s} \) that we obtain will be of size equal to the number of MFCC filters, say \( N \). Let, \( \mathbf{M} \) represent the number of cepstral coefficients (usually 13). We then obtain the \( \mathbf{M} \times \mathbf{M} \) (13x13) matrix as follows:

\[
\mathbf{C}^{\nu_s} = \mathbf{D} \mathbf{T}^{\nu_s} \mathbf{D}^{-1} \mathbf{C}^{1.00}
\]

(8)

We believe that this is the first time an exact LT for frequency warping has been obtained without any modification in the signal processing of conventional MFCC.
3. Experimental Setup

In all the experiments, we present results on Resource Management (RM) task, Number Corpus of OGI and TI-Digits. RM task and TI-Digits are a wide-band speech having a sampling frequency of 16KHz and 20KHz respectively. OGI is narrow-band telephone based continuous digit corpus with a sampling frequency of 8KHz. For the proposed LT approach using bandlimited interpolation, we require two filters positioned at zero and \( F_s / 2 \), where \( F_s \) is the sampling frequency of the speech signal. This is because, the full symmetric spectrum (including the frequencies at zero and \( F_s / 2 \)) is required for obtaining the interpolated spectrum at warped frequencies. In all the experiments we map frequency points zero and \( \pi \) onto themselves using piecewise linear warping.

In TI-Digits and OGI, the digits are modelled as whole word simple left-to-right HMMs without skips and have 16 states per word with 5 diagonal covariance Gaussian mixtures per state. On the RM database we performed the recognition task with 49 monophones. The phone HMM models consist of 3 states with 15 diagonal covariance Gaussian mixtures per state. In both the tasks, we used a silence model having 3 states and a single state short pause model tied to the middle state of the silence model. The features in all tasks are of 39 dimensions comprising normalised log-energy, \( c_1, \ldots, c_{12} \) (excluding \( c_0 \)) and their first and second order derivatives. In all cases, cepstral mean subtraction was applied.

Table 1. shows the performance of proposed linear transformation (LT VTLN) approach along with conventional VTLN. The baseline results for conventional VTLN as well as LT VTLN are same. In all the results presented for VTLN, we follow a two-pass recognition approach. The results in Table 1. indicate that the performance of LT VTLN is comparable to conventional VTLN. In case of OGI, there is slight degradation in LT VTLN when compared to conventional approach. One possible reason could be that, OGI is narrow band telephone speech where some of the information at low and high frequencies are lost due to the bandpass nature of the telephone channel thus effecting the interpolation values.

4. Jacobian Compensation

One of our main motivations in this paper is to study the effect of Jacobian of the transformation. In conventional VTLN, the Jacobian of the transformation is usually ignored as the relation between warped and unwarped cepstral features is not known. However, as shown in Section 2.2, a LT relation can be obtained and in this case the Jacobian will be simply the determinant of the LT matrix.

4.1. Jacobian Compensation in the proposed Approach

It can be easily shown that the likelihood of the observed sequence \( (C_t) \) with respect to the model parameters \( (\mu, \Sigma) \) and the LT matrix \( (B^\alpha) \) is equivalent to applying the LT on the observation vectors and accounting for the Jacobian of the transformation [7]. This is given by:

\[
\mathbb{L}(C_t; \mu, \Sigma, (B^\alpha)^{-1}) = \mathcal{N}(B^\alpha C_t; \mu, \Sigma) + \log |(B^\alpha)| \tag{9}
\]

In Eq. 9 a single LT is applied on the model parameters, both means \( (\mu) \) and covariances \( (\Sigma) \). This is essentially constrained MLLR matrix (CMLLR) [7]. Note that we could either treat Eq. 9 as CMLLR approach to adaptation, except that in this case the CMLLR matrix is already precomputed for each \( \alpha \) or equivalently we can transform the features and account for the Jacobian. Though the LT is derived for the static features, it can be shown that the same relation holds true for \( \Delta \) and \( \Delta \Delta \) using Eq. 8. The transformation between warped and unwarp cepstral features is given by:

\[
\begin{bmatrix}
C^\alpha \\
C^\Delta \\
C^\Delta \Delta
\end{bmatrix} = \begin{bmatrix}
J^\alpha & 0 & 0 \\
0 & J^\alpha & 0 \\
0 & 0 & J^\alpha
\end{bmatrix} \begin{bmatrix}
C \\
C^\Delta \\
C^\Delta \Delta
\end{bmatrix} \tag{10}
\]

We follow the approach in Eq. 9 to linearly transform the unwarped features to obtain warped features and account for the Jacobian in all our experiments using the proposed LT.

Fig. 4 shows how the log determinant values of the proposed LT matrices \( (B^\alpha) \) vary with \( \alpha \). Table 2 shows the recognition performance of LT VTLN with and without Jacobian compensation. In conventional VTLN, the Jacobian cannot be
easily found and hence is ignored. We therefore do not have Jacobian compensation results in conventional VTLN framework. We observe that Jacobian compensation provides improvement in recognition performance. Except for the case of TI-Digits, both OGI and RM-task have shown considerable improvement in the recognition performance.

4.2. Cepstral Variance Normalisation

In literature, cepstral variance normalisation (CVN) has often been used as an approximation for Jacobian compensation. CVN normalises the variance of each element of the feature vector in an utterance to be same. This may be viewed as a crude way of normalising the data so that the Jacobian is roughly the same for each warp factor and hence ignored [8]. The CVN features are computed as follows:

\[
C_{cvn}[i] = \frac{C[i] - C_{mean}[i]}{\sqrt{C_{var}[i]}} \sqrt{C_{globalvar}[i]} \tag{11}
\]

where \(i\) indicates a particular component in the feature vector. Here \(C_{cvn}\) represents the CVN features. \(C_{mean}\) and \(C_{var}\) represent the mean and variance values calculated over the utterance for each speaker. \(C_{globalvar}\) is the global variance of the entire data, which is used to scale all the elements to the same dynamic range of original data. Note that in CVN, the normalisation parameter is estimated from the data while in Jacobian compensation, the determinant is already precomputed using the analytically determined transformation matrix.

Table 3 shows the recognition performance of CVN and Jacobian compensation using the proposed LT matrix. In the results presented here we do CVN at segment level. Note that the baseline results change with CVN when compared with the baseline results in Table 1. We observe that performing CVN on the features generated using the conventional approach or from the proposed LT approach have almost similar performance. We also present the Jacobian compensation results using the proposed LT approach. We observe that in all cases the Jacobian compensation using the LT approach has better or comparable performance to the CVN approach.

5. Conclusion

In this paper, we show that an exact LT for VTLN warping can be obtained without any modification in the existing framework using conventional MFCC features. This is done by exploiting the idea of separating the VTLN warping from the Mel filter bank. The LT is realised using the idea of bandlimited interpolation as the filter bank outputs are frequency limited. This is assured in the current frame work through filter bank smoothing. Since a LT can be obtained for VTLN warping, we are able to study the effect of Jacobian on the recognition performance in ASR. We observe that the use of Jacobian compensation during warp-factor estimation improves the normalisation performance. We also compare the effect of CVN with proposed Jacobian compensation and show that the proposed Jacobian compensation has comparable or better performance. Unlike CVN, Jacobian compensation involves no estimation of parameters from the speech data.

6. Acknowledgement

A part of this work was supported by SERC project funding SR/S3/EECE/0008/2006 from the Department of Science & Technology, Ministry of Science & Technology, India.

7. References