Study and Implementation of Analytically Determined VTLN Warping

A Project Report

submitted by

Akshay Elencheran

in partial fulfilment of the requirements for the

award of the degree of

BACHELOR OF TECHNOLOGY

under the guidance of

Prof. S Umesh

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

JUNE 2016

THESIS CERTIFICATE

This is to certify that the thesis titled **Study and Implementation of Analytically Determined VTLN Warping**, submitted by **Akshay Elencheran (EE12B005)**, to the Indian Institute of Technology, Madras, for the award of the degree of **BACHELOR OF TECHNOLOGY**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. S. Umesh
Professor
Dept. of Electrical Engineering IIT-Madras, 600 036

ACKNOWLEDGEMENT

I take this golden opportunity to express my heartfelt gratitude to Professor S.Umesh, for giving me an opportunity to work in and explore this interesting field of Digital Signal Processing. I also would like to thank him for giving me a valuable chance to meet many great and intellectual minds at Speech Processing Lab of IIT Madras.

I would also like to sincerely thank my highly esteemed supervisor, Mr. Sandeep Kothinti, for his relentless support and guidance during the course of this project. I am grateful for the motivation, inspiration and wholehearted cooperation he provided me with at all stages of this project without which this task would not have been accomplished.

TABLE OF CONTENT

Study and Implementation of Analytically Determined VTLN Warping

Akshay Elencheran, EE12B005

June 2016

Abstract

Vocal Tract Length Normalization (VTLN) for standard filterbank-based Mel Frequency Cepstral Coefficient (MFCC) features is usually implemented by warping the center frequencies of the Mel filterbank, and the warping factor is estimated using the maximum likelihood score (MLS) criterion. A linear transform (LT) equivalent for frequency warping (FW) would enable more efficient MLS estimation. In this study, a novel LT to perform FW for VTLN is analysed. The study is based on "VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC" by S.Umesh and D.R.Sanand, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 5, JULY 2012

1 Introduction

Automatic speech recognition systems must be able to cope with considerable variation among speakers, major sources of this inter-speaker acoustic variation are physiological factors such as gender and vocal tract length. Vocal tract length normalization is one of the easiest ways of doing fast speaker adaptation. The underlying idea is simple.

Resonances in an acoustic tube (such as the vocal tract) are inversely proportional to the length of the tube. Thus, as female vocal tracts are 10-15% shorter than male vocal tracts, resulting female formant positions are higher than the equivalent male formant positions. Given the all in all moderate variations in vocal tract length, the effect of vocal tract length variation can be modeled well by a linear warping of the frequency axis. VTLN tries to normalize the position of the formant peaks by warping the spectrum to represent an average vocal tract. Hence by warping a spectrum by a speaker specific warping factor, typically towards a global average vocal tract length, we obtain a 'normalized' (wrt. vocal tract length) spectral estimate. By normalizing out this physiological influence the obtained spectral estimates are more homogeneous across speakers and hence more suitable for recognizing the acoustic phonetic content. The linear warping itself can be incorporated into the filterbank that is used to convert from linear frequency to mel frequency.

The main advantage of feature normalization is that the number of parameters to be estimated from the adaptation data is generally smaller compared with the standard model based adaptation techniques. Hence, adaptation can be carried out with very little adaptation data.

2 Conventional VTLN

2.1 Computation of Mel-Frequency Cepstral Coefficients

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. Steps involved in determination of MFCCs:

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

Mel Frequency Cepstral Coefficients (MFCCs) are a very popular choice of features used for automatic speech recognition. Standard MFCCs are computed as shown in Fig.1, and the Mel filterbank is shown in Fig.2. The filters are assumed to be triangular and half overlapping, with center frequencies spaced equally apart on the Mel frequency scale. The Mel scale was derived from experiments on pitch perception (frequencies which are spaced equally apart according to pitch) and is calculated from the regular frequency scale using the formula,

$$mel(f) = 1125 * log(1 + \frac{f}{700}) \tag{1}$$

During MFCC feature extraction, the speech signal is pre-emphasized and divided into frames and each frame is first windowed using the Hamming window. The short-time power spectrum vector P is obtained from the squared magnitude of the FFT of the windowed frame. The log of the filterbank outputs is obtained as:

$$L = log(F.P) \tag{2}$$

where F is the Mel filterbank matrix. Here, we use the notation that the log of a vector is the log applied to each component. The MFCCs are then given by

$$c = D.L \tag{3}$$

$$c = D * log(F.P) \tag{4}$$

where D is a type-II DCT matrix. The final feature vector x used for recognition, typically consists of the MFCCs and their first and second time derivatives, often called the deltas and delta-deltas:

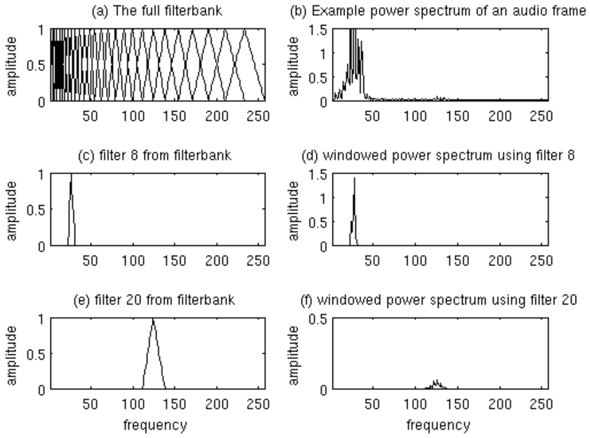

Figure 1: Plot of Mel Filterbank and windowed power spectrum.

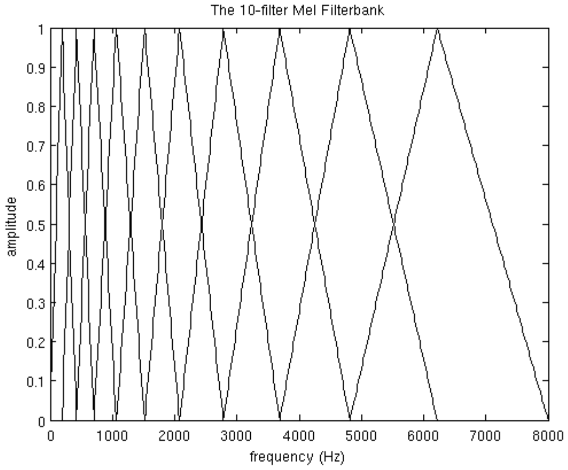

Figure 2: A Mel-Filterbank containing 10 filters. This Filterbank starts at 0Hz and ends at 8000Hz.

$$\begin{vmatrix} c \\ \Delta c \\ \Delta^2 c \end{vmatrix}$$

The delta cepstra are computed using the following formula

$$\Delta c_t = \frac{\sum_{k=1}^{K} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \tag{5}$$

This approximation of the time derivative is obtained by fitting a second order polynomial to a sequence of $2K + 1$ cepstral coefficients. $\Delta 2c$ is similarly calculated from Δc.

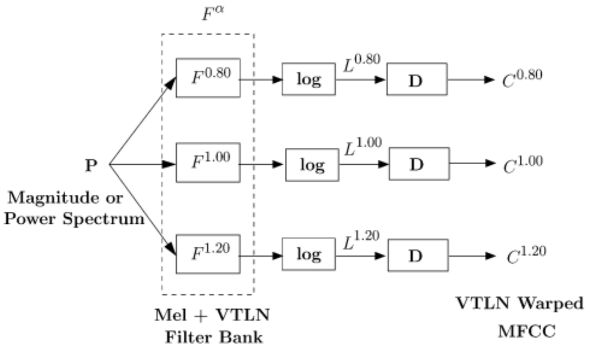

Figure 3: Conventional frame work for generating warped features in VTLN.

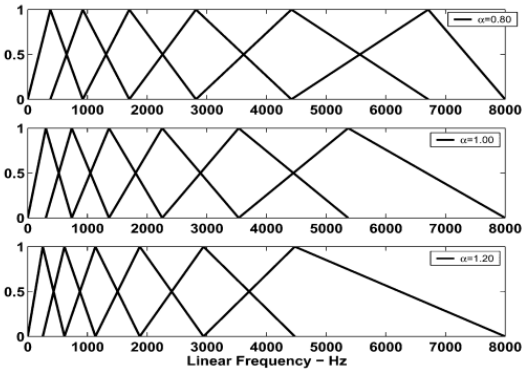

Figure 4: Illustrating the change in the filter-bank structure with VTLN-warping in linear-frequency (Hz) domain.

2.2 Computation of Conventional VTLN Features

VTLN features are originally obtained in the original method of Andreou et al. [1], by frequency-warping the magnitude spectra P to get P^{α} before applying the unwarped Mel filter-bank. This is done by re-sampling the signal. Therefore, in this case the signal is warped for each VTLN warp-factor, while the Mel filter-bank is left unchanged.

Lee and Rose [2] proposed an efficient alternate implementation, where the Mel filter-bank is inverse-scaled for each , while the signal spectra is left unchanged as shown in Fig.3. This is the most popular method of VTLN-warping. Therefore, in the Lee-Rose method, VTLN warping is integrated into the Mel filter-bank and F^{α}denotes the (inverse) VTLN-warped Mel filter-bank. Conventionally the warp-factor, α, used for warping the spectra is in the range of 0.80 to 1.20 based on physiological arguments. For each α, the center frequencies and bandwidths of the Mel filter-bank are appropriately scaled to obtain Mel- and VTLN-warped smoothed spectra [2]. The change in the filter-bank structure for different warp-factors is illustrated in Fig.4. The slope in the last filter has been modified appropriately using piece-wise linear warping, so that the Nyquist

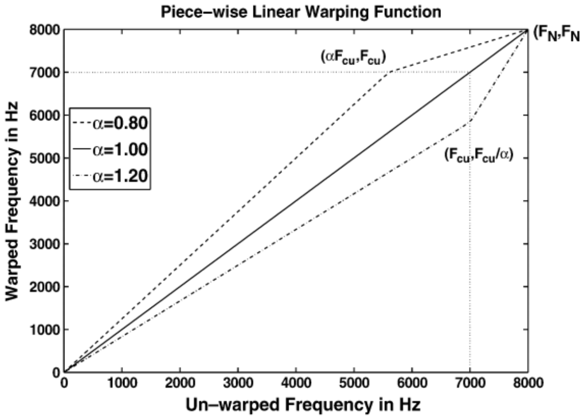

Figure 5: Piece-wise linear warping function used in conventional VTLN.

frequency maps onto itself after frequency scaling. This avoids the bandwidth mismatch that arises due to frequency warping. The piece-wise linear warping function used in our experiments is given by

$$f^{\alpha<1} = \begin{cases} \alpha f, & \text{if } f \leq F_{cu} \\ \alpha F_{cu} + \frac{F_N - F_{cu}\alpha}{F_N - F_{cu}}(f - F_{cu}), & \text{otherwise} \end{cases}$$

$$f^{\alpha>1} = \begin{cases} \alpha f, & \text{if } f \leq F_{cu}/\alpha \\ F_{cu} + \frac{F_N - F_{cu}}{F_N - F_{cu}/\alpha}(f - F_{cu}/\alpha), & \text{otherwise} \end{cases}$$

and is shown in Fig.5. F_{cu} represents the cutoff frequency where the slope is changed and F_N is the Nyquist frequency. The warped cepstral features C^α are given by

$$C^\alpha = D[log(F^\alpha.P)] \tag{6}$$

These are obtained by first warping and smoothing the power spectrum, followed by log and the DCT operations. The filterbank is integrated with both Mel- and VTLN-warping, to perform smoothing as well as scaling of the spectrum. For the case of $\alpha = 1.00$, C^α exactly corresponds to the case of conventional MFCC without VTLN warping. From (4) and (6), the relation between C^α and C is given as

$$C^\alpha = D[logF^\alpha(F^{-1}.exp(D^{-1}.C^{1.00}))] \tag{7}$$

A linear-transformation between C^α and $C^{1.00}$ can be derived if all the intermediate operations can be represented as linear operations, but from (7), it is evident that log is a nonlinear operation and in practice F^{-1} does not exist. This is because, the power-spectrum P cannot be completely reconstructed from the filter-bank outputs because of the smoothing operation.

We need to obtain P, since conventional VTLN warping relations are always specified in the linear-frequency (Hz) domain, usually through a mathematical relation of the type $f^\alpha = g_\alpha(f)$, where f^α is the warped-frequency and $g_\alpha(f)$ is the frequency warping function. Therefore, in this case, it is not possible to completely recover P from the filter-bank output and hence a linear-transformation is not possible.

2.3 Examples of Normalized Frequency Warping Functions

Piecewise Linear: These are the type of Frequency Warping functions that are commonly used in VTLN.

$$\theta_p(\lambda) = \begin{cases} p\lambda, & 0 \leq \lambda \leq \lambda_0 \\ p\lambda_0 + \frac{1-p\lambda_0}{1-\lambda_0}(\lambda - \lambda_0), & \lambda_0 < \lambda \leq 1 \end{cases}$$

Linear: This FW can be used for adaptation from adult models to children's models, where the original models have more spectral information than necessary for children's speech. For $p \leq 1$,

$$\theta_p(\lambda) = p\lambda, 0 \leq \lambda \leq 1 \tag{8}$$

Sine-Log Allpass Transforms (SLAPT): SLAPT frequency warping functions are capable of approximating any 1-1 arbitrary frequency warping function, and are therefore suitable for multi-class adaptation or the adaptation of individual distributions. The K-parameter SLAPT, denoted SLAPT-K, is given by:

$$\theta_p(\lambda) = \lambda + \sum_{k=1}^{K} p_k sin(\pi k\lambda) \tag{9}$$

3 Previous works on VTLN

Recently there have been many types of linear transformation approaches proposed for VTLN. Pitz and Ney[3] have proposed a method for analytical computation of the linear transformation in the continuous frequency domain. In continuation Umesh.et.al[4] have showed that assuming que-frency limitedness and using the idea of band-limited interpolation to implement the method of Pitz and Ney in discrete-domain. They exploit the idea of separating the smoothing and warping operations, which requires modification in the signal processing for extraction of cepstral features. Though the method was basically aimed at the discrete implementation of the method proposed by Pitz and Ney, the method could be easily modified to work in the conventional frame work with out any modification in the signal processing.

For standard MFCC features, because of the non-invertible filterbank with non-uniform filter widths, even with the assumption of quefrency limitedness, the MFCC features after warping cannot even be expressed as a function (linear or non-linear) of the unwarped MFCC features. For a given warping of the linear frequency signal spectrum, there is not a single function (for all possible cepstra) that will give the warped cepstra from the unwarped cepstra. An exact inverse transformation of the mel-FB smoothed spectrum is not possible. We cannot construct the original power spectrum from the smoothed spectrum. Another problem is the presence of a logarithm (non-linear operation), which can not be represented as a matrix operation. Hence, approximate linear transforms have been developed for frequency warping with MFCC features (Claes et al., 1998; Cui and Alwan, 2006; Umesh et al., 2005).

Panchapagesan[5] proposed the idea of incorporating the transformation into the DCT matrix. Here he assumes the linear-warping relation, defined in the frequency Hz domain between the speakers to be true in the mel-frequency domain, which needs clarification and is not valid. Sanand.et.al[6] have proposed the idea of using dynamic frequency warping for obtaining a linear transformation. They also exploited the idea of separating smoothing and warping operation, hence modified the signal processing in the feature extraction.

While the methods in [4] and [5] do not require knowledge of the closed form equation for frequency warping function, they do require knowledge of the exact mapping between unwarped and warped discrete frequencies. This means the correspondences between warped and unwarped frequencies at certain discrete points are assumed to be known, but not the exact functional relation between them.

In the next section, the separation of the frequency warping operation from the filter-bank avoiding the need to invert the filter-bank operation or the logarithm is discussed which allows us to derive a linear transformation on conventional MFCC.

4 Analytically Determined VTLN Warping

4.1 Band Limited Sinc Interpolation

Bandlimited interpolation of discrete-time signals is a basic tool having extensive application in digital signal processing. In general, the problem is to correctly compute signal values at arbitrary continuous times from a set of discrete-time samples of the signal amplitude. In other words, we must be able to interpolate the signal between samples. Since the original signal is always assumed to be bandlimited to half the sampling rate, (otherwise aliasing distortion would occur upon sampling), Shannon's sampling theorem tells us the signal can be exactly and uniquely reconstructed for all time from its samples by bandlimited interpolation.

We review briefly the "analog interpretation" of sampling rate conversion on which the present method is based. Suppose we have samples $x(nT_s)$ of a continuous absolutely integrable signal $x(t)$, where t is time in seconds (real), n ranges over the integers, and T_S is the sampling period. We assume $x(t)$ is bandlimited to $\pm F_S/2$, where $F_S = 1/T_S$ is the sampling rate. If $X(\omega)$ denotes the Fourier transform of $x(t)$, i.e.,

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \tag{10}$$

then we assume $X(\omega) = 0$ for $|\omega| \geq \pi F_S$. Consequently, Shannon's sampling theorem gives us that x(t) can be uniquely reconstructed from the samples $x(nT_S)$ via

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_s)h_s(t - nT_s) \tag{11}$$

where $h_s(t) = sinc(F_s t)$

To resample x(t) at a new sampling rate $F_S = 1/T_S$, we need only evaluate Eq.11 at integer multiples of T_S.

4.2 A Mel-Filterbank Approach

In the Mel-frequency domain, the continuous Mel-warped logcompressed spectrum, $L(v)$, can be interpreted as the convolved output of a triangle function on the Mel-warped magnitude spectrum and followed by a log operation on the amplitudes. We can think of vector as being obtained by uniformly sampling $L(v)$ at $v_i = 2\pi i/N$ where i = 0,1,....,(N-1) and the positions of these samples exactly correspond to the center frequencies of the filter-bank. Because of the triangle smoothing and subsequent log-operation on the output (which reduces dynamic range), the que-frency content of this -compressed smoothed spectrum is only in the low que-frency region.

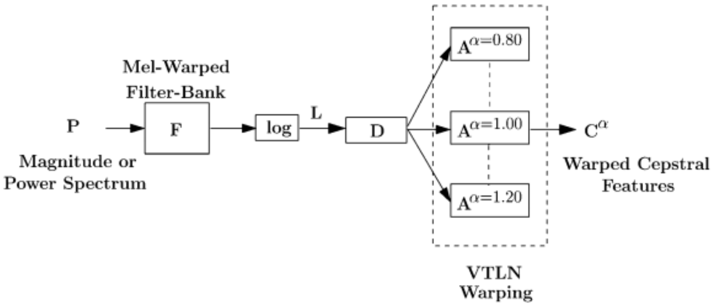

Figure 6: Framework of Analytically Determined VTLN.

During VTLN-warping, the filter center frequencies are appropriately scaled in the linear-frequency (Hz) domain by $inverse - \alpha$. This corresponds to the center frequencies of the filter-bank to be non-uniformly spaced in the Mel-frequency domain. As we represent the log-compressed Mel-warped smoothed magnitude spectrum by the continuous function $L(v)$, the output of the VTLN-warped filterbank corresponds to sampling $L(v)$ nonuniformly, $L[v_i^\alpha]$. These nonuniformly spaced samples exactly correspond to the elements of the vector L^α.

The elements of vector L (i.e., $L[v_i]$) can be interpreted as uniformly spaced samples and elements of L^α(i.e., $L[v_i^\alpha]$) as nonuniformly spaced samples of the same continuous function $L(v)$. The main idea is that, given the samples in L, the samples (or elements) in L^α can be reconstructed using band-limited interpolation provided that the cepstrum is que-frency limited.

Let $L(v)$ and H form a discrete-time Fourier transform (DTFT) pair. Then sampling $L(v)$ would result in periodic repetition of H. As long as H is strictly que-frency limited and the sampling rate is sufficiently high, then there is no aliasing in the cepstral domain. In such a case, the value of $L(v)$ at any Mel-frequency v_i^α can be found from its uniformly-spaced samples at v_i through band-limited interpolation. This is basically exploiting the sampling theorem, where a signal (in this case a frequency domain signal) can be reconstructed from its samples using Sinc-interpolation. H is nowhere used for any calculation purposes and is presented here only for better understanding in the derivation of the band-limited interpolation matrix.

The que-frency limitedness ensures that there is no overlap in the periodic repetition of H (i.e., no aliasing), and hence L^{α} can be exactly recovered. The que-frency limitedness property depends both on the amount of smoothing done by the Mel-filters (which controls the number of significant cepstral coefficients) and on the number of Mel-filters which determines the periodicity. If there is aliasing, there will be differences between Sinc-interpolated L^{α} and the actual values.

4.3 Calculation of Linear Transformation Matrix

Let $v_0, v_1, v_2, v_3 ... v_{N-1}$, represent the uniformly-spaced Mel frequencies with samples of $L(v)$ at these points being elements of vector L. Their corresponding linear-frequencies (Hz) are nonuniformly spaced and are represented by $f_0, f_1, f_2, f_3 ... f_{N-1}$. These are the center frequencies of the Mel-filters in the linear-frequency (Hz) domain and are related through the standard Mel-relation, i.e.,

$$v_i = 1125 log_e (1 + \frac{f_i}{700}) \tag{12}$$

During VTLN-warping, the warping function $g_{\alpha}(f)$ is applied to obtain the warped frequencies. Let, $f_i^{\alpha} = g_{\alpha}(f)$ represent the warped frequencies in the linear-frequency (Hz) domain. The corresponding VTLN-warped center frequencies of the filters in the Mel-frequency domain (v_i^{α}) will not be related through a linear scaling relation, since

$$v_i^{\alpha} = 1125 log_e (1 + \frac{g_{\alpha}(f_i)}{700}) \tag{13}$$

Therefore, while $f_i^{\alpha} = g_{\alpha}(f) = \alpha f_i^{\alpha}$ for the linear scaling relation (i.e., along x axis), along y-axis $v_i^{\alpha} \neq \alpha(v_i)$ as seen from (12) and (13).

The Fourier relation between H and L is given by

$$h_k = \frac{1}{2N-2} \sum_{i=0}^{2N-3} L[v_i] e^{j2\pi \frac{v_i}{2v_s} k} \tag{14}$$

where v_s is the Nyquist frequency in the Mel frequency domain. Here, we assume that the signal is periodic with a period of $2N-2$and symmetric around $N-1$. Therefore, theoretically half-filters are present at indices 0 and $N-1$. The values at these indices are required for performing band-limited interpolation. If, we assume that H is que-frency limited, the elements of L^{α} can be determined as

$$L[v_j^{\alpha}] = \sum_{k=0}^{2N-3} h_k e^{-j2\pi \frac{v_j^{\alpha}}{2v_s} k} j = 0, 1, ...2N-3 \tag{15}$$

Substituting h_k of (14) in (15), we get

$$L[v_j^{\alpha}] = \sum_{i=0}^{2N-3} T_{ji}^{\alpha} L[v_i] \tag{16}$$

where T^{α}_{ji} is,

$$T^{\alpha}_{ji} = \frac{1}{2N-2} \sum_{k=0}^{2N-3} e^{-j2\pi \frac{v^{\alpha}_{j}}{2v_s} k} e^{j2\pi \frac{v_i}{2v_s} k} \tag{17}$$

5 Implementation in KALDI

The Simulation accomplished in MATLAB was taken to a further level in the form of a C code, which could be used in KALDI to automate the process of VTLN and check results.

The main objective of the code was to derive the Linear transformation matrixes for the various values of alpha and store it in a file accessible to the routine code of KALDI.

The inputs that can be given in as arguments are

1. Sampling frequency of the speech signal
2. Low frequency: Lower bound for frequency
3. High frequency: Upper bound for frequency
4. Warp inflection frequency: Cutoff frequency used in the calculation of warping function
5. Num-bins: number of filters in the Filter-Bank
6. Num-ceps: number of cepstral coefficients taken after DCT
7. Start-alpha: Starting value of range of warping coefficients (α)
8. End-alpha: Ending value of range of warping coefficients (α)
9. Alpha-dist: distance/difference between two consecutive α
10. Cepliftering parameter
11. Delta: whether to include delta & delta-delta information in the matrix
12. Linearity: whether to store the matrix in an order linear to alpha or inverse to alpha

The Value of logarithm of the determinant of the matrix was also required in the specified format of KALDI. Since the matrices were large but almost sparse, LU decomposition method was used to calculate the *logdet* value for each warping factor.

The matrices were computed depending on the values input in the command line arguments and were stored in a format compatible with KALDI.

For the experiments, matrices of size 13 as well as 40 bins were tried. Cases of the matrices including and excluding delta and delta-delta features were considered. Hence matrices of size 13x13, 39x39, 40x40 and 120x120 were used for Analytically Determined VTLN. Surprisingly we got the best results for 13x13 case rather than 40x40 and hence 13x13 results are used for inference in the next section.

6 Results

The resemblance between the plot of coefficients generated by the conventional method and the ones generated by Analytically determined VTLN warping method signifies accuracy and efficiency of the new method.

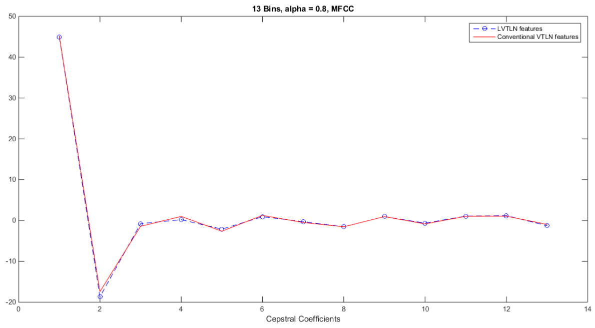

Figure 7: MFCC: Conventional vs Analytically Determined.

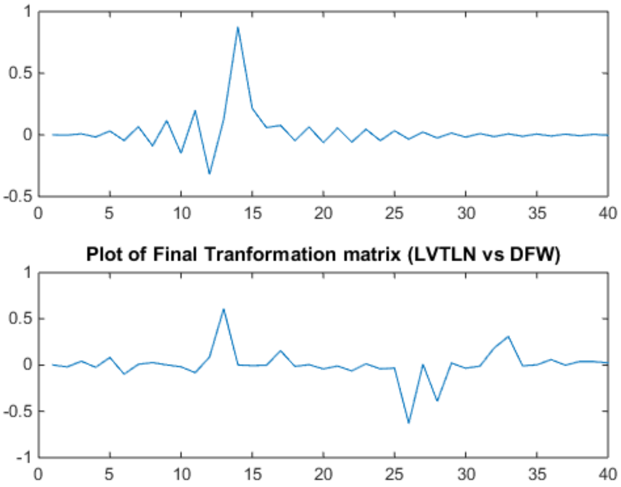

Figure 8: Transformation Matrix: Conventional vs Analytically Determined.

The use of such matrices also enables the warp-factors to be estimated by accumulating the sufficient statistics, there by simplifying the procedure for optimal warp-factor estimation and reducing the computational complexity by 75%. Further, VTLN matrices can be used in regression tree framework to perform VTLN at acoustic class level, allowing estimation of multiple warp-factors for a single utterance which is very difficult to implement in conventional VTLN framework.

Table 1: 13 Bins Dev and Test Results

	DEV	TEST
MONO	31.6	32.4
TRI1	24	26.3
TRI1 KALDI VTLN	23.3	25.1
TRI1 LVTLN	23.6	25.5
TRI2	22.7	24.1
TRI2 KALDI VTLN	21.8	23.3
TRI2 LVTLN	22.5	23.9

The results obtained from the experiments have not been up-to par with the expectations. Upon further investigation, it was found that the inbuilt warping factor (α) estimator in KALDI is not compatible with the way the Linear Transformation Matrices are generated.

The current VTLN model implemented in KALDI determines the Linear Transformation Matrices by generating a set of warped features for various values of alpha and a different set of unwarped features. Using these generated sets, KALDI tries to estimate the possible values of the Linear Transformation Matrix for each value of alpha, which then is later on used for the process of VTLN. During this process many key information and values are calculated and stored which are then used for the estimating the appropriate value of α for a particular segment of speech belonging to a user. Once the warping factor is estimated, it is tagged as the value of α for the particular user.

Since for the process of α estimation, KALDI uses the intermediate information which is not calculated in the New method of VTLN, it is not compatible with the Analytically Determined VTLN warping and is hence results are not upto the mark.

But according to the Fig.7 and Fig.8, the filters generated in the new method should work better as compared to the ones generated by KALDI. Hence a proper α estimator for the new VTLN warping techniques will defintely improve the performance and give better results.

7 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] L. Lee and R. Rose, “Frequency warping approach to speaker normalization,” IEEE Trans. Speech Audio Process., vol. 6, no. 1, pp. 49–59, Jan. 1998.

[3] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 930–944, Sep. 2005.

[4] S. Umesh, A. Zolnay and H. Ney, “Implementing frequency warping and VTLN through linear transformation of conventional MFCC”, Interspeech2005, pp. 269-272.

[5] S. Panchapagesan., "Frequency warping by Linear Transformation of Standard MFCC", Interspeech2006.

[6] D. R. Sanand, D. Dinesh Kumar and S. Umesh, "Linear Transformation Approach to VTLN Using Dynamic Frequency Warping", Interspeech2007.