Revisiting VTLN Using Linear Transformation on Conventional MFCC

D. R. Sanand, R. Schlüter and H. Ney

Lehrstuhl für Informatik 6, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany

`[doddipatla,schlueter,ney]@cs.rwth-aachen.de`

Abstract

In this paper, we revisit the linear transformation for VTLN on conventional MFCC proposed by Sanand *et al.* in [1], using the idea of band-limited interpolation. The filter-bank is modified to include half-filters at *zero* and *nyquist* frequencies, as the full symmetric spectrum is required for performing band-limited interpolation. In this paper, we show that the filter-bank with half-filters does not affect the recognition performance on clean speech (also shown in [1]), but does affect the recognition performance on noisy speech. This motivated us to revisit the linear transformation for VTLN in [1] and propose modifications to undo the affect of half-filters during the feature extraction. We show through recognition experiments that the proposed modifications to the linear transformation have comparable performance as the conventional VTLN approach, still enabling us to perform VTLN using a linear transformation on conventional MFCC.

Index Terms: MFCC, VTLN, Linear Transformation, Speaker Normalization, Automatic Speech Recognition.

1. Introduction

Vocal tract length normalization (VTLN) is an established procedure in the area of automatic speech recognition (ASR) for performing speaker normalization. It normalizes the variations in the spectra of speech signals that arise due to the differences in the vocal tract lengths (VTL) of the speakers uttering the same sound [2]. The normalization is achieved by scaling the spectra, i.e. either compressing or expanding. Usually, the vocal tract is assumed to be a uniform tube and the spectra of different speakers uttering the same sound are assumed to be linearly scaled versions of one another [2, 3].

The conventional Mel frequency cepstral coefficient (MFCC) feature extraction is illustrated in Fig. 1, and can be written as:

$$\mathbf{C} = \mathbf{D}[\log{(\mathbf{F_m}.\mathbf{P})}] \tag{1}$$

where, $\mathbf{P}$ represents the power or magnitude spectrum, $\mathbf{F_m}$ the filter-bank smoothing, $\log$ the logarithm operation on the amplitudes, $\mathbf{D}$ the discrete cosine transformation (DCT) and $\mathbf{C}$ are the cepstral coefficients. For efficient implementation, VTLN-warping is embedded into the Mel filter-bank [3]. The VTLN-warped MFCC features ($\mathbf{C}^\alpha$) are given by:

$$\mathbf{C}^\alpha = \mathbf{D}[\log(\mathbf{F_m^\alpha}.\mathbf{P})] \tag{2}$$

where, $\mathbf{F_m^\alpha}$ represents the piece-wise linearly scaled Mel-warped filter-bank and α (also called warping-factor) represents the amount of piece-wise linear scaling of the spectra.

In practice there is no reference speaker with respect to which the optimal warping-factor is estimated and usually a
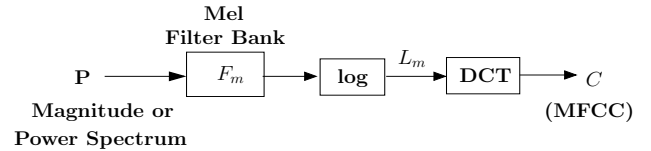

Figure 1: Conventional MFCC features extraction.

maximum likelihood (ML) based grid search is performed over a predefined range of warping-factors [2, 3]. It is given by:

$$\hat{\alpha}_{\mathbf{ML}} = \arg\max_\alpha \mathbf{p}\{\mathbf{X}^\alpha|\lambda; \mathbf{W}\} \tag{3}$$

where, $\mathbf{X}^\alpha$ are the features that are either transformed using linear discriminant analysis (LDA) or appended with differential and acceleration coefficients of $\mathbf{C}^\alpha$, λ is the acoustic model and $\mathbf{W}$ is the transcription. The range of α is usually chosen to be in between 0.80 and 1.20 with increments of 0.02.

Observing Eq. 1 and Eq. 2, it is clear that the only difference in operation is the change in the filter-bank structure between warped and un-warped cepstral features. From Eq. 3, it is evident that all the warped features in the range of α need to be generated before the optimal estimate can be determined. This implies that the structure of the filter-bank needs to be changed for each α and all the signal processing steps following the filter-bank need to be repeated during the feature extraction. This has motivated many researchers to come up with procedures to avoid repeating all the signal processing steps and still be able to generate the VTLN-warped MFCC features.

One such approach that gained interest in the recent years, is to derive a linear transformation (LT) on the conventional MFCC to obtain the VTLN-warped MFCC features [1, 4, 5, 6], i.e.

$$\hat{\mathbf{C}}^\alpha = \mathbf{A}^\alpha \mathbf{C} \tag{4}$$

where, $\mathbf{A}^\alpha$ is the LT matrix for a specific warping-factor α. By doing so, we not only eliminate the need to repeat all the signal processing steps following the filter-bank, but also can generate the warped-features on-the-fly and eliminate the need to store them for obtaining the optimal estimate of α [1].

In this paper, we focus on the linear transformation approach to VTLN proposed by Sanand *et al.* in [1] using the idea of band-limited interpolation. This approach was motivated from the work of Umesh *et al.* [4], where they modified the linear transformation proposed by Pitz *et al.* [5] in the continuous domain and showed that VTLN-warped MFCC features can be obtained using a linear transformation on plain cepstra in the discrete domain. Panchapagesan *et al* [6] also motivated from the work of Umesh *et al.* [4], proposed a linear transformation by incorporating VTLN-warping into the inverse DCT
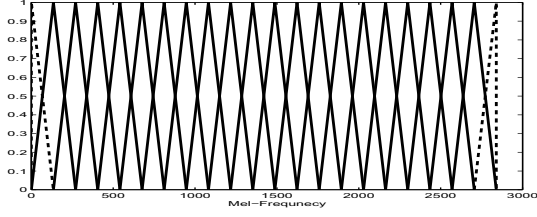
26 – 30 September 2010, Makuhari, Chiba, Japan

Figure 2: Illustrating the change in the filter-bank with the inclusion of half-filters (shown in dotted line).

(IDCT) transformation. By doing so, VTLN-warping is performed on the Mel-warped frequencies and does not correspond to conventional VTLN-warping.

In [1], the VTLN-warped cepstra are obtained by performing band-limited interpolation on the Mel-warped log-smoothed spectra, say $\mathbf{L_m}$(=log($\mathbf{F_m}.\mathbf{P}$)), and followed by a DCT operation, i.e.

$$\hat{\mathbf{C}}^\alpha = \mathbf{D}.\mathbf{L_m^\alpha}, \quad \text{where} \quad \mathbf{L_m^\alpha} = \mathbf{T}^\alpha.\mathbf{L_m}. \quad (5)$$

Here, $\mathbf{T}^\alpha$ represents the interpolation matrix for a specific α and $\mathbf{L_m^\alpha}$ is the VTLN-warped $\mathbf{L_m}$. Assuming that $\mathbf{L_m}$ is obtained by uniformly sampling the continuous smoothed spectrum, say L_m, then $\mathbf{L_m^\alpha}$ can be obtained by non-uniformly sampling L_m. If L_m is band-limited, we can perfectly reconstruct $\mathbf{L_m^\alpha}$ given $\mathbf{L_m}$ using band-limited interpolation. In order to do perfect reconstruction, we require the information of the full symmetric spectrum and hence the filter-bank is modified to include half-filters at *zero* and *nyquist* frequencies as shown in Fig. 2.

In this paper, we show that the presence of these half-filters does *not* affect the recognition performance on clean speech (also shown in [1]), but does affect the performance on noisy speech. Since the presence of half-filters is only a requirement in the interpolation of the spectrum and not a requirement in the feature extraction process, we propose modifications to the linear transformation and show that VTLN can still be performed using a linear transformation on conventional MFCC.

The paper is organized as follows: first, we present a brief review of the linear transformation approach proposed in [1], using the idea of band-limited interpolation. We then propose modifications to the linear transformation to undo the affect of half-filters on the warped cepstra. Later, we present the recognition results on European Parliament Plenary Sessions (EPPS) English and Aurora 4.0 tasks and show that the modified linear transformations have comparable performance with the conventional VTLN approach. Finally, we present our conclusions.

2. Linear Transformation for VTLN

In [1], it is argued that a linear transformation is not possible in the conventional VTLN frame work due to the presence of log and the need to invert the filter-bank smoothing operation. It is also shown that a linear transformation can be realized by separating the VTLN-warping from the Mel filter-bank. The VTLN-warped cepstra are obtained as shown in Eq. 5. Since $\mathbf{L_m} = \mathbf{D}^{-1}\mathbf{C}$ (from Eq. 1), the linear transformation between the warped and un-warped cepstra is given by:

$$\hat{\mathbf{C}}_{hf}^\alpha = \mathbf{A}^\alpha.\mathbf{C}_{hf} \quad \text{where} \quad \mathbf{A}^\alpha = [\mathbf{D}.\mathbf{T}^\alpha.\mathbf{D}^{-1}]. \quad (6)$$

The subscript *hf*, explicitly indicates the presence of half-filters in the filter-bank and are used in the feature generation.

Exploiting the even symmetric property, the NxN interpolation matrix is given by:

$$\mathbf{T}^\alpha{}_{NXN} = \frac{2}{N-1}\left[\mathbf{V}_{NXN}.\mathbf{W}_{NXN}\right] \quad (7)$$

where, N is the number of filters. The matrices $\mathbf{V}$ and $\mathbf{W}$ are given by:

$$\mathbf{V} = [cos(2\pi\hat{l}k)][w_{kk}] \quad \text{and} \quad \mathbf{W} = [cos(2\pi lk)][w_{kk}]$$

where, $0 \le \hat{l}, l \le 0.5$ and $0 \le k \le N-1$. w_{kk} is a diagonal matrix and is given by:

$$w_{kk} = \begin{cases} \frac{1}{2}, & k = 0, N-1 \\ 1, & k = 1, 2, \ldots, N-2. \end{cases}$$

l and $\hat{l}$ are the normalized un-warped and warped center frequencies of the filters in the Mel-frequency domain respectively. Excluding the half-filters, the rest of the filter-bank center frequencies exactly correspond to the original filter-bank. It is *important* to note that VTLN-warping is not valid on the Mel-warped frequencies and the corresponding warped frequencies ($\hat{\nu}$ - not normalized) are obtained using the Mel-warping relation:

$$\hat{\nu} = 2595 \ \log_{10}\left(1 + \frac{\hat{f}^\alpha}{700}\right) \quad \text{and} \quad \hat{l} = \frac{\hat{\nu}}{(2\nu_n)} \quad (8)$$

where, $\hat{f}^\alpha$ corresponds to the warped center-frequency for a particular filter in the linear-frequency (Hz) domain and ν_n is the *nyquist* frequency in Mel's.

The band-limited interpolation matrix ($\mathbf{T}^\alpha$) will be of size equal to the number of filters, N. Let, M represent the number of cepstral coefficients. The linear transformation matrix ($\mathbf{A}^\alpha$) is given by:

$$\begin{bmatrix} \hat{\mathbf{C}}_{hf}^\alpha \\ \text{M x 1} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{D} \\ \text{M x N} \end{bmatrix}\begin{bmatrix} \mathbf{T}^\alpha \\ \text{N x N} \end{bmatrix}\begin{bmatrix} \mathbf{D}^{-1} \\ \text{N x M} \end{bmatrix}}_{\mathbf{A}^\alpha\,(\text{M x M})}\begin{bmatrix} \mathbf{C}_{hf} \\ \text{M x 1} \end{bmatrix} \quad (9)$$

Fig. 3 illustrates the differences in the warped-cepstra obtained using the conventional approach to VTLN with and without half-filters and also the warped-cepstra obtained using the linear transformation ($\mathbf{A}^\alpha$) approach. We observe that there are differences in the cepstra obtained with and without half-filters. The warped-cepstra obtained using the linear transformation approach are very close to the warped-cepstra obtained using the conventional approach with half-filters, which indicates that the linear transformation does a pretty good job.

Before proceeding further, we present results of the above discussed linear transformation approach to VTLN. For all the experiments, we have used $M = 16$ and $N = (20 + 2$ half-filters), i.e. $N = 20$ in the conventional approach and $N = 22$ in the modified filter-bank with half-filters. The experimental setup is presented later in Section. 4. The details of the corpus are presented in Table 1. In the EPPS task, tuning of the parameters is done on the development (*dev07*) set and the same parameters are used for recognition on the evaluation (*eval07*) set. The Aurora 4.0 task consists of 14 (2 clean + 12 noisy) test sets, with Test 01-07 being recorded from a similar microphone as that of the training data and Test 08-14 from a different microphone respectively.

The results are presented in Table 2. We observe that the baseline results are different for conventional (Conv.) and linear transformation (LT-) based approaches due to the difference
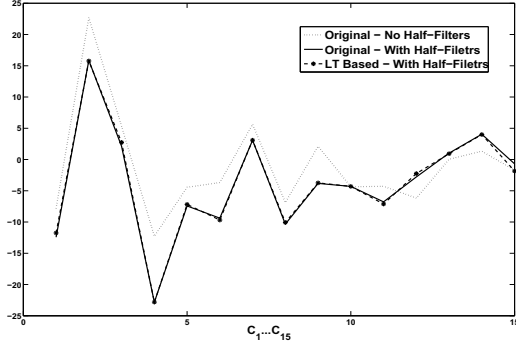
Figure 3: Illustrating the difference in warped-cepstra with and without half-filters along with the LT-based approach with half-filters ($\alpha = 0.88$).

Table 1: Description of the Corpus used for Experiments.

Corpus	Type	Speech [h]	Words
EPPS (ENGLISH)	Train06	87.70	7,61,234
	dev07	3.18	28,975
	eval07	2.85	27,360
AURORA 4.0	Clean Train	15.14	1,26,559
	Test(1 of 14)	0.34	2,769

in the filter-banks. For the EPPS task, the VTLN results using LT approach are comparable with conventional VTLN. For the Aurora 4.0 task, we observe that there are significant differences between the conventional and LT-based approaches both in baseline and after VTLN.

To get a better understanding of the differences in the Aurora 4.0 task, we examined the results for all the test sets individually. The results are presented in Table 3. Due to space constraint, we do not show all the results. We present results for all the noise conditions, which include both the microphones. Similar behavior is present on the test sets not shown. It is clear that the noisy speech data have inferior performance when half-filters are included during feature extraction. This behavior motivated us to re-investigate the linear transformation approach and ask the question: can the affect of half-filters be nullified in the linear transformation during feature extraction?

In the next section, we show that the linear transformation can be modified to undo the affect of half-filters during the feature extraction. We also show that the VTLN-warped cepstra obtained using the conventional and modified linear transformations are very close to each other.

3. Modification in the Linear Transformation for VTLN

In this section, we propose modifications to the linear transformation discussed in the previous section to undo the affect of half-filters in the feature extraction. We look at two different aspects of the linear transformation: one approach modifies the linear transformation and eliminates the affect of half-filters, while the other approach uses the interpolation matrix without the information of half-filters.

Table 2: Recognition results (%WER) on the conventional and linear transformation (with half-filters) approaches to VTLN.

Task	Type	Baseline		VTLN	
		Conv.	LT	Conv.	LT
EPPS (ENGLISH)	dev07	16.7	16.6	16.0	16.0
	eval07	15.7	15.8	14.9	15.1
AURORA 4.0	overall	23.9	26.6	21.8	23.6

Baseline - No VTLN   Conv. - Conventional;   LT - Linear Transformation

Table 3: Detailed recognition results (%WER) on Aurora 4.0 for the conventional and linear transformation (with half-filters) approaches to VTLN.

Test Condition	Baseline		VTLN	
	Conv.	LT	Conv.	LT
Test 01 (Clean)	6.4	5.8	5.3	5.6
Test 03 (Babble)	14.7	16.0	12.1	13.9
Test 05 (Street)	19.4	23.1	17.5	19.2
Test 07 (Train)	22.7	25.1	19.0	21.5
Test 09 (Car)	22.1	25.7	21.5	22.1
Test 11 (Restaurant)	37.7	41.4	34.9	38.3
Test 13 (Airport)	33.3	38.7	31.9	34.9

Baseline - No VTLN   Conv. - Conventional;   LT - Linear Transformation

3.1. Using the Cepstra with Half-filters

Excluding the half-filters in Figure 2, the rest of the filter-bank center frequencies exactly correspond to the conventional filter-bank used for generating MFCC features. Since the information of half-filters is only necessary while performing the interpolation, discarding the information before the DCT transformation will enable us to suppress the influence of half-filters during the feature generation. The modification in the linear transformation is given by:

$$\begin{bmatrix} \mathbf{C}^\alpha_{nhf} \\ {}_{M \times 1} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{D} \\ {}_{M \times L} \end{bmatrix} \begin{bmatrix} \mathbf{TR} \\ {}_{L \times N} \end{bmatrix} \begin{bmatrix} \mathbf{T}^\alpha \\ {}_{N \times N} \end{bmatrix} \begin{bmatrix} \mathbf{D}^{-1} \\ {}_{N \times M} \end{bmatrix}}_{\mathbf{A}^\alpha_{\mathbf{m}}(\mathbf{M} \times \mathbf{M})} \begin{bmatrix} \mathbf{C}_{hf} \\ {}_{M \times 1} \end{bmatrix} \quad (10)$$

The subscript *nhf* explicitly indicates that there are *no* half-filters. TR represents the truncation matrix, that suppresses the half-filters. It will be simply an identity matrix with the first and last columns as zeros and will be a rectangular matrix. Here, $M = 16$, $N = 22$ and $L = 20$. Comparing Eq. 9 and Eq. 10, we observe that both the transformations use the cepstra with half-filters ($\mathbf{C}_{hf}$). The modification is only done in the linear transformation, say $\mathbf{A}^\alpha_{\mathbf{m}}$ (the subscript $\mathbf{m}$ indicates the modified transformation), to undo the affect of half-filters. Now the forward and inverse DCT transformations are different.

3.2. Using the Cepstra without Half-filters

In this case, we assume that there are no half-filters and use the conventional filter-bank without any modification. As we have been arguing in the previous sections that the half-filters were introduced for performing proper interpolation, we also explore how well the interpolation matrix presented in Eq. 7 works without any modification in the filter-bank structure. By doing so, we are considering the first and last filters as the *zero* and *nyquist* frequencies respectively. The linear transformation is given by:

$$\hat{\mathbf{C}}^\alpha_{nhf} = \mathbf{A}^\alpha_{\mathbf{app}}.\mathbf{C}_{nhf} \quad (11)$$
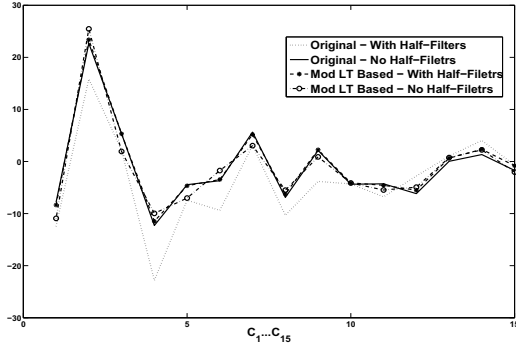
Figure 4: Illustrating the differences in the warped-cepstra with the proposed modifications to the linear transformations with and without half-filters ($\alpha = 0.88$).

This is the case that we would ideally like to have when performing VTLN as a linear transformation on conventional MFCC. The absence of half-filters will make the linear transformation now approximate and we denote it as $\mathbf{A}_{\mathrm{app}}^{\alpha}$. We use $M = 16$ and $N = 20$ in Eq. 9.

The main advantage with the above discussed modifications in the linear transformation is that, the baseline model will be same during warp-factor estimation either in conventional or LT-based VTLN.

Figure 4 illustrates the differences in the warped-cepstra obtained using the modified linear transformations discussed above. We observe that the warped-cepstra obtained using $\mathbf{A}_{\mathrm{m}}^{\alpha}$ match closely to the original warped-cepstra obtained using the filter-bank approach. We also observe that the warped-cepstra obtained using $\mathbf{A}_{\mathrm{app}}^{\alpha}$ have differences when compared to the original warped-cepstra, but follow the structure very closely. This also indicates that the proposed modifications allow us to obtain the warped cepstra very close to the ones obtained using the conventional VTLN approach.

The results comparing the performance of the proposed modifications in the linear transformation are shown in Table 4. We observe that, the baseline results are same for all the approaches. For the EPPS task, the performance of $\mathbf{A}_{\mathrm{m}}^{\alpha}$ on the *dev07* set slightly falls apart, but $\mathbf{A}_{\mathrm{app}}^{\alpha}$ has comparable performance with conventional VTLN. The results for the *eval07* set are all comparable. For the Aurora. 4.0 task, the proposed linear transformations perform comparably with conventional VTLN. The results also indicate that presence of half-filters during feature extraction affected the recognition performance.

4. Experimental Setup

All the experiments were done using the RWTH Aachen Speech Recognition System [7]. While performing feature extraction in the conventional case we have used 20 filters and obtained 16 cepstral coefficients $C_0...C_{15}$. The features were mean and variance normalized at the segment level and LDA over a window of nine consecutive frames was used to derive a 45 dimensional feature vector. The system used a classification and regression tree (CART) state tying, grouping the possible triphones into, 1501 for Aurora and 4501 for EPPS, generalized triphone states respectively. The VTLN warping-factors were estimated in training using forced-alignment and during recognition using a Gaussian mixture model (GMM) classifier.

Table 4: Recognition results (%WER) of the conventional and the modified linear transformation approaches to VTLN.

Task	Type	Base.	VTLN		
			Conv.	Mod LT	
				$\mathbf{A}_{\mathrm{m}}^{\alpha}$	$\mathbf{A}_{\mathrm{app}}^{\alpha}$
EPPS	dev07	16.7	16.0	16.2	15.8
(ENGLISH)	eval07	15.7	14.9	14.8	14.9
AURORA 4.0	overall	23.9	21.8	21.4	21.3

Base. - No-VTLN; Conv. - Conventional; Mod LT - Modified Linear Transformation

5. Conclusion

In this paper, we first presented a brief review of the linear transformation approach discussed in [1], using the idea of band-limited interpolation. We showed through recognition experiments that the presence of half-filters in the feature extraction might affect the recognition performance on noisy speech and proposed modifications to the linear transformation to undo this affect. We have analyzed the differences in the warped-cepstra obtained using the linear transformations discussed in this paper. We also presented the recognition performance of the modified linear transformations and showed that the performance is comparable to the conventional VTLN irrespective of clean or noisy speech. This also indicates that we can still perform VTLN as a linear transformation on conventional MFCC. The main idea behind revisiting the linear transformation proposed in [1] is to show that the inclusion of half-filters should not be considered as a constraint in deriving the linear transformation for VTLN, but depends on how we would like to handle the interpolated spectrum.

6. References

[1] D. R. Sanand and S. Umesh. Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN. In *Interspeech 2008*, pages 1233–1236, Sept. 2008.

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

[3] L. Lee and R. Rose. Frequency Warping Approach to Speaker Normalization. *IEEE Trans. Speech and Audio Processing*, volume 6, pages 49–59, Jan. 1998.

[4] S. Umesh, A. Zolnay, and H. Ney. Implementing Frequency Warping and VTLN through Linear Transformation of Conventional MFCC. In *Interspeech2005*, pages 269-272, Sept. 2005.

[5] M. Pitz; H. Ney Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. In *IEEE Trans. Speech and Audio Processing* , vol.13, no.5, pages 930-944, Sept. 2005.

[6] S. Panchapagesan and A. Alwan. Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC. *Computer Speech & Language*, volume 23, number 1, pages 42 – 64, 2009.

[7] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The RWTH Aachen University Open Source Speech Recognition System. In *Interspeech 2009*, pages 2111-2114, Sept. 2009.