Speaker Variability in Automatic Speech Recognition

Vikrant Tomar

McGill ID: 260394445

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

February 20, 2012

Preface

In the report, the natural logarithm is denoted with $\log$, and the base is stated otherwise. If both time and vector indices are used, the variable $\boldsymbol{h}_j(k)$ is the j'th coefficient at time index k, where the bold face represent a vector. Literature references are represented with numbers in IEEE style, e.g. [number], and a the full list of references is found in the References section.

Synopsis

Speaker Variability or Inter-speaker variability focuses on the variation in feature vectors from one speaker to another. It arises because the speech human produce depends upon a number of speaker specific characteristics, including age, culture, etc. But the major reason behind it is the variation in fundamental physiological characteristics of the speaker, such as Vocal Tract Length (VTL), and Glottal Pulse Rate (GPS). For a given speaker, in natural speech, the VTL is fixed, however, GPR is varied to indicated prosodic distinctions. However, recent researches have shown that, for SI ASR, VTL is far more crucial than GPR [1]. Inter-speaker variability is a major source of performance degradation in speaker independent (SI) automatic speech recognition (ASR). Also, we should note that the average length of female's vocal tract is 12 cms whereas that of male's is 17 cms, thus there is about 33% variability across the VTL's here. Accounting for this variability can improve the performance of SI ASR systems significantly.

The problem of speaker variability has been at focus of speech researchers for a long time, and many approaches have been proposed to address it. One such approach is Vocal Tract Length Normalization (VTLN), which refers to minimizing the effects of VTL variations among different speakers. Speaker Normalization is achieved in VTLN by warping the frequency spectrum of speech from different speakers, which – *in a very crude sense* – refers to adjusting the length of frequency axis by compressing or expanding it to fix the variability in speech [2]. VTLN has been found to be very effective, and achieves very low Word Error Rate (WER). However, to increase the effective of VTLN, or to further reduce the WER, it can be combined with other approaches [3, 4].

The classical methods of frequency warping are very cumbersome and computationally inefficient, and there have been a number of efforts to simplify the procedure or to achieve similar results using different procedures. This report is focused on a recently proposed approach in [5], where the authors has transferred the complex problem of VTLN into a simple Linear Transformation (LT) problem.

Contents

List of Figures

Chapter 1

Introduction

Speech is the fundamental mode of communication among humans. In addition, it also posses information about the language, gender, emotional state *etc.,* of the speaker. Over the last few decades, there have been a lot of emphasis in developing applications related to speech technology such as Automatic Speech Recognition (ASR), speech synthesis, speech understanding, *etc.* The ultimate goal of all these applications is to achieve successful human-machine interaction through speech. In this chapter, basics of ASR and challenges are described very briefly so that the reader could develop the necessary background for the rest of the report.

1.1 Challenges in ASR

ASR systems can be broadly classified as *speaker dependent* and *speaker independent*. A recognizer trained using speech data from a particular speaker is referred to as speaker-dependent (SD) recognition system, and it may not be accessible to other speakers. In contrast, speaker-independent (SI) recognition system in principle can be accessed by any speaker. Therefore, for improved performance it is important to minimize the effects of the individual characteristics and only consider the linguistic information that is common to all speakers uttering the same language. The major challenges in ASR arise due to the variability in the speech signal and can be accounted to the varying acoustic characteristics of speakers, the background noise as well as based on the size of the task, vocabulary and grammar.

The acoustics of the speech signal varies even for the same spoken text utterance,

on the account of physiological differences in the speech production organs, as well as the differences in linguistic style and habits, among others. The difficulty in a speech recognition problem also depends on the nature of the speech being recognized. The gamut of speech recognition systems can vary from an SD isolated word to a large SI vocabulary continuous speech recognition system, where co-articulation and other continuity effects increase the level of difficulty of the task. For instance, the read speech is much easier to transcribe than conversational speech.

1.2 Statistical Viewpoint of ASR

Most of the present day state-of-the-art ASR systems are based on statistical approaches, where are typical ASR problem translates to find the optimal sequence of words $\hat{\boldsymbol{W}}$ from all possible sequences of words $\boldsymbol{W}$ that yields the highest probability for the given acoustic feature sequence $\boldsymbol{X}$.

$$\hat{\boldsymbol{W}} = \arg\max_{W} p(\boldsymbol{W}|\boldsymbol{X}) \tag{1.1}$$

The posterior probability can further be decomposed using Bayes's rule:

$$p(\boldsymbol{W}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{W})p\boldsymbol{W}}{p(\boldsymbol{X})} \tag{1.2}$$

where

- $p(\boldsymbol{X}|\boldsymbol{W})$: probability of observing $\boldsymbol{X}$ under the assumption that $\boldsymbol{W}$ is a true utterance.
- $p(\boldsymbol{W})$: probability that word sequence $\boldsymbol{W}$ is uttered.
- $p(\boldsymbol{X})$: average probability that $\boldsymbol{X}$ will be observed.

Note that the denominator does not depend on the word sequence $\boldsymbol{W}$, and $p(\mathbf{X})$ is same for all candidate word sequences. Therefore, it does not affect the outcome of finding the optimal sequence, and we may safely ignore it and write:

$$\hat{\mathbf{W}} = \arg\max_{W} \left\{ \overbrace{p(\boldsymbol{X}|\boldsymbol{W})}^{acoustic\ probability} \times \underbrace{p(\boldsymbol{W})}_{language\ probability} \right\} \tag{1.3}$$

As denoted in the expression 1.3, the entire decision model thus can be decomposed into acoustic and language model properties.

1.3 Feature Extraction

One of the primary building blocks of an ASR system is signal analysis, which refers to parametrize the speech signal to best suit the ASR model. Since in an ASR system, *what* is spoken is more important than *who* spoke it, the parametrization step should be aimed at capturing the relevant information, while discarding the irrelevant information. One crucial and formidable aspect of this parametrization step is to eliminate any variability in the speech due to speaker and/or environment.

The signal analysis of the present day ASR systems is based on short term spectral analysis [6], usually Fourier analysis. For further processing and smoothing, Mel frequency cepstral coefficients (MFCC) filtering is widely used, [7]. Such a filter bank is referred to Mel-filter bank. Various signal processing steps in the computation of MFCC features are illustrated in Fig. 1.1.

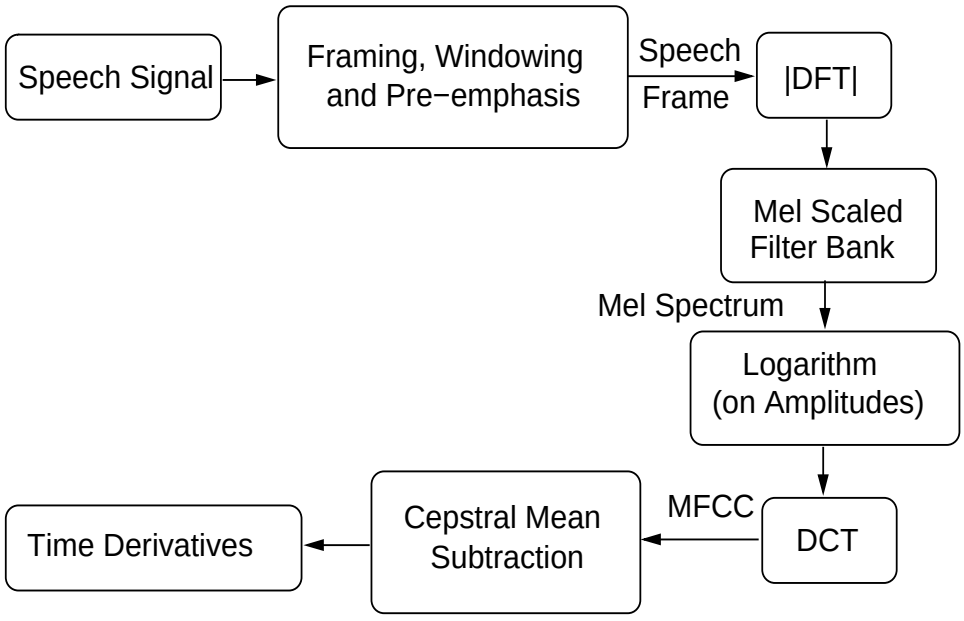

Fig. 1.1 MFCC Feature Extraction

The most important step of the procedure is the Mel filter bank processing. This filter

converts DFT spectrum to Mel frequency spectrum. The DFT coefficients are binned into required number of channels by multiplying them by weights corresponding to Mel-scaled triangular filter mask. The relation between linear frequency (Hz), and Mel frequency is given by

$$\begin{aligned} f_{Mel} &= 1127 \log \left(1 + \frac{f_{Hz}}{700}\right) \\ \text{or,} \quad \mathrm{f_{Mel}} &= 2595 \log_{10} \left(1 + \frac{\mathrm{f_{Hz}}}{700}\right) \end{aligned} \tag{1.4}$$

where, f_{Mel} and f_{Hz} correspond to the frequencies in Mel and linear domains respectively. The structure of a typical Mel filter bank is shown in Fig. 1.2. It should be noted that the filter has higher resolution at lower frequencies and lower resolution at higher frequencies, which is motivated from the processing in the human ear, [6]. The output of the Mel filter bank is then log compressed and converted into cepstral coefficients by applying discrete cosine transform (DCT), and long term mean (typically that of a sentence) is subtracted from the coefficients in order to remove time-invariant distortions introduced by the transmission channel and the recording device.

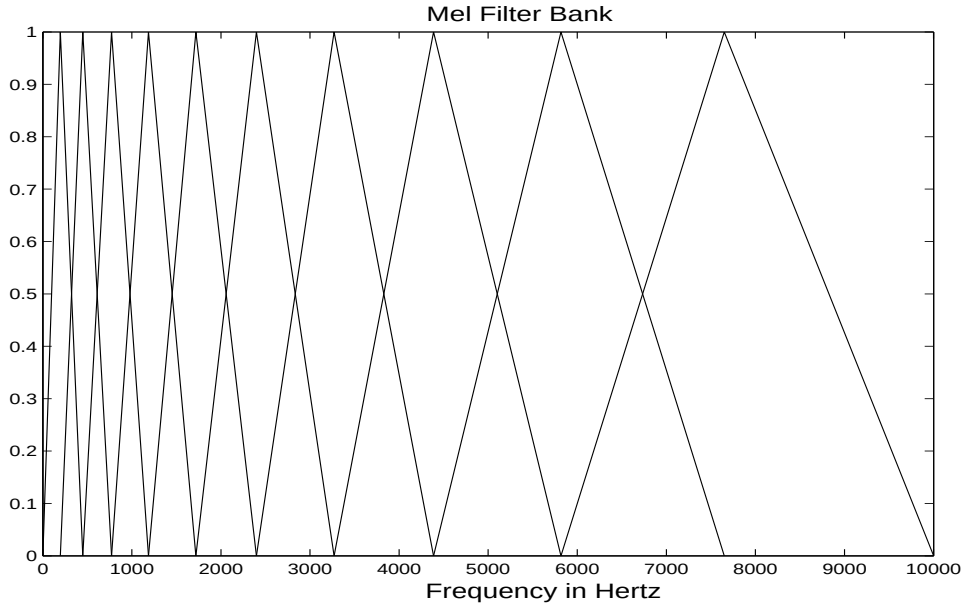

Fig. 1.2 A Typical Mel filter bank

The output vectors of this process are known as MFCC features of the speech signal, and constitutes the basis of modern-day speech recognition systems. Having described the basics of ASR in this chapter, we would look at speaker variability in ASR, and use of frequency warping in the next chapter.

Chapter 2

Speaker Variability

Of the various challenges for ASR, discussed in the previous chapter, the variability due to speaker is considered to be a major source of performance degradation. This variability may be related to the physiological differences among speakers, linguistic variations, or even the psychological and mental conditions. Though it is generally agreed upon that VTL is a major source of speaker variability, even more crucial than Glottal Pulse Rate (GPR), [1]. Since, in statistical models of ASR, the recognition is performed by matching the observed unknown speech to the model which most likely generated the sequence, the system should be able to handle the variabilities in speech.

In general, two speakers enunciating the same sound have different pressure waveforms, and the resulting spectra is very different. As described in previous chapter, the ASR systems use features extracted from the spectra, thus the features themselves would be different for the same utterance by two speakers[1]. Therefore, for an ASR system, recognizing these different features as belonging to the same sound is difficult, and leads to degradation in performance.

2.1 Formant Frequencies

The dominant peaks in the smooth spectrum of an utterance are known as formant frequencies or formants. The positions of these formants uniquely characterize a specific utterance.

[1]It is important to note that there will be a certain amount of variability for the same sound spoken by the same person, *i.e.*, intra-speaker variability, which is handled by the statistical model. However, in general, intra-speaker variability is substantially smaller than the inter-speaker variability.

Formants are considered to be resonances of the vocal tract, with the assumption that vocal tract is a hollow tube that is excited by the air coming out of lungs.

The positions of the resonant frequencies are inversely proportional to the length of the tube (or the vocal tract).

$$F \propto \frac{1}{VTL} \tag{2.1}$$

The smaller VTL in females results in higher formants. As mentioned earlier, this variability in VTL may be as high as 33% between an average male and female pair. And it is increased further for children.

This is a significant variation, and – if accounted for properly – can improve the performance of SI ASR systems. Thus the ASR systems should be trained in a that the acoustic classes are well separated, and it is robust to variability across speakers.

2.2 Accounting for Speaker Variability

In literature, there are two major classes of techniques defined to handle speaker variability, namely Speaker Adaptation and Speaker Normalization.

Speaker Adaptation mainly deals with the parameters of the acoustic models, and – as the name suggests – modifies them in a manner to best suit the test data, and hence reducing the mismatch between the acoustic model and acoustic vectors. Speaker Normalization, on the other hand, transforms the feature vectors to best suit the model. In this report, we will only be focusing on Speaker Normalization.

The fundamentals of Speaker Normalization are based upon the framework of physical model of speech production, which highly depends upon the VTL. The predominant choice for speaker normalization has been vocal tract length normalization (VTLN), which tries to compensate for the variabilities arise due to the differences in VTL [2, 3, 5, 8]. The primary focus of researcher have been to derive a parametric model for vowel production and normalization. Most of the VTLN methods attempt to emulate the effect of resizing the vocal tract by warping the frequency axis, and the simplest warping is to linearly scale the frequency axis as,

$$f_{warped} = \alpha f_{orignal} \tag{2.2}$$

where α is known as the warp factor.

Andreou *et al.* [9] proposed an estimation approach, where the warp factor was esti-

mated by maximum likelihood grid search over the warped version of acoustic vectors for possible values of the warp factor and the warping was done by resampling the speech. Extending the idea of Andreou *et al.*, Lee and Rose [2] proposed a number of novel and successful ideas for VTLN. They proposed a fast text-independent approach to warp factor estimation using Gaussian mixture model (GMM) trained for each of the warp category. For normalization during training, they proposed an iterative procedure and while testing they proposed a multiple-pass procedure for normalization using HMM based speech models. Furthermore they proposed to incorporate linear frequency warping efficiently by inversely scaling the center frequencies and bandwidths of the Mel filter bank.

Before proceeding further, a brief review of VTLN that is followed conventionally in SI ASR is presented.

2.3 Vocal Tract Length Normalization

The main objective of VTLN is to find an optimal warping factor to warp the frequency axis of the speech signal so that the spectra of speaker enunciating the same sound appear similar.

Since, there exists no reference or *golden speaker* with respect to whom an optimal warp factor α can be estimated, a maximum likelihood (ML) based grid search over a discrete set of α's is followed. The optimization is performed with respect to the SI or the previous iteration of the VTLN model. The range of α is usually restricted between 0.80 and 1.20 based on physiological arguments of the vocal tract and an incremental step size of 0.02 is usually followed in practice.

The optimal warp-factor estimation for the i^{th} utterance C_i is given as following.

$$\hat{\alpha}_i = \arg\max_{\alpha} p(C_i^{\alpha}|\lambda, W_i) \tag{2.3}$$

Where, C_i^{α} represents the warped features of the i^{th} utterance, λ is the model, and W_i is the true transcription during the training or first pass recognition during testing. In expression 2.3, the likelihood has to be calculated for all values of α before finding an optimal estimate, hence it is computationally expensive.

Following are the steps in VTLN based speaker normalization:

- Generate warped features for all values of α in the search range. These features

are obtained by changing the filter bank structure for each value of α as shown in Fig. 2.1. The filter bank incorporates both Mel and VTLN warping for efficient implementation. The entire signal processing step need to be repeated for all the warp factors in the search range to generate the VTLN warped features. The entire process of generating warped features is summarized in Fig. 2.2 where, P represents the power or magnitude spectrum, and F_m the scaling of the filter bank.

- Estimate the optimal warp factor $\hat{\alpha}$ using equation 2.3. The features corresponding to the optimal warp factor are then taken as normalized features.
- Using the normalized features obtained above, the model parameters are updated for the case of training or used for recognition during testing.

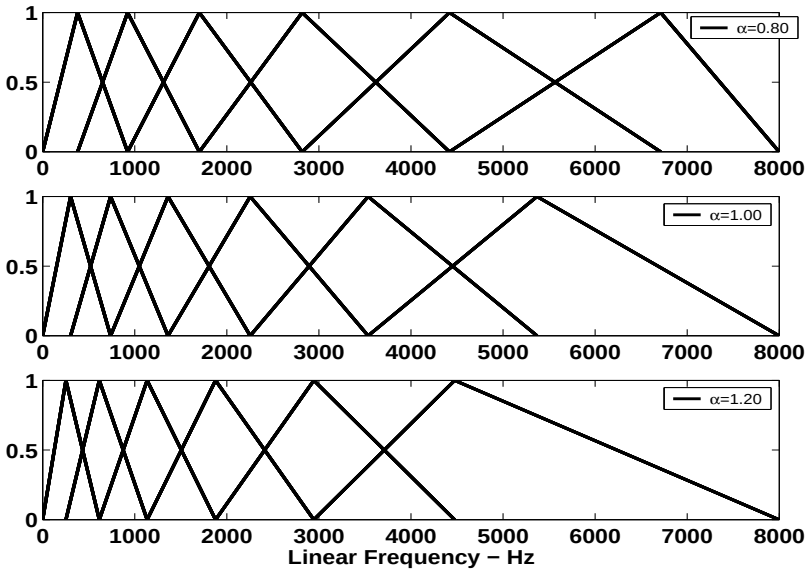

Fig. 2.1 A Mel filter bank with VTLN warping. The filters have non-uniform center frequencies with non-uniform bandwidths.

Algorithm Analysis

VTLN has following merits and demerits:

- It requires very low amount of training data.

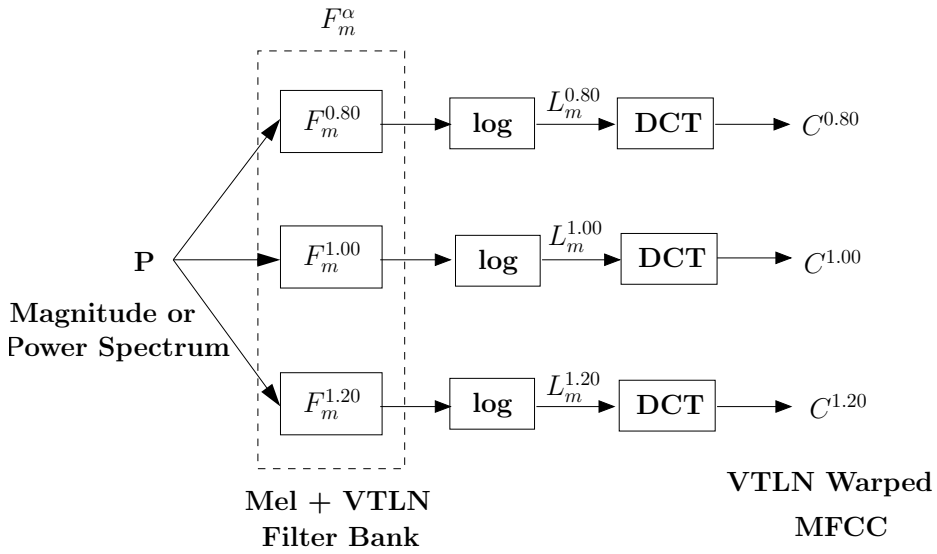

Fig. 2.2 A framework for generating warped features. The filter bank in inversely scaled instead of resampling the speech signal for each warp factor for efficient implementation.

- Most SI ASR systems use only one warping parameter for the normalization of whole frequency spectrum. This is inconsistent with the fact that the vocal tract is not a single uniform tube; in particular, there are greater individual differences in pharynx length than in oral cavity length. However, this linear scaling model is found to be equally competitive with all the proposed non-linear warping functions.

- For each frame, VTLN requires generation of all the warped feature before hand (using equation 2.3).

- Computationally expensive, as it requires decoding for all possible warping factors.

Thus, despite of being a robust and effective algorithm, VTLN is very expensive computationally. There is a need for methods to reduce the computational complexity of VTLN, while retaining its competitiveness. The authors in [3, 5] proposed a linear approach to VTLN using simple matrix multiplications, which is investigated in the next chapter.

Chapter 3

Linear Transform based VTLN

This chapter details the approach proposed in [5], which uses only a linear transformation on conventional MFCC coefficients to obtain VTLN warped features, *i.e.,*

$$C^{\alpha} = A^{\alpha} C \tag{3.1}$$

where, C represents the MFCC features, C^{α} represents the VTLN warped features, and A^{α} is the linear transformation (LT) matrix. The authors have argued that separating the VTLN warping from the Mel filter bank helps in deriving the linear transform. One apparent advantage of this approach is that it eliminates the need of updating the filter bank structure for every warping factor in range, thus promises a significant improvement in complexity.

3.1 Framework

For a bandlimited continuous-time signal $x(t)$, given uniformly spaced samples of the signal that are appropriately sampled (above Nyquist rate) $x(tn)$, the continuous-time signal, can be exactly reconstructed. This idea is exploited to obtain the LT for VTLN-warping, except that the signals are considered to be quefrency limited instead of being frequency-limited, as cepstrum domain. The conventional Mel warped smoothed spectral output is obtained in the linear frequency (Hz) domain by applying the triangular averaging Mel filter bank on the linear frequency (Hz) magnitude spectrum. The filter bank has non-uniformly spaced and non-uniform bandwidth filters as shown in Fig. 1.1. In the approach proposed in

[5], the Mel-warped smoothed spectral output is obtained in the Mel-frequency domain by applying uniformly spaced and uniform bandwidth filters on the Mel-warped magnitude spectrum as shown in Figure 3.1.

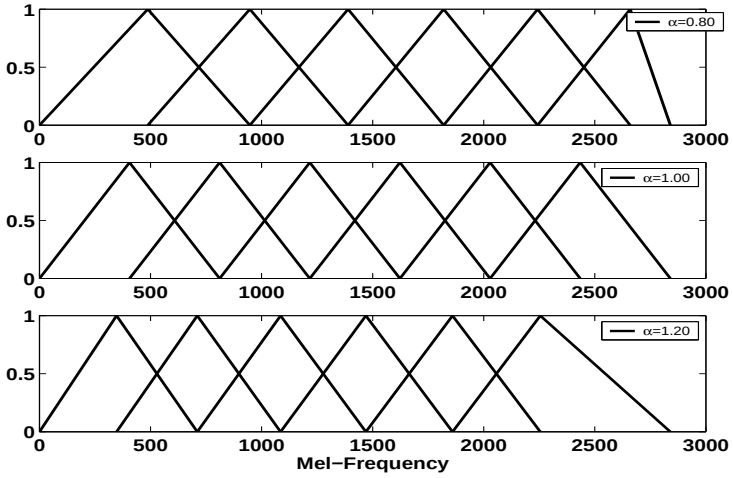

Fig. 3.1 New Mel filter structure. Filters have uniformly spaced center frequencies with uniform bandwidth for $\alpha = 1.00$. However, they are non-uniformly distributed spaced where α is not unity.

Therefore, the Mel-warped magnitude spectrum can be interpreted as being convolved with a triangle function to obtain the Mel-warped smoothed spectrum followed by the log compression on the amplitude to get the continuous function in Mels, *i.e.*, $L_m(\nu)$. This function is then uniformly sampled at $\nu_l = 2\pi l/N$ where $l = 0, 1, , (N-1)$ in the normalized-Mel domain. These uniformly spaced samples correspond to the output of the triangular filters centered at those particular Mel frequencies and exactly correspond to the elements of the vector $\boldsymbol{L_m}$. Because of the triangle smoothing and subsequent log-operation on the output (which reduces dynamic range), this smoothed log filter-bank output is content within the low quefrency region.

During VTLN-warping, the filter center frequencies are appropriately scaled in the linear frequency (Hz) domain as a function of inverse of α. This corresponds to the center frequencies of the filter-bank to be non-uniformly spaced in the Mel frequency domain as shown in Fig. 3.1. Since the log-compressed Mel-warped smoothed magnitude spectrum is represented by the continuous function $L_m(\nu)$, the output of the VTLN-warped filter-bank corresponds to sampling $L_m(\nu)$ non-uniformly, giving $L_m[\tilde{\nu}_l]$. These non-uniformly spaced samples exactly correspond to the elements of the vector $\boldsymbol{L}_m^{\alpha}$.

3.2 Principal

The basic idea is that, with no VTLN warping, the conventional log-compressed Mel filter bank outputs $\boldsymbol{L}_m$ are uniformly spaced samples in the Mel-domain. During VTLN warping, the outputs of the log-compressed Mel filter-bank ($\boldsymbol{L}_m^{\alpha}$) are non-uniformly spaced in the Mel-domain. The problem is one of estimating the non-uniformly ($\boldsymbol{L}_m^{\alpha} = L_m[\tilde{\nu}_l]$) spaced samples given the uniformly spaced samples ($L_m[\nu_l]$). This can be easily achieved through bandlimited interpolation as Mel filter-bank outputs are quefrency limited. This is assured in this work through filter-bank smoothing. Let $L_m(\nu)$ and $\boldsymbol{C}_{DFT}$ (these are not MFCC coefficients since DCT is not used) form a discrete Fourier transform (DFT) pair. Then sampling $L_m(\nu)$ would result in periodic repetition of $\boldsymbol{C}_{DFT}$. As long as $\boldsymbol{C}_{DFT}$ is strictly quefrency limited, and the sampling rate is sufficiently high, there is no aliasing in the cepstral domain. In such a case, the value of $L_m(\nu)$ at any Mel-frequency $\tilde{\nu}_l$ can be found from its uniformly-spaced samples at ν_l through bandlimited interpolation. This is basically exploiting the sampling theorem, where a signal (in this case a frequency domain signal) can be reconstructed from its samples by using Sinc-interpolation. Note that while implementing the bandlimited (Sinc) interpolation, the uniformly spaced samples are assumed to range from index 0 to $(N-1)$. Hence the outputs at zero-Mel and Nyquist-Mel are needed, which are obtained by having extra (half-)filters centered at *zero-Mel* and *Nyquist-Mel*, as illustrated in Fig. 3.2.

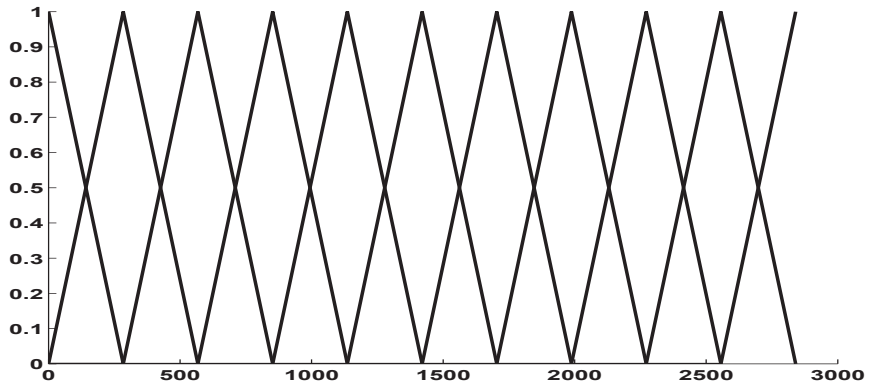

Fig. 3.2 Modification in the filter-bank construction (shown in Mel-frequency domain) for performing bandlimited interpolation. *There are half-filters at zero-Mel and Nyquist-Mel frequencies.*

3.3 Obtaining the Transformation Matrix

Let $\nu_0,\ \nu_1,\ \nu_2,\ \ldots.,\ \nu_{N-1}$ represent the uniformly-spaced Mel-frequencies of L_m. Their respective linear-frequencies (Hz) are non-uniformly spaced and are represented as $f_0,\ f_1,\ \ldots,\ f_{N-1}$. These are the center frequencies of the N Mel-filters in the linear frequency (Hz) domain. They are related through the standard Mel-relation, *i.e.*,

$$\nu_l = 2595 log_{10}(1 + \frac{f_l}{700}); \quad \forall l \tag{3.2}$$

During VTLN-warping, the warping function $g_a(f)$ is applied to obtain the warped frequencies. Let $\hat{f}_l = \alpha f_l$, denote the linear warping in the linear frequency (Hz) domain. Note that this would not be linear-warping in the Mel-frequency domain. The corresponding center frequencies of the filters in the Mel-domain ($\hat{\nu}_l$) are related to $\hat{f}_l$ through a similar relation as in 3.2, with $\hat{\nu}_l = \nu_l$ for $\alpha = 1.00$.

The inverse Fourier relation between $\boldsymbol{C}_{DFT}$ and $\boldsymbol{L}_m$ is given by,

$$c_k = \frac{1}{2N-2} \sum_{l=0}^{2N-3} L_m \left[\frac{\nu_l}{\nu_s}\right] e^{j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k} \tag{3.3}$$

where, ν_s is the sampling frequency in the Mel frequency domain. Here the signal is assumed to be periodic with a period of $2N-2$ and symmetric around $N-1$. The half-filters are present at indices 0 and $N-1$. As discussed previously, the values at these indices are required for performing bandlimited interpolation. If c_k is assumed to be quefrency limited, the elements of $\boldsymbol{L}^a_m$ can be determined as following.

$$L_m \left[\frac{\hat{\nu}_l}{\nu_s}\right] = \sum_{k=0}^{2N-3} c_k e^{-j\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s})k} \tag{3.4}$$

Substituting equation 3.3 in 3.4:

$$\begin{aligned} L_m \left[\frac{\hat{\nu}_l}{\nu_s}\right] &= \sum_{k=0}^{2N-3} \frac{1}{2N-2} \sum_{l=0}^{2N-3} L_m \left[\frac{\nu_l}{\nu_s}\right] e^{j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k - j\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s})k} \\ &= \sum_{l=0}^{2N-3} L_m \left[\frac{\nu_l}{\nu_s}\right] \left[\frac{1}{2N-2} \sum_{k=0}^{2N-3} e^{j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k - j\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s})k}\right] \end{aligned} \tag{3.5}$$

Thus, the transformation matrix between $L_m[\nu_l]$ and $L_m[\hat{\nu}_l]$ is given by,

$$\boldsymbol{T}^{\alpha} = \frac{1}{2N-2} \sum_{k=0}^{2N-3} e^{j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k - j\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s})k} \tag{3.6}$$

where $i = 0, 1, ..., 2N-3$. Using the even-symmetry property, the $N \times N$ interpolation matrix $\boldsymbol{T}^{\alpha}$ is given by,

$$\hat{\boldsymbol{T}}^{\alpha} = \frac{1}{2N-2} \sum_{k=0}^{N-1} 2a_l cos\left(\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s}k\right) cos\left(\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s}k\right) \tag{3.7}$$

where,

$$a_l = \begin{cases} \frac{1}{2}; & l = 0, N-1 \\ 1; & l = 1,\ 2, ..., N-2 \end{cases} \tag{3.8}$$

And, we can obtain the warped feature vectors as,

$$\boldsymbol{L}_m^{\alpha} = \hat{\boldsymbol{T}}^{\alpha}.\boldsymbol{L}_m \tag{3.9}$$

Thus the complex problem of frequency warping has been reduced to a mere linear transformation task. Note that $\boldsymbol{C}_{DFT}$, or c_k for that matter, has been completely eliminated from the equations, and has not been used in the final calculation. It was presented only for better understanding in the derivation of the bandlimited interpolation matrix.

Chapter 4

Conclusion

The work discussed in this report has proposed computationally efficient methods for Vocal Tract Length Normalization (VTLN). In particular, a method based on linear transformation was presented.

It is argued that a linear transformation in the conventional MFCC frame work can be realized by separating the VTLN warping from the Mel filter bank smoothing. This completely eliminates the inversion of the filter bank smoothing and bypasses the non-linear log operation. Obtaining a linear transformation completely eliminates the need to generate the warped features in advance and can be generated on the fly similar to the adaptation based approaches.

4.1 Future Work

- Many studies indicate that non-linear scaling model better describes the relationship between spectra of speakers enunciating the same sound. However, the nature of nonlinear relation is still an unsolved problem and is actively pursued.
- Obtaining a linear transform also enables to use VTLN matrices in the adaptation framework. There have been recent efforts to use class specific VTLN warping matrices using regression class tress for performing speaker normalization [3].

References

[1] E. Gaudrain, S. Li, V. S. Ban, and R. D. Patterson, "The role of glottal pulse rate and vocal tract length in the perception of speaker identity," in *Interspeech '09*, ISCA, Brighton, UK. 2009.

[2] L. Lee and R. Rose, "A Frequency Warping Approach to Speaker Normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan 1998.

[3] S. P. Rath and S. Umesh, "Acoustic Class Specific VTLN-Warping using Regression Class Trees," in *Interspeech '09*, ISCA, Brighton, UK. 2009.

[4] S. Demange and D. V. Compernolle, "Speaker Normalization for Template based Speech Recognition," in *Interspeech '09*, ISCA, Brighton, UK. 2009.

[5] D. R. Sanand, *Linear Transformation Approaches to Vocal Tract Length Normalization for Automatic Speech Recognition.* Phd Thesis, Dept. of Electrical Engineering, IIT Kanpur, July 2009.

[6] D. O'Shaughnessy, *Speech Communication: Human and Machine.* New York, NY, USA: IEEE, 2002.

[7] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357 – 366, Aug 1980.

[8] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *ICASSP '96: IEEE International Conference on Acoustics, Speech, and Signal Processing 1996*, pp. 346–349, IEEE.

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