Sādhanā Vol. 36, Part 5, October 2011, pp. 853–883. © Indian Academy of Sciences

Studies on inter-speaker variability in speech and its application in automatic speech recognition

S UMESH

Department of Electrical Engineering, Indian Institute of Technology-Madras, Chennai 600 036, India
e-mail: umeshs@iitm.ac.in

Abstract. In this paper, we give an overview of the problem of inter-speaker variability and its study in many diverse areas of speech signal processing. We first give an overview of vowel-normalization studies that minimize variations in the acoustic representation of vowel realizations by different speakers. We then describe the universal-warping approach to speaker normalization which unifies many of the vowel normalization approaches and also shows the relation between speech production, perception and auditory processing. We then address the problem of inter-speaker variability in automatic speech recognition (ASR) and describe techniques that are used to reduce these effects and thereby improve the performance of speaker-independent ASR systems.

Keywords. Vowel-normalization; vocal-tract length normalization; speech-scale; frequency-warping; linear transformation of cepstra; speaker-adaptation.

1. Introduction

In this review article, we study the variability in the acoustic signal among different speakers enunciating the same sound. Although, humans can quite often correctly recognize same sound enunciated by different speakers, there are significant differences in the pressure variations and therefore the frequency content of these signals corresponding to the same perceived sound. The classic study of Peterson & Barney (1952) found that the same vowel produced by different speakers have very different formant frequencies (or dominant resonances) while different vowels spoken by different speakers *can* have very similar formant (or dominant) frequencies. This is shown in figure 1. Understanding these variations has important implications in many areas of speech processing.

Ladofoged & Broadbent (1957) classified the sources of variation in speech signal into linguistic or phonemic variation and speaker-related variation. They classify speaker-related information itself into personal variation and socio-linguistic variation. Personal variations are due to differences between speakers in the shape and size of their vocal tract and larynx and correspond to anatomical or physiological variations. Socio-linguistic variations originate from differences in regional background, education level and gender of speaker. A more detailed analysis can be

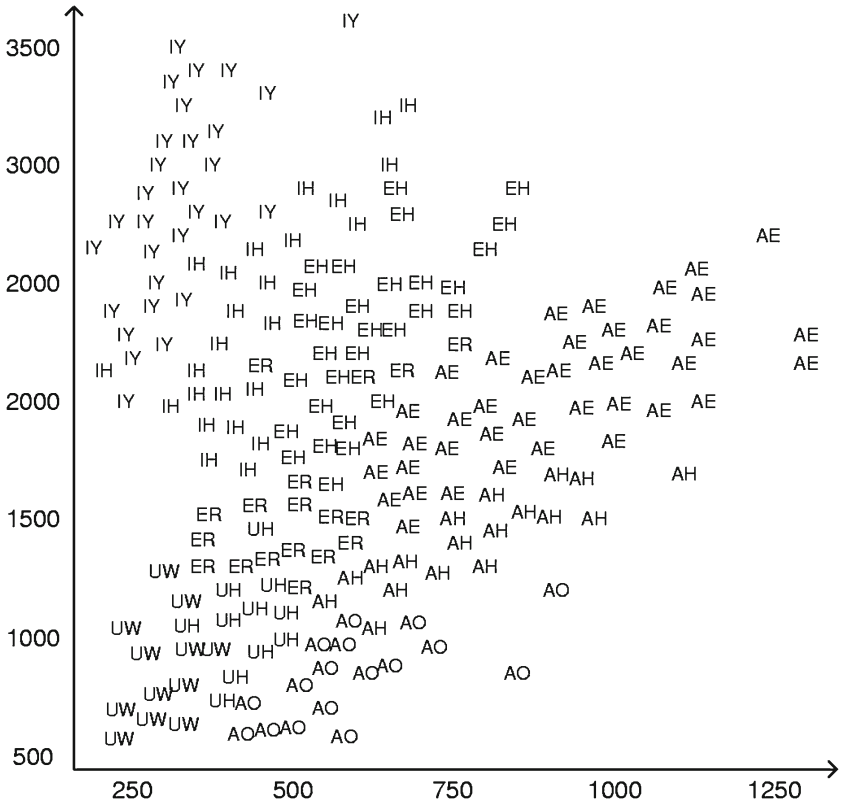


Figure 1. This figure is taken from the classic study of Peterson & Barney (1952) which shows that the formant frequencies F_1 and F_2 have significant variations among different speakers enunciating the same vowel. It also shows the overlap between the different vowel classes.

found in Adank (2003) and Adank *et al* (2004). Some researchers also have a third source of variation attributed to the emotional state of the speaker. Physiological variations are often regarded as the major source of inter-speaker variations and most studies in literature have focussed on reducing the variations due to anatomical differences. In this review, therefore, we will focus only on studies and methods that attempt to reduce inter-speaker variations by *normalizing* the physiological differences.

Much work in normalization is focussed on developing procedures to improve the classification performance of vowels by reducing inter-speaker variations and thereby improving the separability between vowel classes. These are usually referred to as vowel normalization procedures. There are two broad normalization procedures – formant-based and whole-spectrum-based. In formant-based procedures, normalization is done to minimize the variations within a vowel class spoken by different speakers, while maximizing the separability between vowel classes by transforming the values of the formant (or dominant) frequencies. Whole-spectrum-based normalization procedures use the entire frequency-spectrum, instead of estimating dominant frequencies or formants. Vowel normalization procedures are also classified as intrinsic and extrinsic procedures. Intrinsic vowel normalization procedures use information *only* from that acoustic

realization of that vowel and typically involve non-linear transformation of the frequency scale (Syrdal & Gopal 1986). Extrinsic procedures, on the other hand, use information distributed across many vowel categories for improving vowel classification (Gerstman 1968; Lobanov 1971; Nordström & Lindblom 1975; Nearey 1978). A related question that has been subject of much research is to find the relationship between spectra (or formants) for the same sound enunciated by different speakers. These topics are discussed in the initial sections of this review paper.

Characterizing speaker-variability is also very important in many of today's speech technologies – and we focus on this topic in the later sections of this paper. For example, in speaker-independent automatic speech recognition (SI-ASR), we are interested in 'what is said' rather than 'who said it'. In such cases, the phonemic or linguistic information is useful, while the speaker-related information is noise or unwanted. On the other hand, in text-independent speaker verification task where one identifies a speaker based on his/her voice signal, phonemic information is ignored and speaker-related information is used to help improve the performance. In this review paper, we will describe two standard procedures that are used to reduce inter-speaker variations to improve the performance of SI-ASR, namely, vocal-tract length normalization (VTLN) and speaker-adaptation. In particular, we will discuss the advantages and disadvantages of these two approaches for speaker-normalization and their computational complexity. We also discuss our recent work in developing an approach to combine the advantages of these two approaches and also our efforts in reducing computational complexity of VTLN.

The paper is organized as follows: In section 2, we first review studies done in the area of vowel normalization which try to find a representation that is insensitive to acoustic variations for different speakers enunciating the same sound. We then review our previous work on universal warping function approach to normalization. We show that the universal warping function is remarkably similar to a psycho-acoustic scale called Mel-scale. The Mel-scale has been found to be important in many areas of speech-processing and it is interesting that Mel-scale is also useful in speaker-normalization. Section 4 discusses the affine-model that describes the relationship between two speakers enunciating the same sound. In section 5, we introduce VTLN which is commonly used in most state-of-the-art SI-ASR systems to overcome the problem of inter-speaker variability. We also discuss the advantages and disadvantages of VTLN approach to normalization and the computational complexity involved in implementing it in practice. In section 6, we review many approaches that have been proposed to obtain a linear-transformation for VTLN. These approaches are motivated by a desire to improve the computationally efficiency of VTLN. We also review an alternate approach to reduce speaker-variability in SI-ASR, where the model parameters are linearly transformed to better match the test-speaker data. These techniques are referred to as speaker-adaptation. Finally, we discuss the advantages and disadvantages of VTLN and speaker-adaptation methods and show how our recently proposed transform-based VTLN (T-VTLN) combines the best of both approaches.

2. Vowel normalization

As mentioned in the introduction, phonetically identical utterances show a great deal of acoustic variation across speakers, even though human listeners are able to correctly identify them despite these variations. It is well known that the first two formants, namely F_1 and F_2, are sufficient to perceptually identify a particular vowel. A dramatic illustration of the acoustic variation is the study by Peterson & Barney (1952) where there is significant overlap between vowel classes

in the $F1 - F2$ plane and yet the human listeners correctly identified all the vowels. Vowel-normalization research tries to find the commonality between these acoustic variations which help listeners correctly identify them.

2.1 *Formant-ratio theory*

One of the earliest approaches to vowel normalization is the formant-ratio theory which is based on the idea that the vowels are relative patterns and not absolute frequencies. Therefore, it is only the ratio of the formants that are important and not the actual frequency values. For example, although absolute values of formant frequencies for a vowel may be different for different speakers, the ratio $\frac{F_2}{F_1}$, $\frac{F_3}{F_2}$, etc. would be similar. Equivalently, $\log(F_2) - \log(F_1)$, etc. would be similar among different speakers for the same vowel. There have been a number of different variations of this idea proposed by many researchers (Peterson 1961; Sussman 1986; Syrdal & Gopal 1986; Miller 1989).

Some of formant-ratio formulations include those by Syrdal & Gopal (1983) and Miller (1989).

$$\text{Syrdal–Gopal: } \begin{cases} \mathcal{B}(F_1) - \mathcal{B}(F_0),\ \mathcal{B}(F_2) - \mathcal{B}(F_1), \\ \mathcal{B}(F_3) - \mathcal{B}(F_2) \end{cases}$$

$$\text{Miller: } \log\left(\frac{F_1}{\text{SR}}\right),\ \log\left(\frac{F_2}{F_1}\right),\ \log\left(\frac{F_3}{F_2}\right),$$

where $\mathcal{B}(F_i)$ is the 'Bark' equivalent of i^{th} formant frequency F_i and SR is the sensory reference derived from the geometric mean of F_0 over an interval of time. The Bark scale is a non-linear function of physical frequency, f, in Hz and is approximately logarithmic for high frequencies and almost linear for low frequencies. Nearey (1978) used constant log-interval normalization given by

$$\text{Nearey: } \log(F_1) - \mu_{\text{L}},\ \log(F_2) - \mu_{\text{L}},\ \log(F_3) - \mu_{\text{L}},$$

where μ_{L} is mean log-value of the speakers' F_1 and F_2. Bladon *et al* (1983) proposed a normalization method based on the observation that the average frequency difference between formants of vowels produced by men and those produced by women approximately differ by 1 Bark. This method uses the whole spectrum and not just the formant frequencies. Therefore, their method of normalization involved shifting the women spectrum by 1 Bark unit.

An alternate approach to normalization involves algorithms which assume that all the information about *all* the vowels and the formant frequencies are available for a speaker before normalization is applied. Some of these approaches include those of Gerstman (1968), Lobanov (1971) and Nearey (1978). Although, it does not seem reasonable that in practice a listener has information about all vowels spoken by an unfamiliar speaker to recognize speech to perform this kind of normalization, nevertheless, these methods are useful when automatically classifying vowel classes since they often increase the separability of the overlapping clusters of vowels.

2.2 *Normalization using vocal-tract length – uniform and non-uniform scaling*

VLTN is another approach to reduce inter-speaker variability, where the formants are on a speaker-specific co-ordinate system which is determined by the length of the speaker's vocal-tract. The frequency of the third formant, i.e. F_3, is causally linked to vocal tract length and was used explicitly by Nordström & Lindblom (1975) in a vowel normalization algorithm. Nordström &

Lindblom (1975) proposed a normalization procedure in which *all* the formants of *all the vowels* are scaled by a constant scale factor. This is referred to as uniform scaling. The scale-factor is based on an estimate of the speaker's average vocal-tract length with respect to that of a reference speaker. The vocal-tract length is related to third formants, F_3 of open vowels. Therefore, in this approach, all the formant frequencies of the subject to be normalized are simply divided by the factor $\alpha_{\mathrm{rs}} = \frac{\widehat{F}_{3,\mathtt{s}}}{\widehat{F}_{3,\mathtt{r}}}$, where $\widehat{F}_{3,\mathtt{s}}$ and $\widehat{F}_{3,\mathtt{r}}$ are the *average* F_3 of open vowels of the subject, $\mathtt{s}$ and the reference speaker, $\mathtt{r}$ respectively.

Fant (1975) argued that uniform scaling is a very crude approximation and proposed that the scale factor be made a function of *both formant number and vowel category.* This is referred to as non-uniform scaling. With this approach, Fant claims to reduce the female–male variance to one-half of that remaining after simple uniform-scaling-based normalization suggested by Nordström & Lindblom (1975). Nearey (1978) extensively studied the validity of uniform-scaling model and concluded that there may be some systematic speaker-dependent variation supporting some of Fant's observations. However, his efforts to find a better additive transform than log transform (which corresponds to uniform scaling) using generalized linear models did not yield any alternate scale (Nearey 1992).

Fant's non-uniform scaling requires each speaker to have a separate scale-factor for each formant frequency of each vowel category. In other words, if we have 10-vowel categories and each vowel is characterized by three formants, we have 30 scale factors. Fant (1975) uses the scale factor expressed in percentage in his studies, i.e. $\alpha_{\mathrm{s}} = (1 + \frac{\Gamma_{\mathrm{s}}}{100})$, and hence the same notation is followed in this discussion. Therefore, the scale factor of the j^{th} vowel and n^{th} formant for a subject speaker, s, is given by $A^j_{n,\mathcal{F}s} = \left(1 + \frac{\Gamma^j_{n,\mathrm{s}}}{100}\right)$, where

$$\Gamma^j_{n,\mathrm{s}} = \left(\frac{\Gamma^j_{\mathrm{n},\mathcal{FM}}}{k_{\mathcal{FM}}}\right) \underbrace{k_{\mathrm{s}}}_{\text{speaker-dependent}} \tag{1}$$

$\left(\frac{\Gamma^j_{n,\mathcal{FM}}}{k_{\mathcal{FM}}}\right)$ is obtained from vowel data and is *independent of any speaker*. The procedure to estimate $\left(\frac{\Gamma^j_{n,\mathcal{FM}}}{k_{\mathcal{FM}}}\right)$ is described in Fant (1975) and basically corresponds to scale factors between *average male* and *average female* of the database. $k_{\mathcal{FM}}$ is a constant that depends on the average male and average female of the database and is found to be -12.18 for Hillenbrand database (see Fant (1975) for more details). The speaker-dependence comes from the term k_{s} which is estimated from formant frequencies of that speaker. It should be noted that in Fant's normalization scheme, one needs to know the vowel and formant number which may not be practically possible in many real-world applications.

Recently, Umesh *et al* (2002c), have proposed the idea of frequency-dependent scale factor. The basic idea is to model the vowel-category and formant-number-specific scale factor, $\Gamma^j_{n,\mathcal{FM}}$ as a function of frequency. This is reasonable since $\Gamma^j_{n,\mathcal{FM}}$ is associated with n^{th} formant-frequency (in Hz) of j^{th} vowel, and therefore, we can plot $\Gamma^j_{n,\mathcal{FM}}$ as a function of frequency as shown in figure 2. It should be noted that there is a strong correlation between the value of the formant frequency and the value of the scale factor, i.e., lower the formant frequency the larger the deviation from $k_{\mathcal{FM}}$ $(= -12.18)$. The formant frequencies from different speakers are averaged over small frequency bands to obtain a frequency-dependent scale function $\gamma(f)$

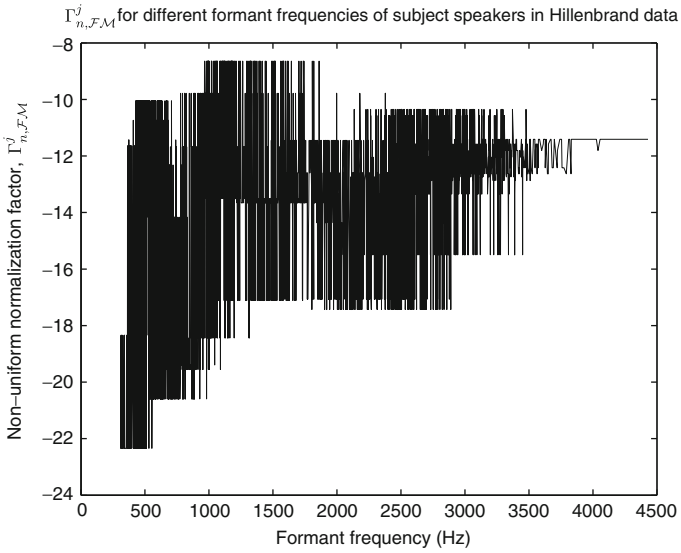


Figure 2. The non-uniform scale factors, $\Gamma^j_{n,\mathcal{F}s}$, plotted as a function of the corresponding formant frequency for different speakers in the Hillenbrand database. It should be noted that there is a strong correlation between the value of the formant frequency and the value of the scale-factor ($\Gamma^j_{n,\mathcal{FM}}$), i.e., lower the formant frequency, the larger the deviation of $\Gamma^j_{n,\mathcal{FM}}$ from $k_{\mathcal{FM}}(= -12.18)$.

which is independent of the speaker as shown in figure 3. The proposed normalization scheme is given by

$$\rho(k_{\mathrm{s}}, f) = \gamma(f).\left(\frac{k_{\mathrm{s}}}{k_{\mathcal{FM}}}\right) = \left(\frac{\gamma(f)}{k_{\mathcal{FM}}}\right).\underbrace{k_{\mathrm{s}}}_{\text{speaker-dependent}} \tag{2}$$

where k_{s} is a speaker-specific scale-factor (independent of formant-frequency, vowel category, etc.) and $\gamma(f)$ if the frequency-dependent weighting that is applied on it. $k_{\mathcal{FM}}$ is a constant that is found to be -12.18 for Hillenbrand database (see Fant (1975) for more details). $\rho(k_{\mathrm{s}}, f)$ is the speaker and frequency-dependent scaling function, but is independent of vowel category and

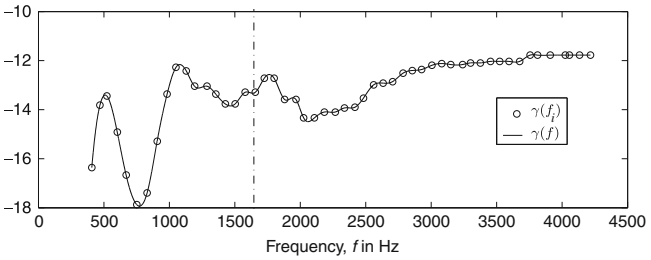
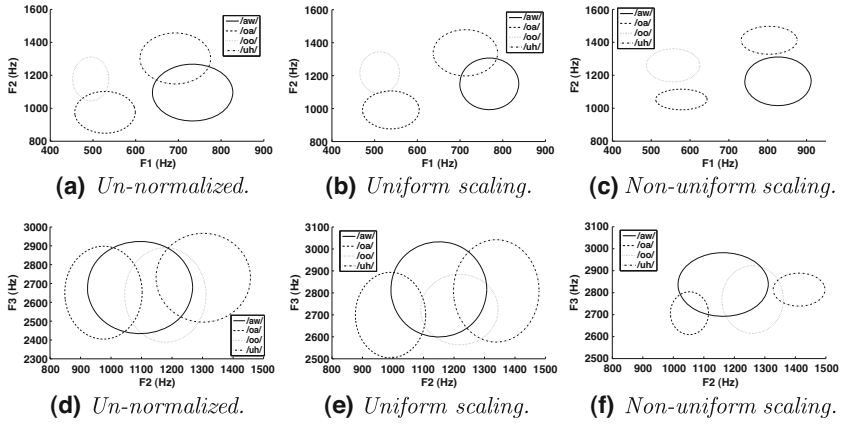


Figure 3. Frequency-dependent scale factors, $\gamma(f_i)$ and frequency-dependent scaling function, $\gamma(f)$ for Hillenbrand databases. It might be easier to understand the weighting function if the reader considers $\frac{\gamma(f)}{k_{\mathcal{FM}}}$, where $k_{\mathcal{FM}} = -12.18$ for Hillenbrand databases.



(a) *Un-normalized.* **(b)** *Uniform scaling.* **(c)** *Non-uniform scaling.*

(d) *Un-normalized.* **(e)** *Uniform scaling.* **(f)** *Non-uniform scaling.*

Figure 4. $F_1 - F_2$ and $F_2 - F_3$ normalization of vowels /aw/, /oa/, /oo/ and /uh/ in Hillenbrand data.

formant number. Although, $\gamma(f)$ is used for most part of the discussion, it might be easier to understand the weighting function in figure 3 if the reader considers $\frac{\gamma(f)}{k_{\mathcal{FM}}}$. The function $\frac{\gamma(f)}{k_{\mathcal{FM}}}$ weights the speaker-specific scale factor k_s differently at different frequencies and hence the overall warp-factor varies as a function of frequency. As seen in figure 3, $\frac{\gamma(f)}{k_{\mathcal{FM}}}$ tends to unity for $f \geq 1600$ Hz. This agrees with the uniform scaling (i.e., constant scale factor) assumption made for higher formants which are mostly affected by the open-vowel scale-factor, k_s. The frequency dependence comes from the *weighting* of $\gamma(f)$ on k_s.

In Harish *et al* (2009), separate frequency-dependent scaling function is found for different vowels. These functions are used to normalize the vowels in Hillenbrand data since the database provides carefully estimated formant frequencies for every vowel utterance. Figure 4 shows $F_1 - F_2$ and $F_2 - F_3$ space normalization for the *back* vowels */aw/*, */oa/*, */oo/* and */uh/* in Hillenbrand data. It is clear from the figure that non-uniform scaling not only reduces the standard deviation of the formants but also improves the separation among vowels that are closely packed in $F_1 - F_2$ and $F_2 - F_3$ planes, as compared to uniform scaling. The improvement is significant especially in the $F_2 - F_3$ plane.

3. Vowel-normalization using universal warping function

In this section, we review our previous work (Umesh *et al* 2002a,b), where we have approached the problem of vowel normalization through the concept of *universal* warping function. The basic idea of universal-warping function is to find a warping function of the frequency-axis that maps the physical frequency, f, to an alternate domain, ν, such that, in the alternate domain speaker-dependent parameter separates out as a pure translation factor. It is 'universal' in the sense that the same mapping should be applicable to *all speech data* irrespective of the speaker. Throughout this paper, when we talk of spectra (i.e., functions of frequency), we use f to denote the frequency variable, and when we talk of formants (i.e., specific frequency values) we use F_i to denote the i^{th} formant frequency. The 'universal'-warping approach assumes that the speaker dependencies can be modelled through a single translation factor. For example, the commonly

assumed uniform scaling model where all formant frequencies (F_i) between speakers r and s are scaled by one constant α_{rs} is given by

$$F_{i,r} = \alpha_{rs} F_{i,s}. \tag{3}$$

In this case, α_{rs} is the speaker-dependent parameter that relates speaker s with the reference speaker r. This can be equivalently expressed in the log-warped domain as

$$\log\left(F_{i,r}\right) = \log\left(\alpha_{rs}\right) + \log\left(F_{i,s}\right). \tag{4}$$

For the uniform scaling model of Eq. 3, we see from Eq. 4 that the universal-warping function is the log-warping, i.e., $\Lambda_i = \log(F_i)$ and in this case, the speaker-dependent scale factor separates out as a translation factor in the log-warped domain, i.e.,

$$\Lambda_{i,r} = \tau_{rs} + \Lambda_{i,s}. \tag{5}$$

Equivalently, if we take the ratio of formants (say i^{th} and j^{th}) for the same speaker, we have

$$\begin{aligned}\log\left(\frac{F_{i,r}}{F_{j,r}}\right) &= \log\left(F_{i,r}\right) - \log\left(F_{j,r}\right) = \log\left(\frac{\alpha_{rs} F_{i,s}}{\alpha_{rs} F_{j,s}}\right) \\ &= \log\left(F_{i,s}\right) - \log\left(F_{j,s}\right).\end{aligned} \tag{6}$$

Hence, Miller's approach to normalization, which is similar to the above equation (except for $\frac{F_1}{SR}$) is equivalent to uniform scaling. Nearey's approach is also a variation of the uniform scaling model, with μ_L corresponding to speaker-dependent shift-factor. We recently became aware of the work of Nearey (1978), where he introduces the use of log-additive hypothesis. Our concept of universal-warping function is similar to this concept, except that we are looking in a generalized framework of translation for any transformation and not necessarily log transformation. It should be noted that since the shift factor does not depend on any specific phoneme, and only on the speaker, it is often referred to as *extrinsic* normalization factor. Extending the model in Eq. 3 to spectral envelopes, we can assume that the spectral envelopes of r and s are scaled versions of one another, i.e., $P_r(f) = P_s(\alpha_{rs} f)$. In the case of spectral envelopes, log-warping the frequency-axis, i.e., $\lambda = \log(f)$ results in

$$_r(\lambda) = P_r(e^{\lambda}) = P_s(e^{\lambda + \log \alpha_{rs}}) = \ _s(\lambda + \log \alpha_{rs}), \tag{7}$$

where $_r$ and $_s$ are log-warped versions of P_r and P_s, respectively. Therefore, the frequency-warped spectral envelopes are *shifted* versions of each other, if the model in Eq. 3 is indeed true. As discussed in the previous section, from experiments of Fant and some of our previous experiments by Umesh *et al* (2002a,b,c), it has been observed that there are deviations from the uniform scaling model. Since there are deviations from the uniform scaling model, log-warping is *not* the appropriate universal-warping function to separate the speaker-dependent parameter. In Umesh *et al* (2002b,c), a piece-wise approximation to the universal warping function was found empirically from speech data, such that in the universal-warped domain, the same sound enunciated by different speakers were a translated version of one another. This empirically obtained universal-warping function is referred to as the *speech scale*. Interestingly, the speech scale was found to be 'very similar' to the Mel-scale which is briefly described below.

Stevens & Volkman (1940) experimentally obtained a non-linear mapping between perceived and physical frequency of a tone and referred to it as the Mel-scale. In their original work, Stevens and Volkman had experimentally obtained the mapping at discrete set of points, for

which various closed-form curves have been fitted by researchers. The widely accepted closed-form approximations to Mel-scale have the functional form

$$\eta = a \log_{10}\left(1 + \frac{f}{b}\right), \tag{8}$$

where f is in Hz and η is in Mels. Fant's technical Mel-formula is defined with $a = \frac{1000}{\log 2}$ and $b = 1000$, whereas in speech recognition the widely used formula is defined with $a = 2595$ and $b = 700$, i.e.

$$\eta_{\text{mel}} = \Theta(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{9}$$

Although in speech recognition, the Mel scale is the most commonly used psychoacoustic scale, in many other areas of speech, the Bark scale or Equivalent Rectangular Bandwidth (ERB) scale is usually used. It should be noted that Bark and ERB scales also have functional form similar to Eq. 8. Therefore, for the purposes of this paper, we will only compare with the Mel scale.

It should be noted that since Bark scale is similar to Mel scale (and hence speech scale), the normalization method of Bladon *et al* (1983) also uses a similar idea which involves shifting down the auditory spectrum produced by women by about 1 Bark. Although Bladon *et al* (1983) uses gender-specific shift, every speaker has a speaker-specific relative shift in the speech scale in the approach of Umesh *et al* (2002b,c). Further, this normalization approach is very similar to that proposed by Syrdal & Gopal (1983). It should be noted that the empirical speech scale, the Mel scale and the Bark scale are very similar, and hence, we have the following normalization.

$$\mathcal{S}\left(F_{i,\mathtt{r}}\right) - \mathcal{S}\left(F_{i,\mathtt{s}}\right) \approx \mathsf{Mel}\left(F_{i,\mathtt{r}}\right) - \mathsf{Mel}\left(F_{i,\mathtt{s}}\right) = c_{\text{rs}}, \tag{10}$$

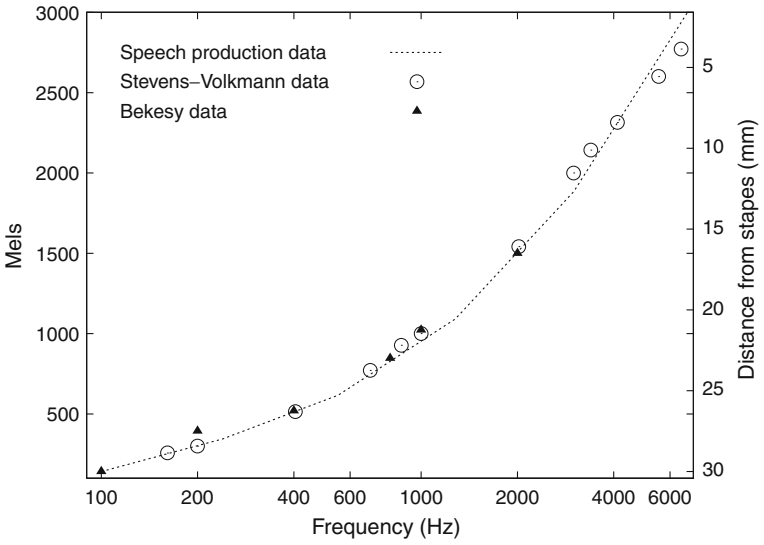


Figure 5. The figure shows the speech scale, Stevens and Volkmann data and Békésy's data. The speech scale has been obtained empirically from actual speech data. The Steven and Volkmann data have been obtained from psycho-physiological study, while the Békésy data has been obtained from experiments on the basilar membrane. The fact that all the three curves are similar shows a strong connection between speech and hearing.

where $\mathcal{S}(.)$ is the speech scale and c_{rs} is a speaker-dependent constant which is independent of i, i.e., formant number.

The Mel scale was obtained by Stevens & Volkman (1940) based on *perceptual studies*. We have obtained a speech scale which has been purely estimated *only from speech data* with the primary purpose to show the commonality *between* speakers for the same utterance. In addition, there is another scale that is of relevance which we now discuss. It is the frequency to physical place mapping on the basilar membrane which is based entirely on physical aspects of the basilar membrane. This place map, representing the behaviour of the basilar membrane as a function of frequency was discovered by von Békésy in his studies, for which he won a Nobel prize (von Békésy & Rosenblith 1951). In his experiments, the stapes was vibrated with a constant amplitude sinusoid, and the frequency response at various points along the basilar membrane was examined under the microscope. The mapping relating the frequency of the stimulation to the position of maximum response on the basilar membrane was thus established.

We now compare the speech scale, the Mel scale and the place map of von Békésy. The results are shown in figure 5. The fact that the three, independent, experimentally derived scales, namely the place map, the Mel scale and the speech scale, are so similar is an indication that they may be the basic fundamental experimental linking of the speech-hearing connection.

4. Relationship between spectra of different speakers enunciating the same sound

As discussed in the previous section, for the uniform scaling model of $\mathrm{P_r}(f) = \mathrm{P_s}(\alpha_{\mathrm{rs}} f)$, the universal warping function corresponds to log-warping the frequency-axis, i.e. $\lambda = \log(f)$. This results in the speaker-dependent factor separating out as translation term in the λ domain as shown below.

$$_{\mathrm{r}}(\lambda) = \mathrm{P_r}(e^{\lambda}) = \mathrm{P_s}(e^{\lambda + \log \alpha_{\mathrm{rs}}}) = \ _{\mathrm{s}}(\lambda + \log \alpha_{\mathrm{rs}}). \tag{11}$$

However, empirical experiments in the previous section using actual speech data indicate that speech scale is a more appropriate universal warping function, and that the speech scale is very similar to the perceptually obtained Mel scale. If the speech scale is indeed the universal-warping function, then what is the corresponding relation between spectra of speakers enunciating the same sound?

Kumar & Umesh (2008) proposed the following affine model to describe the relation between the formant frequencies of subject and reference speakers, i.e.,

$$(F_{\mathrm{r}} + \kappa) = \alpha_{\mathrm{rs}} \, (F_{\mathrm{s}} + \kappa) \, , \tag{12}$$

where F_{r}, F_{s} are the formant frequencies of reference speaker, r and subject speaker, s respectively. α_{rs} is a speaker-dependent parameter relating the reference and subject speakers. κ is a constant in the model and is *not* dependent on the speakers. In Kumar & Umesh (2008), we show that the universal warping function corresponding to the model of Eq. 12 matches the speech scale. Equivalently, we can also re-write Eq. 12 as

$$F_{\mathrm{r}} = \alpha_{\mathrm{rs}} F_{\mathrm{s}} + \kappa \, (\alpha_{\mathrm{rs}} - 1) \, . \tag{13}$$

We can clearly see the affine relation in the above equation, and hence we refer to the proposed model as the affine-model. It should be noted that unlike conventional affine equation, the shift factor is *also* a function of the scaling factor α_{rs}.

The value of κ_{mean} has been empirically obtained to be 508.04 for Peterson & Barney (PnB) database and 495.67 for Hillenbrand (HiL) database (figure 6). The corresponding universal-warping functions ν_{pnb} and ν_{hil} for PnB and HiL data, respectively are given by

$$\nu_{\text{pnb}} = \log\left(1 + \frac{f}{508.04}\right) \tag{14a}$$

$$\nu_{\text{hil}} = \log\left(1 + \frac{f}{495.67}\right). \tag{14b}$$

The fact that these are again close to the usually accepted form of Mel-scale is remarkable, indicating a strong connection between speech production and hearing.

As we mentioned previously, Nearey (1978) also investigated the use of a transformation of the type $\log(f + b)$ which is similar to our affine model. The main difference between our approach and his (Nearey 1978) is that he considers only the *average* formant values for males, females and children for several languages including the American English database of PnB. On the other hand, we have considered *pair-wise* all the speakers in the PnB and HiL databases. We have then averaged the estimates. Further, unlike Nearey (1978), we have considered *all* the formants and not just first or second formant. One of the reasons Nearey was motivated to use the transform of the type $\log(f + b)$ was the systematic speaker-dependent variation in acoustic parameters. However, his analysis using first and second formants in Nearey (1978, 1992) did not show any substantial improvement by using other frequency-scales over the log-scale. However, we have shown that there is improvement in speech recognition performance using our proposed affine model (or equivalently shift in Mel scale) when compared to the uniform-scaling model (Umesh & Sinha 2007; Kumar & Umesh 2008).

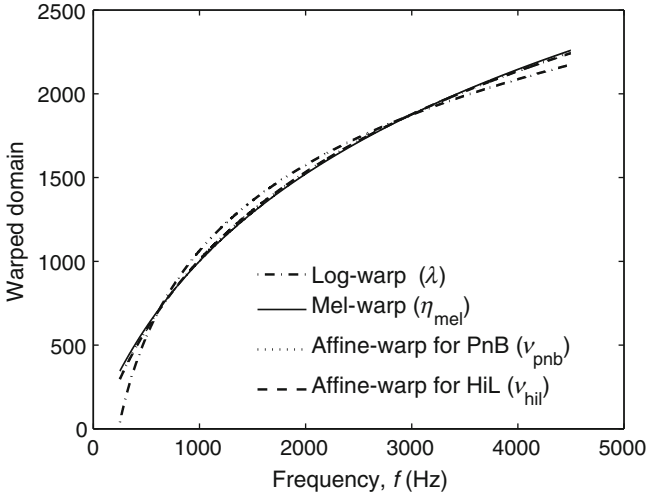


Figure 6. Comparison of different *universal* warping functions including the affine-warp, the log-warp and the Mel-warp functions. The affine-warp functions for Peterson & Barney (PnB) and Hillenbrand (HiL) databases almost overlap, since they are functionally similar as seen from Eq. 14.

5. VTLN in ASR

In the next few sections, we will discuss methods used to reduce inter-speaker variability to improve the performance of ASR systems. If we model the vocal-tract as a uniform tube, then the formant frequencies are inversely proportional to the length of the vocal tract and the n^{th} formant frequency is given by

$$F_n = \frac{(2n+1)c}{4L}, \tag{15}$$

where c is the speed of sound and L is the length of the vocal-tract. Therefore, for two speakers with different vocal-tract lengths, L_{A} and L_{B}, their formants are related through a uniform-scaling factor given by the *ratio* of the vocal-tract lengths. Sondhi (1986) has shown that this relation is true even for a *bent* tube with an uniform cross-section. In practice, when different sounds are produced, the cross-section of the vocal-tract is non-uniform, and is often modelled as a concatenation of many uniform tubes with different cross-sectional areas. However, even in such cases, it has been empirically observed that there is an inverse relationship with the overall length especially for the third, fourth and higher formants. Since the vocal tract length can vary from approximately 13 cm for adult females to over 18 cm for adult males, formant centre frequencies can vary by as much as 25% between speakers. Figure 7 shows the smoothed-spectra of vowel /eh/ spoken by a male and female speaker. As seen from the figure, there is considerable variation between the two spectra for the same perceived sound. This source of variability results in a significant degradation from speaker dependent to speaker independent speech recognition performance. The main objective of VTLN is to find an optimal warping factor to warp the frequency axis of the speech signal so that variations in formant frequencies of speakers enunciating the same sound is reduced. This in turn reduces the variability of features among different speakers enunciating the same sound greatly improving the recognition performance.

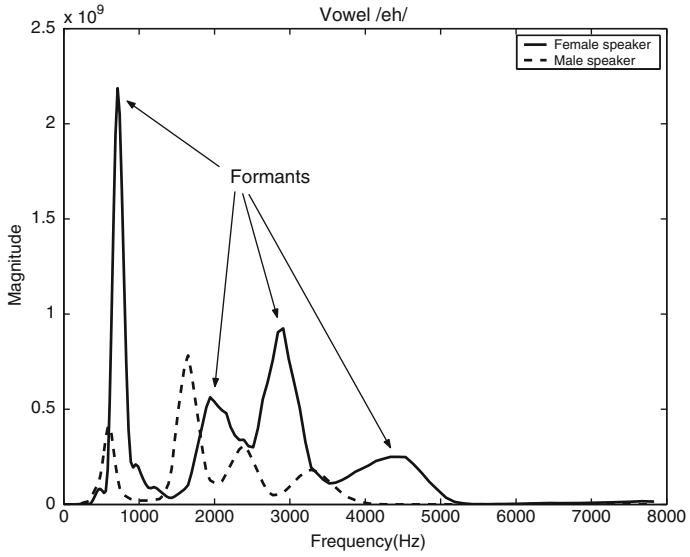


Figure 7. This figure shows the variability in the spectra of a male and female speaker for the same sound /eh/ spoken by them.

Based on the previous discussion, the relation between spectra of two speakers enunciating the same sound is assumed to be uniform scaling, i.e.,

$$\mathrm{P_r}(f) = \mathrm{P_s}(\alpha_{\mathrm{rs}} f). \tag{16}$$

Since in practice, there exists no ideal or reference speaker, the warp factor for each speaker is found using a maximum-likelihood framework as discussed in the next section.

5.1 *Maximum likelihood based approach for warp-factor estimation*

In ASR, a maximum-likelihood (ML)-based method for warp-factor estimation was proposed and investigated in e.g., Kamm *et al* (1994), Wegmann *et al* (1996) and Lee & Rose (1998). The ML-based approach, however, does not perform VTLN in the strict sense. The warping factor is estimated to increase the matching score of the acoustic model and it may not exactly reflect the difference in VTLs. Most state-of-the-art ASR systems use this approach.

The ML-based grid search over a discrete set of αs is performed with respect to the acoustic model – usually a set of hidden Markov models (HMMs). The range of α is usually restricted to be between 0.80 and 1.20 based on physiological arguments that the the vocal tract lengths can vary only so much and an increment of 0.02 is usually followed in practice.

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

$$\hat{\alpha_i} = \arg\max_{\alpha} \Pr(\mathbf{C}_i^{\alpha} | \lambda, W_i), \tag{17}$$

where $\mathbf{C}_i^{\alpha}$ represents the warped features for the i^{th} utterance, λ being the HMM model and W_i is the true transcription during training or first pass recognition during testing. The use of transcription allows us to get the sequence of HMM models, and the optimal α is found to be the one that provides best *alignment* between the warped-features and the HMM models. The steps to calculate warped Mel frequency cepstral coefficients (MFCC) features $\mathbf{C}_i^{\alpha}$ are described below.

Conventional MFCC feature extraction is usually implemented as shown in figure 8. Let $\mathbf{P}$ represent the power or magnitude spectrum of a frame of speech. Let $\mathbf{F_m}$ represent the filter-bank smoothing operation along with Mel warping, which can be represented through a linear-transformation matrix. Further, let $\mathbf{D}$ represent the DCT transformation which is also linear. The MFCC features, $\mathbf{C}$, are obtained by applying the Mel-warped filter-bank on the power spectrum of the speech signal, followed by applying logarithm on the output of the filter-bank and finally a DCT transformation. This can be written mathematically as:

$$\mathbf{C} = \mathbf{D}[\log(\mathbf{F_m}.\mathbf{P})]. \tag{18}$$

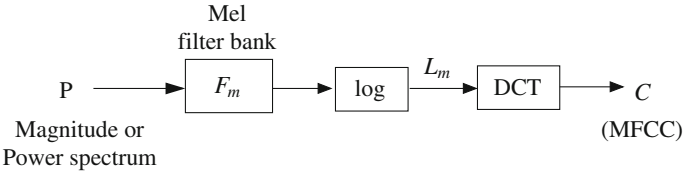


Figure 8. Steps involved in generating conventional MFCC features.

As an illustration, let s be a speech frame consisting of 320 samples. A 512-point DFT is applied to obtain the 256-dimensional vector $\mathbf{P}$ whose elements are the magnitude of the DFT coefficients for one half of the spectrum. This is because the magnitude spectrum has even symmetry. If a 24-filter Mel filter-bank smoothing is applied, then $\mathbf{F_m}$ is a 24×256 matrix that operates on $\mathbf{P}$ to obtain the Mel-warped smoothed spectra. $\mathbf{D}$ is the 24×24 DCT-matrix applied on log-compressed Mel-warped smoothed spectra to obtain the MFCC feature vector $\mathbf{C}$. In practice, only the first 13 cepstral coefficients are used and one may use a 13×24 DCT transformation.

VTLN-warped features were obtained in the original method of Andreou *et al* (1994), by frequency-warping the magnitude spectra $\mathbf{P}$ to get $\mathbf{P}^{\alpha}$ 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 & Rose (1998) proposed an efficient alternate implementation, where the Mel filter-bank is inverse scaled for each α, while the signal spectra is left unchanged as shown in figure 9. 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 $\mathbf{F}_{\mathbf{m}}^{\alpha}$ 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 centre frequencies and bandwidths of the Mel filter bank are appropriately scaled to obtain Mel- and VTLN-warped smoothed spectra (Lee & Rose 1998). The change in the filter bank structure for different warping factors is illustrated in figure 10. The slope in the last filter has been modified appropriately using piece-wise linear warping (Wegmann *et al* 1996) so that the Nyquist frequency maps onto itself after frequency scaling. This avoids the bandwidth mismatch that arises due to frequency warping. The warped cepstral features $\mathbf{C}^{\alpha}$ are given by

$$\mathbf{C}^{\alpha} = \mathbf{D}[\log(\mathbf{F}_{\mathbf{m}}^{\alpha}.\mathbf{P})], \tag{19}$$

which are obtained by first warping and smoothing the power spectrum, followed by log and the DCT operations. The filter bank is embedded with *both* Mel and VTLN warping, to perform smoothing as well as scaling of the spectrum. For the case of $\alpha = 1.00$, $\mathbf{C}^{\alpha}$ *exactly* corresponds to the case of conventional MFCC *without* VTLN warping.

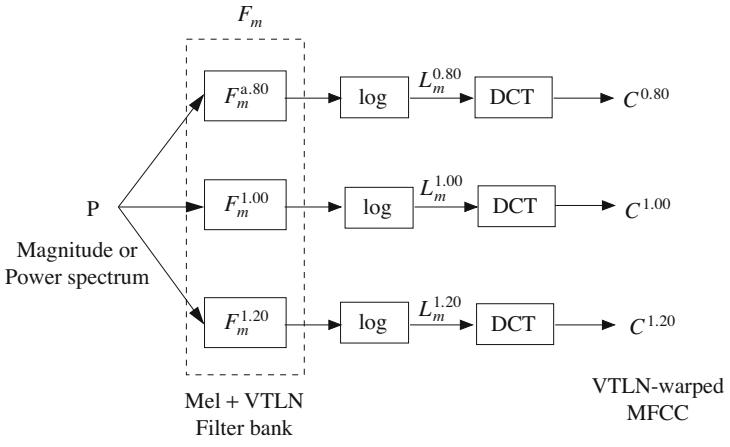


Figure 9. Conventional framework for generating warped features in VTLN. The filter bank is inversely scaled instead of resampling the speech signal for each warp factor for efficient implementation.

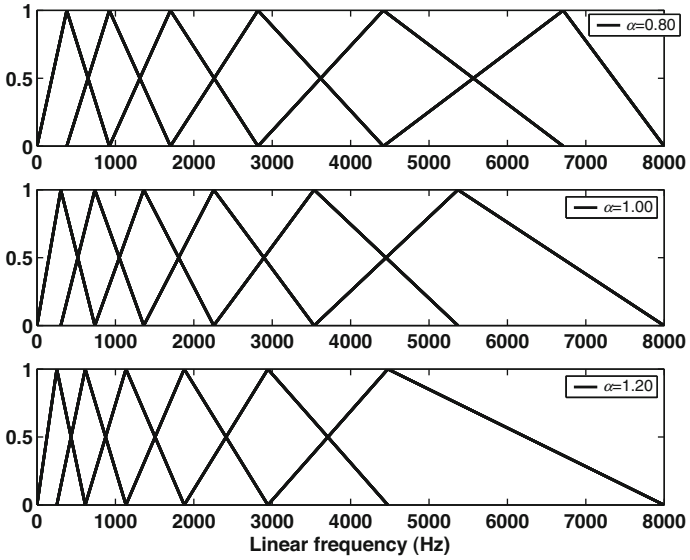


Figure 10. Illustrating the change in the filter-bank structure with VTLN warping in linear frequency (Hz) domain. The filters have non-uniform centre frequencies with non-uniform bandwidths.

The 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 figure 10. The filter bank incorporates both Mel and VTLN warping for efficient implementation (Lee & Rose 1998). The entire signal processing step needs 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 figure 9, where, **P** represents the power or magnitude spectrum, F_m the filter bank scaling.
- Estimate the optimal warp factor $\hat{\alpha}$ using Eq. 17, which requires all the warped features to be generated beforehand. The features corresponding to the optimal warp factor are taken as normalized features.
- Using the normalized features obtained above, the model parameters are updated in the case of training or used for recognition during testing.

The major disadvantage of ML search is that it is very computationally expensive, since one needs to carry out the alignment for all possible warping factors and select the one that has the best matching score. Although, there have been many attempts to improve it by using Brent search, etc., nevertheless, the computational costs are high. Therefore, there have been attempts to find a linear transformation between warped and unwarped cepstra, so that computational costs can be reduced.

6. Linear transform approaches to VTLN

This section deals with approaches for performing VTLN which involves *only* a linear transformation (LT) on conventional MFCC to obtain VTLN-warped features, i.e.,

$$C^{\alpha} = A^{\alpha}.C, \tag{20}$$

where C^{α} represent the VTLN-warped features and C represent the conventional MFCC features and A^{α} is the LT matrix. Obtaining such a relation eliminates the need to change the filter bank structure for every warping factor of interest and VTLN-warped features can be directly obtained using a matrix transformation.

From Eqs 18 and 19, the relation between $\mathbf{C}^{\alpha}$ and $\mathbf{C}$ is given as:

$$\mathbf{C}^{\alpha} = \mathbf{D}[\log .\mathbf{F}_{\mathbf{m}}^{\alpha}.\{\mathbf{F}_{\mathbf{m}}^{-\mathbf{1}}.\exp .\mathbf{D}^{-\mathbf{1}}.(\mathbf{C}^{1.00})\}]. \tag{21}$$

A LT between $\mathbf{C}^{\alpha}$ and $\mathbf{C}^{1.00}$ (or $\mathbf{C}$) can be derived if all the intermediate operations can be represented as linear operations. However, from Eq. 21, it is evident that log is a non-linear operation and in practice $\mathbf{F}_{\mathbf{m}}^{-\mathbf{1}}$ *does not* exist. This is because, the power-spectrum $\mathbf{P}$ *cannot* be completely reconstructed from the filter-bank outputs because of the smoothing operation (Claes *et al* 1998). The knowledge of $\mathbf{P}$ is important, since conventional VTLN warping relations are always specified in the linear frequency (Hz) domain, usually through a mathematical relation of the type $\tilde{f} = g_{\alpha}(f)$, where $\tilde{f}$ is the warped frequency and $g_{\alpha}(f)$ is the frequency-warping function. Therefore, in this case, it is not possible to recover $\mathbf{P}$ from the filter-bank output and hence a LT is not possible.

The main aim for obtaining a LT for VTLN warping is to directly transform $\mathbf{C}$ to $\mathbf{C}^{\alpha}$ rather than rescaling the filter-bank for each α to obtain the warped cepstral features. This will not only provide computational advantage over the conventional approach during VTLN feature generation but will also help in accounting for the Jacobian of the transformation which is discussed later in this paper. Before proceeding further, a brief review of existing LT approaches is presented.

6.1 *Review of existing approaches to obtain LT*

One of the early attempts to obtain a LT was by Acero *et al* (Acero 1990; Acero & Stern 1991). They proposed the use of bilinear warping for achieving variable frequency warping for speaker normalization. Formulating the bilinear transform as a linear filtering operation and having the time reversed cepstrum sequence as an input, they have shown that the warped cepstral coefficients can be obtained at the outputs of the filters at time $n = 0$. They use the complex cepstrum instead of the real cepstrum because the sequence $c[n]$ should be causal (Oppenheim & Johnson 1972). They showed that each warped cepstral coefficient can be represented as a linear combination of the unwarped cepstral coefficients.

Motivated by the work of Acero (1990) and Acero & Stern (1991) and based on the observation that frequency warping functions used in most VTLN methods can be approximated to a reasonable degree by the bilinear transform, McDonough *et al* (1998) suggested the use of conformal maps such as bilinear transform and its generalizations for speaker normalization. Since the unit circle is mapped back onto the unit circle, McDonough refers to these conformal maps as all-pass systems; such systems have uniform frequency response and thus pass signals of all frequencies with neither attenuation nor amplification. The all-pass transforms are classified as rational and sine-log all-pass transforms (RAPT and SLAPT).

Claes *et al* (1998) proposed an approach to transform the HMM-based acoustic models trained for a certain group of speakers (say adults) for use on speech from different group of speakers (say children). The transformations are generated to linearly warp the spectral characteristics of the features from a speaker and the required warping factors are estimated based on the average third formant F_3. As MFCC features were used which also involved a additional nonlinear mapping, a linear approximation for the exact mapping was computed by *locally linearizing* it. It

was also reported that the linear approximation was accurate as long as warping was reasonably small. Although, they did compute LTs in MFCC feature domain for VTLN, they transformed the parameters of the HMM model instead of transforming the acoustic observation vectors and the required warp factor was determined through formant estimation.

Cui & Alwan (2006) have proposed a modification to the approach of Claes *et al* (1998), where they derived a mapping matrix to align formant-like peaks. The peaks are estimated by Gaussian mixtures using the Expectation–Maximization algorithm (Zolfaghari & Robinson 1996).

6.2 *LT approach of Pitz* et al

The LT proposed by Pitz *et al* (2001), Molau *et al* (2001) and Pitz (2005) can be used for deriving a LT for any arbitrary invertible warping function unlike McDonough *et al* (1998). All the derivations were made using the modified signal processing approach proposed in Molau *et al* (2001), where the Mel and VTLN warping are integrated into the DCT transformation matrix and does not have filter-bank smoothing. The methods proposed by McDonough, Pitz, etc. do the signal processing assuming *continuous-frequency spectra* and further do not use the conventional DCT-II for decorrelating the features.

Umesh *et al* (2005) showed that sampling the continuous-frequency spectra would introduce aliasing in the above methods and therefore lead to degradation of performance. They argued that it is important to have filter-bank smoothing to ensure the approximate que-frency limitedness of cepstra and further showed that by using bandlimited interpolation, the linear-transformation matrix can be obtained directly in the discrete domain.

Motivated by the LT approach of Umesh *et al* (2005), Panchapagesan (2006), Panchapagesan & Alwan (2009) and Sanand *et al* (2007) have attempted to derive a LT on conventional MFCC features. However, Panchapagesan assumes linear warping relation to be true in the Mel-domain which does not correspond to conventional VTLN warping. The method of Sanand *et al* (2007) which is exactly equivalent to conventional VTLN warping is described next.

6.3 *LT of conventional MFCC features*

The approaches described above do not derive a LT of conventional MFCC, and in some cases also require changes in signal processing. Sanand *et al* (2007) and Sanand & Umesh (2008) argue that separating the VTLN-warping from the Mel filter bank helps in deriving a LT between, warped and unwarped cepstral features *within* the conventional MFCC framework.

Let $\mathbf{L_m} = \log(\mathbf{F_m}.\mathbf{P})$ be the conventional log-compressed Mel-warped filter-bank outputs. From Eq. 18, the knowledge of $\mathbf{C}$ implies the knowledge of $\mathbf{L_m}$ as they form a DCT pair, i.e.,

$$\mathbf{L_m} = \mathbf{D^{-1}}.\mathbf{C}. \tag{22}$$

However, $\mathbf{P}$ cannot be recovered from $\mathbf{L_m}$ because of the filter-bank smoothing operation. The knowledge of $\mathbf{P}$ is required in the conventional approach to obtain the VTLN-warped spectrum, which in turn is required for generating VTLN-warped cepstral features, $\mathbf{C}^{\alpha}$. Since $\mathbf{P}$ cannot be recovered, the problem is reframed as follows: can ${\mathbf{L_m}}^{\alpha}$ $(= \log(\mathbf{F}_{\mathbf{m}}^{\alpha}.\mathbf{P})$ be obtained by applying a LT on $\mathbf{L_m}$ without recovering $\mathbf{P}$, i.e.,

$$\mathbf{L}_{\mathbf{m}}^{\alpha} = \hat{\mathbf{T}}^{\alpha}.\mathbf{L_m} = \hat{\mathbf{T}}^{\alpha}.\log(\mathbf{F_m}.\mathbf{P}), \tag{23}$$

where $\hat{\mathbf{T}}^{\alpha}$ is the transformation that is applied on $\mathbf{L_m}$ to obtain $\mathbf{L}^{\alpha}_{\mathbf{m}}$. The above equation states that the filter bank performs only Mel warping and the the transformation $\hat{\mathbf{T}}^{\alpha}$ performs VTLN warping. This means that the VTLN warping embedded in the filter bank for efficient implementation in the conventional approach is now performed separately and is not a part of the filter bank construction. This is illustrated in figure 11.

In discrete implementation, the values of $\mathbf{L_m}$ are known only at N equally spaced points, say $\nu_l = \frac{2\pi l}{N}$, in the Mel frequency domain. N is the number of Mel filters. Now the problem is to obtain $\mathbf{L}^{\alpha}_{\mathbf{m}}$, that is defined at a new set of warped frequencies, say $\tilde{\nu}_l = \frac{2\pi \tilde{l}}{N}$, which may not correspond to any of the discrete frequencies of ν_l. So, the problem may be re-phrased as follows: given the values of $L_m[\nu_l]$, can the values of $L_m[\tilde{\nu}_l]$ be derived. In Sanand & Umesh (2008), they obtain $\mathbf{L}^{\alpha}_{\mathbf{m}}$ through a LT of $\mathbf{L_m}$ vector, i.e.,

$$\mathbf{L}^{\alpha}_{\mathbf{m}} = \hat{\mathbf{T}}^{\alpha}\mathbf{L_m}. \tag{24}$$

If such a relation is obtained, then from Eqs 18 and 19, the relation between $\mathbf{C}$ and $\mathbf{C}^{\alpha}$ is given by:

$$\mathbf{C}^{\alpha} = [\mathbf{D}.\hat{\mathbf{T}}^{\alpha}.\mathbf{D^{-1}}].\mathbf{C}^{1.00}. \tag{25}$$

By defining a LT between $\mathbf{L_m}$ and $\mathbf{L}^{\alpha}_{\mathbf{m}}$, the method completely avoids the inversion of filter bank for obtaining the raw magnitude spectrum $\mathbf{P}$ and also bypasses the log operation. The details of obtaining such a LT is given below.

6.3a *Bandlimited (sinc-) interpolation*: For a bandlimited continuous-time signal, $x(t)$, given uniformly spaced samples of the signal that are appropriately sampled, i.e., $x(t_n)$, the continuous-time signal can be exactly reconstructed. This implies that the values of the time signal at time instants *other* than those at the uniformly spaced samples can be recovered. 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. The conventional Mel-warped smoothed spectral output is obtained in the linear frequency (Hz) domain by applying the Mel-warped triangular averaging filter bank on the linear frequency (Hz) magnitude spectrum. The filter bank has non-uniformly spaced and non-uniform bandwidth filters as shown in figure 10. Alternatively, 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 12.

Therefore, in the Mel frequency domain, the Mel-warped magnitude spectrum can be interpreted as being convolved with a triangle function to obtain the Mel-warped smoothed spectrum

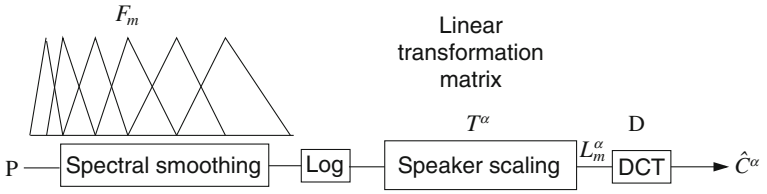


Figure 11. Modification in the signal processing steps (separating the Mel- and VTLN-warping) for realizing a linear transformation. The filter bank performs only Mel warping of the spectra and the proposed bandlimited interpolation matrix performs the VTLN warping.

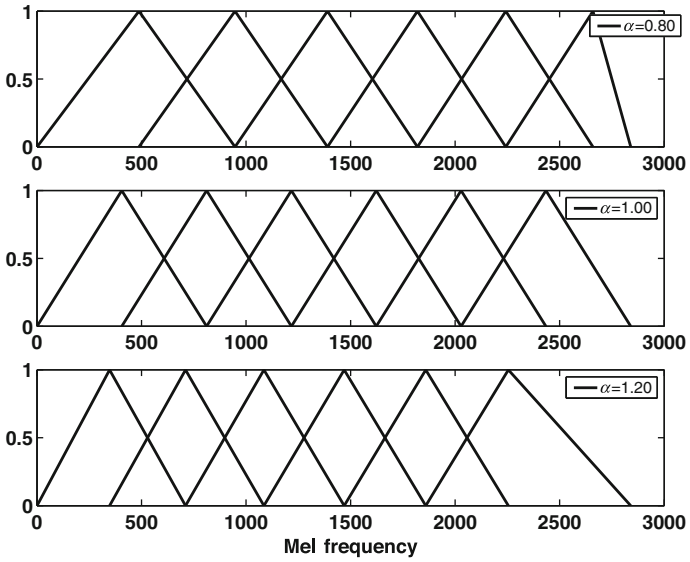


Figure 12. Illustrating the change in the filter-bank structure with VTLN warping in Mel frequency domain. The filters have uniformly spaced centre frequencies with uniform bandwidth for $\alpha = 1.00$. However, they are non-uniformly spaced for α different from unity.

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 $\nu_l = \frac{2\pi l}{N}$, where $l = 0, 1, \cdots, (N-1)$ in the normalized-Mel domain. These uniformly spaced samples correspond to the output of the triangular filters centred at those particular Mel frequencies and exactly correspond to the elements of the vector $\mathbf{L_m}$. Because of the triangle smoothing and subsequent log-operation on the output (which reduces dynamic range), the quefrency content of this smoothed log filter-bank output is only in the low quefrency region. The effect of filter bank smoothing on the cepstral values is shown in figure 13.

During VTLN warping, the filter centre frequencies are appropriately scaled in the linear frequency (Hz) domain by inverse-α as described in Lee & Rose (1998). This corresponds to the centre frequencies of the filter bank to be non-uniformly spaced in the Mel frequency domain as shown in figure 12. 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*, i.e. $L_m[\tilde{\nu}_l]$. These non-uniformly spaced samples exactly correspond to the elements of the vector $\mathbf{L}^{\alpha}_{\mathbf{m}}$.

6.3b *Obtaining the bandlimited interpolation matrix*: The steps to obtain the transformation matrix are as follows:

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

$$\nu_l = 2595 \log_{10}\left(1 + \frac{f_l}{700}\right); \quad \forall l. \tag{26}$$

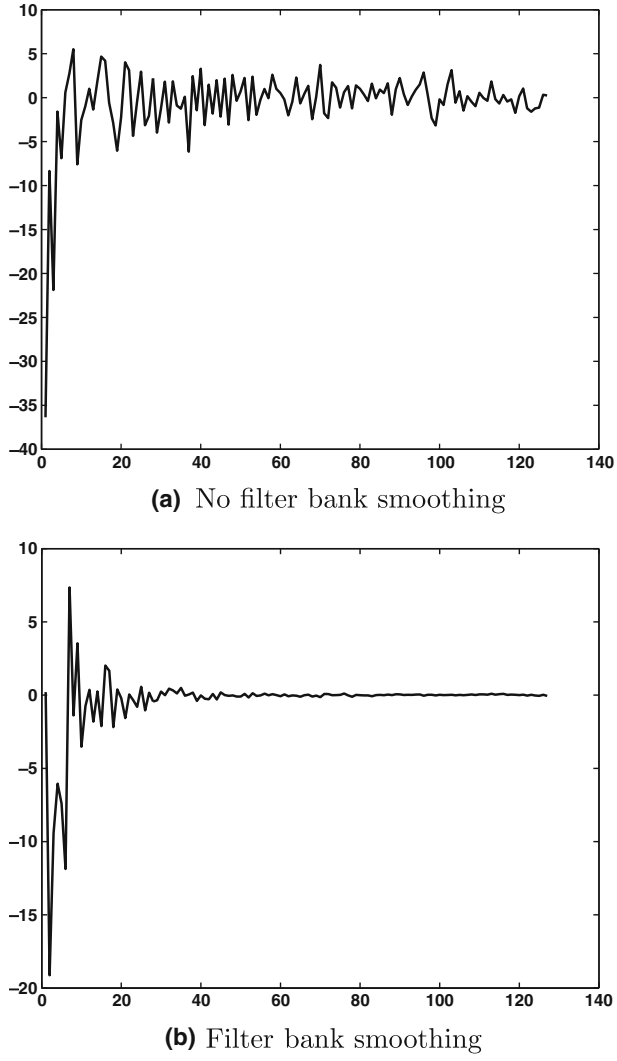


Figure 13. Effect of filter-bank smoothing on the cepstral coefficients. Filter-bank smoothing assures quefrency limitedness.

(ii) During VTLN warping, the warping function $g_\alpha(f)$ is applied to obtain the warped frequencies. Let $\hat{f}_l = \alpha f_l$, denote the linear warping in the linear frequency (Hz) domain. It should be noted that this is *not* linear-warping in the Mel frequency domain. The corresponding centre frequencies of the filters in the Mel domain, i.e., $\hat{\nu}_l$, are related to $\hat{f}_l$ through Eq. 26, with $\hat{\nu}_l = \nu_l$ for $\alpha = 1.00$. This is illustrated in figure 14.

(iii) The Fourier relation between $\mathbf{C_{DTFT}}$ and $\mathbf{L_m}$ is given by:

$$c_k = \frac{1}{2N-2} \sum_{l=0}^{2N-3} L_m \left[\frac{\nu_l}{\nu_s}\right] e^{+j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k}, \tag{27}$$

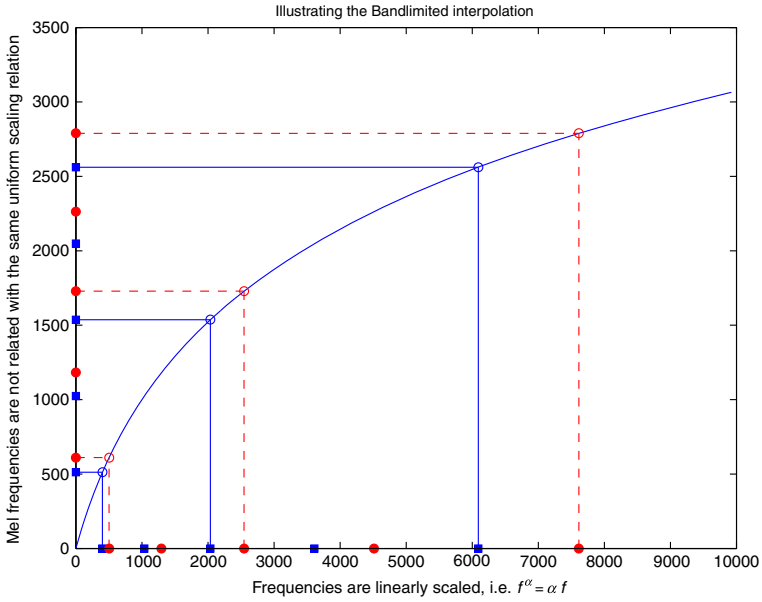


Figure 14. Linear scaling relation is defined in the linear frequency (Hz) domain, i.e. $\tilde{f} = \alpha f$. $\nu_0, \nu_1, \ldots, \nu_{N-1}$ are the outputs of uniformly spaced filter bank corresponding to $\alpha = 1.00$ in the Mel domain. Similarly, $\tilde{\nu}_0, \tilde{\nu}_1, \ldots, \tilde{\nu}_{N-1}$ are the outputs of warped filter bank and are non-uniformly spaced in the Mel domain. The bandlimited interpolation matrix is defined to obtain $\tilde{\nu}_i$ given ν_i. ■ represents the unwarped frequencies both in linear frequency (Hz) and Mel-frequency domains, • represents the warped frequencies in both the domains.

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 $\mathbf{L}_{\mathbf{m}}^{\alpha}$ can be determined as:

$$L_m\left[\frac{\hat{\nu}_l}{\nu_s}\right] = \sum_{k=0}^{2N-3} c_k e^{-j\frac{2\pi}{2N-2}\left(\frac{\hat{\nu}_l}{\nu_s}\right)k}. \tag{28}$$

Substituting Eq. 27 in Eq. 28:

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

The bandlimited interpolation matrix between $L_m[\nu_l]$ and $L_m[\hat{\nu}_l]$ is given by:

$$\mathbf{T}^{\alpha} = \frac{1}{2N-2} \sum_{k=0}^{2N-3} e^{-j\frac{2\pi}{2N-2}(\frac{\hat{\nu}_l}{\nu_s})k} e^{+j\frac{2\pi}{2N-2}(\frac{\nu_l}{\nu_s})k}, \tag{29}$$

where

$$l = 0, \ldots, 2N-3.$$

Using the even-symmetry property, the $N \times N$ interpolation matrix $\hat{\mathbf{T}}^{\alpha}$, i.e., $\mathbf{L}_{\mathbf{m}}^{\alpha} = \hat{\mathbf{T}}^{\alpha}.\mathbf{L}_{\mathbf{m}}$ is given by:

$$\hat{\mathbf{T}}^{\alpha} = \frac{1}{2N-2} \sum_{k=0}^{N-1} 2a_l \cos\left(\frac{2\pi}{2N-2}\left(\frac{\hat{\nu}_l}{\nu_s}\right)k\right)\cos\left(\frac{2\pi}{2N-2}\left(\frac{\nu_l}{\nu_s}\right)k\right), \tag{30}$$

where

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

It is believed that this is the first time an exact LT, without any significant modification in the signal processing of conventional MFCC for VTLN is shown. Although, the transformation here is derived for the case of piece-wise linear warping, the same procedure can be used for any arbitrary warping function.

The idea of LT presented here will be a special case of the LT proposed in Umesh *et al* (2005), where they proposed a LT by separating *both* Mel- and VTLN-warping from the filter bank. The main differences between the two approaches are:

- The filters are uniformly spaced in the Mel frequency domain for the approach proposed in this paper, i.e., ν_l are uniformly spaced. In Umesh *et al* (2005), the filters are uniformly spaced in the linear frequency (Hz) domain, i.e., f_l are uniformly spaced.
- The interpolation matrix proposed in Sanand & Umesh (2008) is defined as:

$$\mathbf{L}_{\mathbf{m}}(\tilde{\nu}_l) = \hat{\mathbf{T}}_{\mathbf{m}}^{\alpha}.\mathbf{L}_{\mathbf{m}}(\nu_l), \tag{31}$$

i.e., it performs only VTLN warping on the Mel-warped spectra. In Umesh *et al* (2005), the interpolation matrix is defined as:

$$\mathbf{L}_{\mathbf{m}}(\tilde{\nu}_l) = \hat{\mathbf{T}}_{\mathbf{m}}^{\alpha}.\mathbf{S}(f_l), \tag{32}$$

where $\mathbf{S}$ is the smoothed spectrum without Mel warping and the transformation matrix performs both Mel and VTLN warping to obtain VTLN-warped MFCC features.

6.3c *Transformation of cepstra*: The LT matrix ($\hat{\mathbf{T}}^{\alpha}$) will be of size equal to the number of MFCC filters, **N**. Let **M** represent the number of cepstral coefficients (usually 13) used as static features the feature vector. We then obtain the $\mathbf{M} \times \mathbf{M}$ (13×13) matrix as follows:

$$\begin{bmatrix} C^{\alpha} \\ \mathrm{M} \times 1 \end{bmatrix} = \underbrace{\begin{bmatrix} DCT \\ \mathrm{M} \times \mathrm{N} \end{bmatrix} \begin{bmatrix} \hat{T}^{\alpha} \\ \mathrm{N} \times \mathrm{N} \end{bmatrix} \begin{bmatrix} IDCT \\ \mathrm{N} \times \mathrm{M} \end{bmatrix}}_{J^{\alpha}(\mathbf{M} \times \mathbf{M})} \begin{bmatrix} C^{1.00} \\ \mathrm{M} \times 1 \end{bmatrix}. \tag{33}$$

The feature generation process using the proposed LT approach is illustrated in figure 15.

Although, the LT is derived for the static features, it can be shown that the same relation holds true for Δ and $\Delta\Delta$. Now the relation between warped and unwarped features is given by the relation:

$$\underbrace{\begin{bmatrix} C^{\alpha} \\ C^{\alpha}_{\Delta} \\ C^{\alpha}_{\Delta\Delta} \end{bmatrix}}_{\mathbf{X}^{\alpha}_{i}} = \underbrace{\begin{bmatrix} J^{\alpha} & 0 & 0 \\ 0 & J^{\alpha} & 0 \\ 0 & 0 & J^{\alpha} \end{bmatrix}}_{\mathbf{B}^{\alpha}} \underbrace{\begin{bmatrix} C \\ C_{\Delta} \\ C_{\Delta\Delta} \end{bmatrix}}_{\mathbf{X}_{i}}, \tag{34}$$

where, $\mathbf{X}^{\alpha}_{i}$ (includes the static, velocity and acceleration coefficients) is the feature vector of i^{th} utterance whose spectra is warped with the scale factor α and the LT matrix is denoted by $\mathbf{B}^{\alpha}$.

6.4 *Jacobian compensation in VTLN*

In conventional VTLN, the warp factor is estimated as follows:

$$\hat{\alpha_i} = \arg\max_{\alpha} \Pr(\mathbf{X}^{\alpha}_{i} | \mathbf{W}_i; \lambda), \tag{35}$$

where λ is the SI model, $\mathbf{W}_i$ is the known transcription and $\mathbf{X}^{\alpha}_{i}$ are the warped features for the scale factor α. The above equation is *not* completely accurate. This is because, the likelihood of the warped utterance is evaluated with respect to the SI model which is built using *unwarped data*, and the Jacobian of the transformation has not been accounted. In order to account for the Jacobian, the above equation needs to be modified as:

$$\hat{\alpha_i} = \arg\max_{\alpha} \Pr(\mathbf{X}^{\alpha}_{i} | \mathbf{W}_i; \lambda).|d\mathbf{X}^{\alpha}_{i}/d\mathbf{X}_i|. \tag{36}$$

Since the transformation across the cepstral coefficients is highly complex, the Jacobian of the transformation is usually neglected and the warp factor estimation using Eq. 35 is followed in

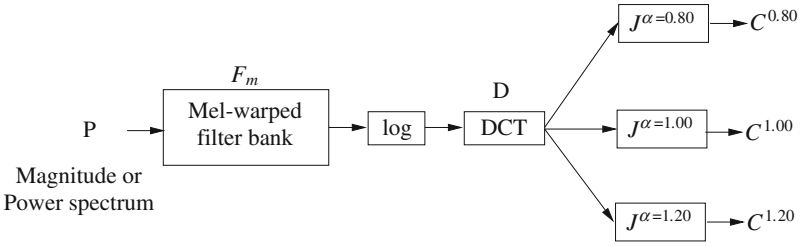


Figure 15. Framework of the proposed LT approach. It should be noted that only the unwarped features are generated and warped features are obtained using the LT matrices (J^{α}).

practice. There have been very few studies that reported the performance of Jacobian in VTLN. Pitz (2005) reported in his thesis that Jacobian provided marginal improvements. Panchapagesan & Alwan (2009) have reported that Jacobian degraded the performance and hence was ignored during recognition. There have been other LT approaches such as McDonough *et al* (2004), Claes *et al* (1998) and Cui & Alwan (2006) but none have reported any study using the Jacobian.

Using the LT relation derived in the previous section to obtain the VTLN-warped MFCC features, the Jacobian of the transformation during the warp-factor estimation can be easily accounted for. In the LT approach to VTLN, the Jacobian will be simply the determinant of the transformation matrix. So, Eq. 36 will now modify as:

$$\hat{\alpha_i} = \arg\max_{\alpha} \left\{ \Pr(\mathbf{X}_i^{\alpha} | \mathbf{W}_i; \lambda) \; |\mathbf{B}^{\alpha}| \right\}, \tag{37}$$

where $\mathbf{B}^{\alpha}$ (see Eq. 34) is the respective transformation matrix for a particular α. It can be easily shown that the likelihood of the observed sequence ($\mathbf{X_t}$) with respect to the model parameters (μ, Σ) and the LT matrix ($\mathbf{B}^{\alpha}$) is equivalent to applying the LT on the observation vectors and accounting for the Jacobian of the transformation (Gales 1998). This is given by:

$$\mathcal{L}(\mathbf{X_t}; \mu, \Sigma, (\mathbf{B}^{\alpha})^{-1}) = \mathcal{L}(\mathbf{B}^{\alpha}\mathbf{X_t}; \mu, \Sigma) + \log(|\mathbf{B}^{\alpha}|). \tag{38}$$

In Eq. 38, the *same* LT is applied on both the means (μ) and the covariances (Σ).

7. Comparison of ASR performance of conventional and LT VTLN

The recognition experiments include three different sets of speech data: TIDIGITS, RM-Task and OGI. Monophone models are used for RM-Task and word models for both TIDIGITS and OGI. All the experiments use 128 filters that are uniformly spaced with uniform bandwidth (with bandwidth equal to that 21 filters for OGI, 24 filters for RM-Task and 29 filters for TIDIGITS) in the Mel domain. This is because 128 filters provided improved performance with Jacobian compensation (discussed later). Intuitively, using more number of filters is equivalent to increased sampling in the spectral domain and hence should reduce the effect of aliasing in the cepstral domain. It should be noted that the cepstra still has only 12 cepstral coefficients along with energy. The filter bandwidths are not scaled and only the centre frequencies are scaled during VTLN (Umesh & Sinha 2007). In all the experiments, the frequency points *zero* and *pi* are mapped onto themselves using piecewise linear warping (Wegmann *et al* 1996).

Table 1 compares the recognition performance of the discussed linear-transformation approach to VTLN along with the conventional approach of filter-bank scaling. The performance of the proposed LT approaches is comparable to the conventional approach of filter bank scaling for VTLN.

Table 2 shows the recognition performance of VTLN with and without Jacobian compensation using the proposed LT approaches. In conventional VTLN, the Jacobian cannot be easily found and is hence ignored. Therefore, there are no Jacobian compensation results for conventional VTLN.

The following observations are made from table 2.

- In the proposed *sinc*- interpolation approach (Sanand & Umesh 2008), Jacobian has provided improvement. For both OGI and RM-Task we see considerable gain in performance. For the case of TIDIGITS, there is not much change in the performance with Jacobian compensation.

Table 1. Recognition performance of conventional and the LT approaches to VTLN. Note that LT approaches provide comparable performance to conventional VTLN and yet are computationally efficient.

Method	OGI	TIDIGITS	RM-Task
	A–A	A–A	A–A
Baseline (No-VTLN)	96.95	99.52	93.71
Conventional VTLN	97.64	99.62	94.81
Sinc interpolation	97.41	99.60	94.85
Cosine interpolation	97.35	99.58	94.73

- For the case of *cosine* interpolation (Panchapagesan 2006), there is a marginal improvement in performance for OGI. TIDIGITS and RM-Task have not shown any improvement. This is because Panchapagesan (2006) incorrectly assumes linear-scaling in the Mel domain in contrast to linear-scaling in the frequency-domain. Linear scaling in frequency domain is physiologically motivated and is used in conventional VTLN.

Although, Jacobian compensation might not provide huge improvements, it should not degrade the recognition performance when properly accounted during the warp factor estimation.

When Jacobian compensation is used under mismatched train and test speaker condition, e.g. male trained model and child test data, then there is usually a degradation in performance. The reason for this degradation is the improper calculation of likelihood while estimating the optimal warp factor using Eq. 36 (Sanand *et al* 2009). This can happen when there is a mismatch between the model and the warped features. If the likelihood calculation is not proper, Jacobian can overcompensate the likelihood and hence result in the degradation of recognition performance. In Eq. 34, the mismatched data X may be assumed to come from a distribution with model parameters μ_T and Σ_T. Applying VTLN linear transform A^α to get warped features X^α would imply that X^α has model parameters $A^\alpha \mu_T$ and $A^\alpha \Sigma_T (A^\alpha)^T$. These model parameters should match the parameters of the SI model (or previous iteration VTLN model) since the data is warped to match this model. When these assumptions do not hold, then Jacobian compensation is not appropriate. Recently (Sanand *et al* 2009) have tried to address this problem.

Table 2. Performance of VTLN with and without Jacobian compensation using the proposed LT approaches. The use of Jacobian ensures mathematical correctness and provides marginal improvement.

Method	OGI	TIDIGITS	RM-Task
	A–A	A–A	A–A
Baseline (No-VTLN)	96.95	99.52	93.71
Sinc	97.41	99.60	94.85
Sinc + Jacobian	97.66	99.60	95.43
Cosine	97.35	99.58	94.73
Cosine + Jacobian	97.48	99.60	94.73

• A–A: Adult train – Adult test.

8. Speaker adaptation approach to normalization

In VTLN, the features are normalized by first warping the spectra resulting in reduced variability for different speakers enunciating the same sound. Such an approach is usually referred to as speaker-normalization or feature-normalization. In literature, there is also a different approach used to improve the performance of speaker-independent (SI) speech recognition systems by modifying the parameters of the system specifically for a particular speaker. This approach is often referred to as speaker-adaptation. SI ASR systems are built using data from all the speakers. These systems are not aimed at modelling the speech from any specific speaker. Therefore, to make the system perform better for a new speaker, a small amount of the new speaker's speech (adaptation data) is used to 'tune' the parameters of the SI model to better match the acoustic data coming from that speaker. This procedure is termed as 'adaptation'. In this method, we are modifying the parameters of the model to better suit the speaker. This approach is different from VTLN where we reduce the inter-speaker variability by normalizing the features, and then building a compact VTLN model.

Speaker adaptation techniques can be broadly classified as:

- Bayesian or MAP approach.
- ML transformation-based approach.
- Speaker cluster/space-based approach.

MAP essentially uses Bayesian approach to parameter re-estimation of the updated model given the adaptation data. One of the major problems with the MAP-based approaches is that the re-estimation formulae *apply only* to individual model parameters. Thus, if a mixture component is *not* observed in the adaptation data, then it cannot be adapted. If we wish to update all the model parameters, then we require a fairly large amount of training data (Gauvain & Lee 1994; Lee *et al* 1990). Therefore, this approach is less commonly used when compared to the ML approach discussed below.

ML transformation approaches estimate a set of transformations that can be applied to the model parameters including unobserved mixture components. These transformations capture the general relation between the original model and the current speaker or the acoustic condition to adapt all the HMM distributions.

Some of the early attempts to estimate the transforms were using the least squares criterion (Hewett 1989). Jaschul (1982) uses a tridiagonal frequency shift matrix, which was later extended by Hewett (1989) to a full transformation matrix. Hewett also showed that his method performs better than the canonical correlation approach (Choukri *et al* 1986), which projects parameters of both reference speaker and the new speaker to a new acoustic space where they are similar. The least square linear regression calculation by Hewett (1989) was replaced by ML estimation taking into account different state distributions. This method is referred to as maximum likelihood linear regression (MLLR) adaptation (Leggetter & Woodland 1995), which estimates a set of LTs for the mean parameters of a mixture Gaussian HMM system to maximize the likelihood of the adaptation data, i.e.,

$$\hat{\mu} = \mathbf{A}\mu + b, \tag{39}$$

where $\mathbf{A}$ is the estimated transformation matrix and b is the bias term. In order to deal with data sparsity, usually several HMM states are grouped together to form adaptation classes that share the same transform. The adaptation classes are determined using a data-driven clustering approach or manually-defined based on broad phonetic classes (Young & Woodland 1994). At

the very least, usually there is a transformation for speech and a separate transformation for silence. In order to reduce the number of parameters to be estimated from the adaptation data of a particular speaker, **A** is often assumed to be block-diagonal and only the block-diagonal elements of the matrix are estimated.

In MLLR, the transformations are applied only on the means of the Gaussian components. Extension of this approach, where separate transforms that are applied to the means and variances separately have also been proposed (Gales & Woodland 1996). If a *single* transformation is used to transform both means and variances, it is called constrained MLLR (CMLLR) (Digalakis *et al* 1995; Gales 1998). The modified parameters of CMLLR are:

$$\hat{\mu} = \mathbf{A}\mu + b \quad \hat{\Sigma} = \mathbf{A}\Sigma\mathbf{A}^T. \tag{40}$$

When they were first proposed, the above LTs of model parameters were done only for test data using the SI model. In Anastasakos *et al* (1996), this approach was also used to linearly transform using data from training speakers to build a canonical model and is called *speaker-adaptive training*.

The speaker cluster/space-based adaptation approach unlike the previously discussed approaches explicitly uses the information about the characteristics of an HMM set for a particular speaker and the two dominant examples of this class are cluster-adaptive training (CAT) (Gales 2000) and eigenvoice technique (Kuhn *et al* 2000). In these methods, a weighted sum of 'speaker cluster' HMMs are estimated, and this interpolated model is used to represent the current speaker. The major difference between CAT and eigenvoice approaches is how the cluster models are estimated.

9. Merits and demerits of VTLN and speaker adaptation

VTLN and speaker adaptation are two different approaches used to improve the performance of speaker-independent (SI) ASR systems. Most state-of-the-art systems have both these components, since speaker adaptation can also compensate for any environment mismatch.

Although, both conventional VTLN and speaker-adaptation methods try to improve the performance of SI-ASR systems by reducing speaker-variability, each method has certain advantages/disadvantages when compared to the other. The T-VTLN method proposed by Sanand *et al* (2007) and Sanand & Umesh (2008) and subsequent enhancements (Akhil *et al* 2008; Rath and Umesh 2009; Rath *et al* 2009) try to combine the advantages of both the approaches and provide the best solution. We discuss below some of these issues:

- *Amount of adaptation data*: As mentioned earlier, in both speaker adaptation and VTLN, the parameters are estimated from the test-adaptation data before a final recognition is done. In the case of VTLN, only a single parameter, namely the warp parameter, needs to be estimated. Therefore, VTLN requires very little adaptation data (possibly few seconds) to robustly estimate the warp factor. On the other hand, speaker-adaptation-based approaches estimate at least two block-diagonal (for silence and speech) matrices for adaptation. If we assume typical MFCC features and their dynamic coefficients, the feature dimension is usually 39. In such a case, we need to estimate $13 \times 13 \times 3$ parameters for each block-diagonal matrix. In practice, we need at least 30 to 40 s of adaptation data to robustly estimate the parameters in speaker-adaptation approaches. Therefore, in applications where there is very little adaptation data, VTLN may be preferred over speaker-adaptation-based approaches.

Table 3. Advantages and disadvantages of the different speaker adaptation and normalization approaches.

	VTLN	MLLR/CMLLR/SAT	T-VTLN
Transform-based (computationally efficient)	×	√	√
Use sufficient statistics (computationally efficient)	×	√	√
No. of estimated parameters	**1**	2 Block-diag matrices	**1**
Rapid (requires little adaptation data)	√	×	√
Robust to transcription errors	√	×	√

- *Robustness to transcription errors*: In both conventional VTLN and speaker-adaptation approaches, an initial transcription is used to obtain frame-state alignment in order to estimate the adaptation/normalization parameters. This is often called 'unsupervised adaptation'. It has been observed that speaker-adaptation methods are very sensitive to transcription errors and often the performance deteriorates dramatically when the transcription is bad. On the other hand, conventional VTLN is robust to errors in transcription (Rath *et al* 2009).
- *Computationally efficiency*: Conventional VTLN requires all the warped features for each frame to be generated in advance before the optimal warping factor is estimated (using Eq. 17). This also requires alignment to be done for all possible warping factors. The alignment is done using the trained model and true transcription during training and first pass recognition during testing. This value of the likelihood is calculated for all the warped features before the optimal estimate can be found. Therefore, conventional VTLN is expensive to implement. On the other hand, speaker-adaptation-based approaches can be estimated very efficiently using sufficient statistics and an EM approach.

Sanand *et al* (2007) have proposed a linear-transformation approach to compute the VTLN-warped features as discussed in section 6 and which we refer to as T-VTLN. This approach overcomes the problems of scaling the filter-bank for each warp factor and subsequently generating the features as is done in conventional VTLN. Using T-VTLN, Akhil *et al* (2008) have also proposed EM-based warp-factor estimation methods that are computationally efficient and yet retain the advantage of conventional VTLN namely robustness to transcription errors and the requirement of only a few seconds of adaptation data to estimate the warp-parameter. Table 3 summarizes the advantages and disadvantages of the various methods discussed above.

10. Conclusion

This paper addresses the issue of acoustic variations among different speakers enunciating the same sound. Although humans can recognize speech from different speakers independent of these variations, researchers have struggled for more than 50 years to come up with a representation that is insensitive to such variations. In this paper, we have given an overview of the research work in this direction. We have first reviewed work in the area of vowel-normalization, and we have also discussed some of our previous work in this area using the concept of universal-warping function and the affine-transform which unify observations made from perception, auditory modelling and speech production studies. We have also discussed the approaches of VTLN and speaker-adaptation which are most commonly used in state-of-the-art systems to reduce inter-speaker variability and thereby improve the performance of SI-ASR systems. In particular, we have discussed, in detail, recent work by many researchers to find

a linear-transformation to VTLN warping and our own work in this area. Finally, we compare the advantages and disadvantages of VTLN and speaker-adaptation, and show that T-VTLN can help get the best of both methods.

Acknowledgments

This work was supported in part by the Department of Science & Technology, Ministry of Science & Technology, India under the DST-SERC project SR/S3/EECE/0058/2008. Most of the work reported is based on the work done by my students and some of the work on speech scale was done jointly with Prof Leon Cohen of the City University of New York and Dr. Nelson of the U.S. Department of Defense. I would like to especially acknowledge the work of Sanand and Shakti since most of the work reported here is a culmination of their efforts. I would also like to acknowledge the work of Bharath, Rohit Sinha and Harish which have been reported here. Finally, I would like to thank Shakti Prasad Rath, Achintya Sarkar, Raghavendra Bilgi, Vikas Joshi and Vinay Chakravarthi for reading through the manuscript.

References

Acero A 1990 Acoustical and environmental robustness in automatic speech recognition. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA

Acero A, Stern R M 1991 Robust speech recognition by normalization of the acoustic space. *Proc. IEEE ICASSP*, Toronto, Canada, 893–896

Adank P, Smits R, van Hout R 2004 A comparison of vowel normalization procedures for language variational research, *J. Acoust. Soc. Am.* 116(5): 1–9

Adank P M 2003 Vowel-normalization – a perceptual-acoustic study of dutch vowels. Ph.D. thesis, University of Nijmegen, The Netherlands

Akhil P T, Rath S P, Umesh S, Sanand D R 2008 A computationally efficient approach to warp factor estimation in VTLN using EM algorithm and sufficient statistics. *Proc. of Interspeech*, 1713–1716

Anastasakos T, McDonough J, Schwartz R, Makhoul J 1996 A compact model for speaker adaptive training. *Proc. Int. Conf. Spoken Lang. Process.*

Andreou A, Kamm T, Cohen J 1994 Experiments in vocal tract normalization. *Proc. CAIP Workshop: Frontiers in Speech Recognition II*

Bladon R A W, Henton C G, Pickering J B 1983 Towards an auditory theory of speaker normalization. *Lang. Commun.* 4: 59–69

Choukri K, Chollet G, Grenier Y 1986 Spectral transformations through canonical correlation analysis for speaker adaptation in ASR, vol. 11, 2659–2662

Claes T, Dologlou I, Bosch L, Compernolle D V 1998 A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Trans. Speech Audio Process.* 6(6): 549–557

Cui X, Alwan A 2006 Adaptation of children's speech with limited data based on formant-like peak alignment. *Comput. Speech Lang.* 20(4): 400–419

Digalakis V, Rtischev D, Neumeyer L, Sa E 1995 Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Trans. Speech Audio Process.* 3: 357–366

Fant G 1975 A non-uniform vowel normalization. Tech. Rep. 2-3, Speech Transmiss. Lab. Rep., Royal Inst. Tech., Stockholm, Sweden

Gales M, Woodland P 1996 Mean and variance adaptation within the MLLR framework. *Comput. Speech Lang.* 10: 249–264

Gales M J F 1998 Maximum likelihood linear transformations for HMM-based speech recognition. *Comput. Speech Lang.* 12(2): 75–98

Gales M J F 2000 Cluster adaptive training of hidden Markov model. *IEEE Trans. Speech Audio Process.* 8(4): 417–428

Gauvain J L, Lee C H 1994 Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Trans. Speech Audio Process.* 2: 291–298

Gerstman L 1968 Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust.* AU-16: 78–80

Harish A N, Sanand D R, Umesh S 2009 Characterizing speaker variability using spectral envelopes of vowel sounds. *Proc. Interspeech*, Brighton, UK, 1107–1110

Hewett A J 1989 Training and speaker adaptation in template-based speech recognition. Ph.D. thesis, Cambridge University

Jaschul J 1982 Speaker adaptation by a linear transformation with optimised parameters, vol. 7, 1657–1660

Kamm T, Andreou G, Cohen J 1994 Vocal tract normalization in speech recognition compensation for systematic speaker variability. *Proc. of the 15th Annual Speech Research Symposium*, CSLP, Johns Hopkins University, Baltimore, MD, 175–178

Kuhn R, Junqua J C, Nguyen P, Niedzielski N 2000 Rapid speaker adaptation in eigenvoice space. *IEEE Trans. Speech Audio Process.* 8(6): 695–707

Kumar S V B, Umesh S 2008 Nonuniform speaker normalization using affine transformation. *J. Acoust. Soc. Am.* 124(3): 1727–1738

Ladofoged P, Broadbent D 1957 Information conveyed by vowels. *J. Acoust. Soc. Am.* 29: 98–104

Lee C-H, Lin C-H, Juang B-H 1990 A study on speaker adaptation of continuous density HMM parameters. *ICASSP-90. 1990 Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 145–148

Lee L, Rose R 1998 Frequency warping approach to speaker normalization. *IEEE Trans. Speech Audio Process.* 6: 49–59

Leggetter C J, Woodland P 1995 Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.* 9(2): 171–185

Lobanov B 1971 Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am.* 49: 606–608

McDonough J, Byrne W, Luo X 1998 Speaker normalization with all-pass transforms. *Proc. Int. Conf. Spoken Lang. Process.* vol. 6, 2307–2310

McDonough J, Schaaf T, Waibel A 2004 Speaker adaptation with all-pass transforms. *Speech Commun.* 42(1): 75–91

Miller J D 1989 Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.* 85(5): 2114–2134

Molau S, Pitz M, Schluter R, Ney H 2001 Computing mel-frequency cepstral coefficients on the power spectrum. *Proc. (ICASSP '01), 2001 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 73–76

Nearey T M 1978 Phonetic feature systems for vowels. Tech. rep., Indiana University Linguistics Club

Nearey T M 1992 Applications of generalized linear modeling to vowel data. *Proc. ICSLP '92*, Canada

Nordström P E, Lindblom B 1975 A normalization procedure for vowel formant data. *Int. Cong. Phonetic Sci.*, Leeds England

Oppenheim A, Johnson D 1972 Discrete representation of signals. *Proc. IEEE* 60(6): 681–691

Panchapagesan S 2006 Frequency warping by linear transformation of standard MFCC. *Proc. Interspeech*, Pittsburgh, Pennsylvania

Panchapagesan S, Alwan A 2009 Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC. *Comput. Speech Lang.* 23(1): 42–64

Peterson G E 1961 Parameters of vowel quality. *J. Speech Hearing* 4: 10–29

Peterson G E, Barney H L 1952 Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24(2): 175–184

Pitz M 2005 Investigations on linear transformations for speaker adaptation and normalization. Ph.D. thesis, RWTH Aachen

Pitz M, Molau S, Schlter R, Uter R S, Ney H 2001 Vocal tract normalization equals linear transformation in cepstral space. *Proc. Eurospeech*, 2653–2656

Rath S P, Umesh S 2009 Acoustic class specific vtln-warping using regression class trees. *Proc. Interspeech*, Brighton, UK, 556–559

Rath S P, Umesh S, Sarkar A K 2009 Using VTLN matrices for rapid and computationally-efficient speaker adaptation with robustness to first-pass transcription errors. *Proc. Interspeech*, Brighton, UK, 572–575

Sanand D R, Kumar D D, Umesh S 2007 Linear transformation approach to vtln using dynamic frequency warping. *Proc. Interspeech*, Antwerp, 1138–1141

Sanand D R, Rath S P, Umesh S 2009 A study on the influence of covariance adaptation on Jacobian compensation in vocal tract length normalization. *Proc. Interspeech*, Brighton, UK, 584–587

Sanand D R, Umesh S 2008 Study of jacobian compensation using linear transformation of conventional MFCC for VTLN. *Proc. Interspeech*, Brisbane, Australia, 1233–1236

Sondhi M 1986 Resonances of a bent vocal tract. *J. Acoust. Soc. Am.* 79: 1113–1116

Stevens S S, Volkman J 1940 The relation of pitch to frequency. *Am. J. Psychol.* 53: 329

Sussman H 1986 A neuronal model of vowel-normalization and representation. *Brain Lang.* 28: 12–23

Syrdal A K, Gopal H S 1983 Perceived critical distances between $F_1 - F_0$, $F_2 - F_1$, $F_3 - F_2$. *J. Acoust. Soc. Am.* 74(S1): S88–S89

Syrdal A K, Gopal H S 1986 A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.* 79(4): 1086–1100

Umesh S, Cohen L, Nelson D 2002a Frequency warping and the Mel scale. *IEEE Signal Process. Lett.* 9(3): 104–107

Umesh S, Cohen L, Nelson D 2002b The speech scale. Acoustics Research Letters Online of the *J. Acoust. Soc. Am.* 3(3): 83–88

Umesh S, Kumar S V B, Vinay M K, Sharma R, Sinha R 2002c A simple approach to non-uniform vowel normalization. *Proc. IEEE ICASSP '02*, Orlando, USA, 517–520

Umesh S, Sinha R 2007 A study of filter-bank smoothing in MFCC features for recognition of children's speech. *IEEE Trans. Speech Audio Process.* 15(8): 2418–2430

Umesh S, Zolnay A, Ney H 2005 Implementing frequency warping and VTLN through linear transformation of conventional MFCC. *Interspeech*, Lisbon, Portugal, 269–272

von Békésy G, Rosenblith W A 1951 in: S S Stevens (ed.), *Handbook of experimental psychology* (New York: John Wiley) 985–1039

Wegmann S, McAllaster D, Orloff J, Peskin B 1996 Speaker normalization on conversational telephone speech. *IEEE ICASSP '96*, Atlanta, USA, 339–341

Young S J, Woodland P C 1994 State clustering in hidden Markov model-based continuous speech recognition. *Comput. Speech Lang.* 8(4): 369–383

Zolfaghari P, Robinson T 1996 Formant analysis using mixtures of gaussians. *ICSLP 96. Proc. Fourth Int. Conf. on Spoken Language 1996*, vol. 2, 1229–1232