

Contents

1	Introduction	1
2	Non-Uniform Vowel Normalization	4
2.1	Simple Linear Scaling	4
2.2	Non-Linear Scaling	6
2.3	Proposed Non-Linear Scaling	9
2.3.1	Weighting Factor as a Function of Frequency	9
2.3.2	Experiments and Results	12
2.4	Summary	16
3	A Warping Function for Non-uniform Vowel Normalization	18
3.1	Non-linear Scaling of Continuous Vowel Spectra	18
3.1.1	Vocal Tract Length Estimation	19
3.1.2	Experiments and Results	19
3.2	Approximate Scaling Function	22
3.3	Warping-Function	22
3.4	Summary	28
4	Recognizers With Scale Factor Estimation	29
4.1	Hidden Markov Model Based Speech Recognizer	30
4.1.1	Feature Extraction	30
4.1.2	Pattern Recognition	33
4.2	Non-uniform Normalization on the Recognizer	34
4.2.1	Scaling Factor Estimation in ML Sense	34
4.2.2	Training and Testing Procedure	35

4.2.3	Non-Linear Scaling With Filterbank Analysis	36
4.2.4	Experiments and Results	39
4.3	Summary	41
5	Recognizers With Non-Linear Scale Invariance	42
5.1	Non-linear Scale Invariance on a Recognizer	42
5.1.1	Non-Linear Scale Invariant Transformation	43
5.1.2	Experiments and Results	46
5.2	Summary	48
6	Conclusions	49

List of Figures

2.1	Deviations from linear scaling.	5
2.2	Weighting factors for females and children with reference male.	7
2.3	Weighting factors for children with reference male and female.	8
2.4	Weighting factor distribution for /IY/ and /UH/ categories.	11
2.5	Proposed non-linear scaling function.	12
2.6	Effect of normalization on female cluster.	14
2.7	Scatter plots of $F_1 - F_2$ for 10 Vowels.	17
3.1	Scatter plot of estimated VTL.	20
3.2	Various scaling functions as applied to the front vowel /IH/.	21
3.3	Various scaling functions as applied to the open vowel /AE/.	22
3.4	Various scaling functions as applied to the rounded vowel /UH/.	23
3.5	Approximate scaling function	24
3.6	Approximate scaling function as applied to the front vowel /IH/.	25
3.7	$\beta(f)$ for non-linear scaling schemes.	26
3.8	Warping functions for various non-linear scaling schemes	27
4.1	Block diagram of a continuous speech recognizer	30
4.2	Feature Extraction stages	32
4.3	Recognition with speaker normalization	37
4.4	Mel filterbank analysis with frequency warping	38
5.1	Log warping of two scaled signals	44

List of Tables

2.1	Formant and vowel specific scale factors, K_{nfm}	10
2.2	Cluster discriminability in terms of F-Ratio	15
2.3	Residual variance after normalization	16
3.1	β values as used by Umesh et al	28
4.1	Mel filter boundaries and $\gamma(f)$	38
4.2	Recognition performance	40
5.1	Eight band non-linear frequency warping Implementation	46
5.2	Fifteen band non-linear frequency warping Implementation	46
5.3	Recognition performance of non-linear scale-invariant transformations	48

Chapter 1

Introduction

Automatic speech recognition (ASR) system enables a computer (or a machine) to recognize the words spoken by a person. ASR is essentially the process of converting speech to text. An ideal ASR system should be able to recognize with 100% accuracy all the words (as defined by the dictionary) that are spoken by any person independent of vocabulary size, speaker characteristics, accent, noise and channel characteristics. But the present day ASR systems are quite far from being ideal. There are broadly two classes of ASR systems (1) Speaker dependent and (2) Speaker Independent systems. This distinction is purely based on the kind of training dataset used. Speaker dependent (SD) systems are trained from speech data collected from a single user, who is the sole user of the system. Speaker independent systems (SI) on the other hand are trained from speech collected from many different users. Typical applications of SD systems are in desk-top applications, word processing, etc. While SI systems are typically used at public interfaces like airline interface system, telephone directory service, etc where there are varied type of speakers. The parameters of the acoustic models of a SI system are estimated from speech collected from a large population of speakers, in order to model the the large speaker variability in the user population. While the SI systems yield better recognition rates for test speakers who are not in the training dataset than speaker dependent systems, they are less accurate (have 2-3 times higher word error rates) than *adequately trained* speaker dependent systems for a given speaker who has contributed to the training dataset. This degradation in the performance of SI systems over SD systems for a given speaker is mainly due to large speaker variability present in the training set of

SI systems. Speaker normalization techniques aim to remove these speaker specific variabilities from the SI systems.

Because of the superior performance of SD systems over SI systems for a given task and speaker it is clear that speaker variabilities play a major role in recognition performance. Speaker variabilities can be either due to non-physiological factors like dialect, emotions, speaking idiosyncronacies, background noise etc., or due to intrinsic factors like changes in vocal tract physiological characteristics, pitch, etc. which are usually dependent on the age and the gender of the speaker.

There are two broad approaches to being robust to some of these variations (1) **Model based** approaches like Maximum likelihood linear regression (MLLR) [1], Speaker Adaptive Training (SAT) [2], etc., derive a transformation for the model parameters to arrive at a compact model sans speaker and other variabilities. (2) **Feature based** approaches like affine transformations in feature space (cepstral mean subtraction [3]), features derived from scale-invariant transformations [4], scale factor estimation and subsequent compensation schemes [5] [6], etc., aim to extract features which are invariant to speaker dependent variations.

Vocal tract length (VTL) variation has been one of the major contributors to these speaker-variabilities [7]. It has been found that VTL variation causes scaling in the spectral domain [8]. Formant (spectral peak) position variations are found to be closely associated with the length variations of the vocal tract components [7] [9] [10]. Most of the normalization techniques assume that, the linear scaling of formants (spectra in general) to be the main source of variation and hence needs to be compensated. Compensating for this variability by re-scaling the frequency axis provides substantial improvements in speech recognition performance [6] [8] [11]. But with non-proportional variations of the vocal tract component lengths among speakers, scaling has been found to be highly context dependent [12] and hence non-linear. These experiments suggest that a more detailed modelling of “scaling function” of the frequency axis may be helpful in understanding vowel (or in general phoneme) perception [13] [14] [15].

In the first half of this thesis we have attempted to model these non-linearities in scaling as a function of frequency alone and to decouple it from context dependence. An approximate non-linear scaling function for vowel normalization motivated from the works of Fant [12] has been derived. This function is general-

ized to be applicable to continuous vowel spectra. This generalized non-linear scaling function has been applied to a “Hidden Markov Model (HMM)” based recognizer with scale factor estimated in maximum likelihood (ML) sense. The second half of the thesis is motivated from the scale-invariant transformations [15]. A warping function is derived from our proposed non-linear scaling function. This warping function is incorporated into a non-linear scale invariant transformation so as to be applicable on a HMM-based recognizer. We have used different analysis methods such as formant data analysis, spectral alignments and HMM based recognizers, to compare the performance of the proposed methods with other similar techniques.

The thesis is organized as follows. In Chapter 2 the motivation for non-linear scaling is provided. This non-linearity is expressed as a pure function of frequency. A simple procedure for non-uniform vowel normalization on “formant frequencies” is presented. The procedure is evaluated in terms of F-ratio, residual variance and other separability measures. In Chapter 3 the proposed non-uniform normalization scheme is generalized to continuous spectra and is evaluated for subjective spectral alignment of vowel data. A warping function is defined from this non-linear scaling model so that it can be incorporated in a non-linear scale invariant transformation for HMM-based recognizer. In Chapter 4 a method for incorporating the proposed non-linear scaling function in a HMM-based recognizer has been presented. In Chapter 5 a non-linear scale invariant transformation from the proposed warping function is derived and used in a HMM based recognizer. Performance of these algorithms are evaluated, analysed and compared with respect to that of other similar techniques, using percentage correct and percentage accuracy as performance measures. Finally, in Chapter 6 using all the experimental results conclusions are drawn about the effectiveness of the proposed non-linear warping function in a recognizer/classifier framework.

Chapter 2

Non-Uniform Vowel Normalization

In speech recognition, speaker dependence of the speech signal is mainly due to the variation in the vocal tract length of the speakers. An average adult male has a vocal tract length (VTL) of around 17cm, while an average female VTL measures around 14.5cm. With uniform tube as a first order approximation to the vocal tract, it is obvious that on an average the formants of an average female speaker are scaled up by 20% with respect to that of an average male speaker. Hence, it is commonly assumed that differences in formant patterns between male and female speakers, are related by a pure scale factor which is inversely proportional to the “*overall*” vocal tract length.

2.1 Simple Linear Scaling

Nordstrom & Lindblom [8] have suggested a simple normalization procedure, based on the estimate of the speaker’s average vocal-tract length in open vowels, as determined from measurement of the third formant F_3 . In their procedure of uniform scaling, the formant frequency of the subject to be normalized is simply divided by,

$$\alpha = (1 + k/100) = \frac{F_{3av}}{F_{3ref}} = \frac{l_{ref} + 1}{l_{av} + 1} \quad (2.1)$$

where k is the scale factor in percentage, l_{av} is the vocal-tract length associated with the subject’s average F_3 of open vowels (vowels with F_1 greater than 600Hz) and l_{ref} is the vocal-tract length of the reference “male” speaker.

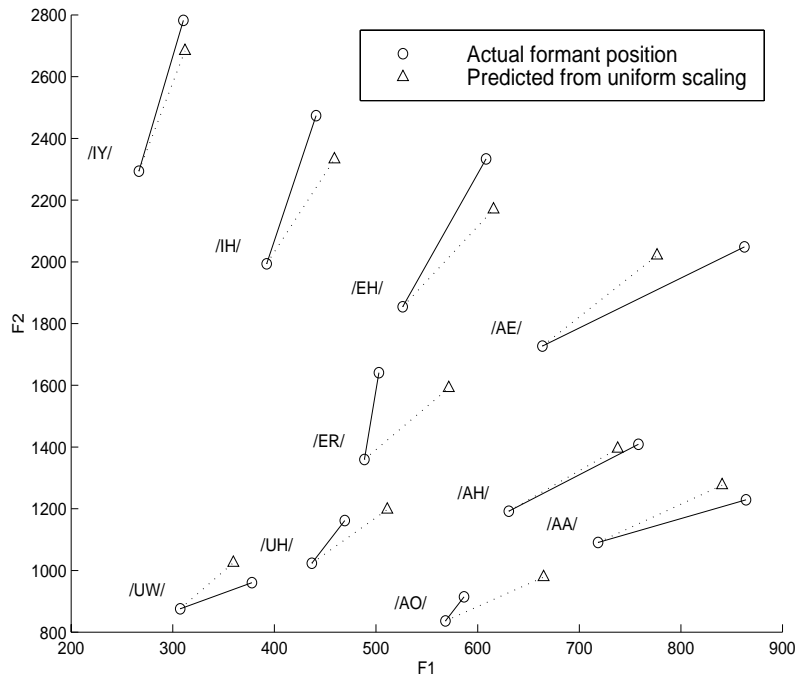


Figure 2.1: Deviations from linear scaling.

Figure shows the actual values of average male (F_1, F_2) and average female (F_1, F_2) points for various vowel categories [16]. Dashed line indicates predicted (F_1, F_2) point, with a linear scaling $\alpha = 1.17$. Variation in the distances, between predicted and actual points, over vowel categories is evident. F_1 & F_2 are in Hz.

But in general formant frequency locations for vowels are affected by three factors: the overall length of the pharyngeal-oral tract, the location of constrictions along the tract, and the narrowness of the constrictions[10]. Simple scale rule neglects both the location of the constrictions and the vocal tract shape. The actual nature of scaling among vowel categories can be seen in Fig 2.1. It can be seen from the figure, that the scaling is not uniform, but is a function of both formant number and vowel category.

2.2 Non-Linear Scaling

The observations suggest gross deviations from linear scaling. “*Why are these deviations present?*” Various vowel articulatory constraints depend upon different vocal cavity dimensions. It has been found that F_2 is pharynx dependent while F_3 is mouth cavity dependent. These “formant-cavity” affiliations, coupled with unequal ratios of pharyngeal to oral cavity lengths between males and females, results in these gross deviations. A non-uniform vowel normalization procedure suggested by Fant [12], where in, a weighting factor k_{nf} , which is both a function of formant number and vowel category, is applied to the ratio of the subject’s particular k (Eqn 2.1) to the $k = 17\%$ of the reference female. With this non-uniform normalization, Fant showed substantial reduction in speaker differences between males and females.

For a given vowel category, the formant specific weighting factor is denoted by k_{nf} in general. To identify the subject and reference speaker classes additional suffixes are added, for eg: K_{ncf} denotes the n^{th} formant specific weighting factor with children as the subjects and female as the reference speaker.

For female speakers,

$$K_{nfm} = \left(\frac{F_n \text{ female}}{F_n \text{ male}} - 1 \right) 100\% \quad (2.2)$$

For children speakers,

$$K_{ncm} = \left(\frac{F_n \text{ child}}{F_n \text{ male}} - 1 \right) 100\% \quad (2.3)$$

where “ $F_n \text{ female}$ ”, “ $F_n \text{ child}$ ” and “ $F_n \text{ male}$ ” are the average n^{th} formants of female, children and male cluster of speakers respectively, for a vowel category.

If the linear-scale rule were to be true these factors should have turned out to be constant ($\alpha - 1$), over the formant numbers and the vowel categories. But Fig 2.2 shows, that there are systematic and substantial variations of these factors with respect to vowel categories and formant numbers. To analyze the child-female scaling variations let us define a new set of formant specific weighting factors,

$$K_{ncf} = \left(\frac{F_n \text{ child}}{F_n \text{ female}} - 1 \right) 100\% \quad (2.4)$$

with female as the reference speaker and child as the subject speaker. These factors are plotted in Fig 2.3. It can be observed from the figure that, these factors are less formant and vowel specific as compared to child-male weighting factors.

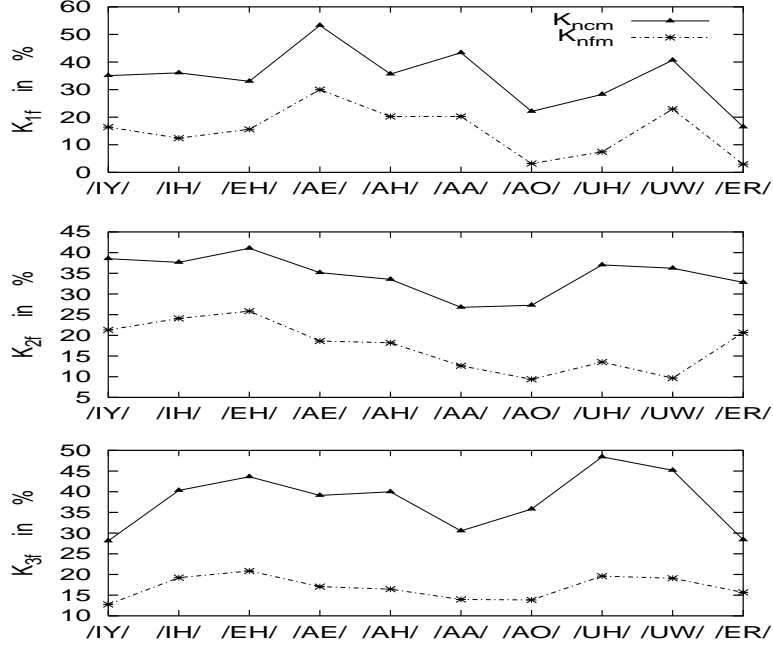


Figure 2.2: Weighting factors for females and children with reference male. Figure shows the weighting factors in percentage for female and children groups with male as the reference speaker. K_{nfm} plots Eqn 2.2, K_{ncm} plots Eqn 2.3. It can be observed that K_{3fm} is less vowel specific and is hence used in scale factor, k , estimation.

Fant's Normalization Procedure:

1. Determine the subject's F_{3av} .
2. Calculate the ratio $F_{3av}/F_{3ref} = (1 + k/100)$.
3. For each vowel to be normalized, select the corresponding female-male formant specific scale factor set, $k_{nf} = k_{1f}, k_{2f}, k_{3f}, \dots$.
4. To normalize the n^{th} formant, apply a weighting to the scale factors k_{nf} with the ratio of the subject's particular k to the $k = 17\%$ of the reference speaker.

$$k_n = k_{nf} \cdot \frac{k}{17} \quad (2.5)$$

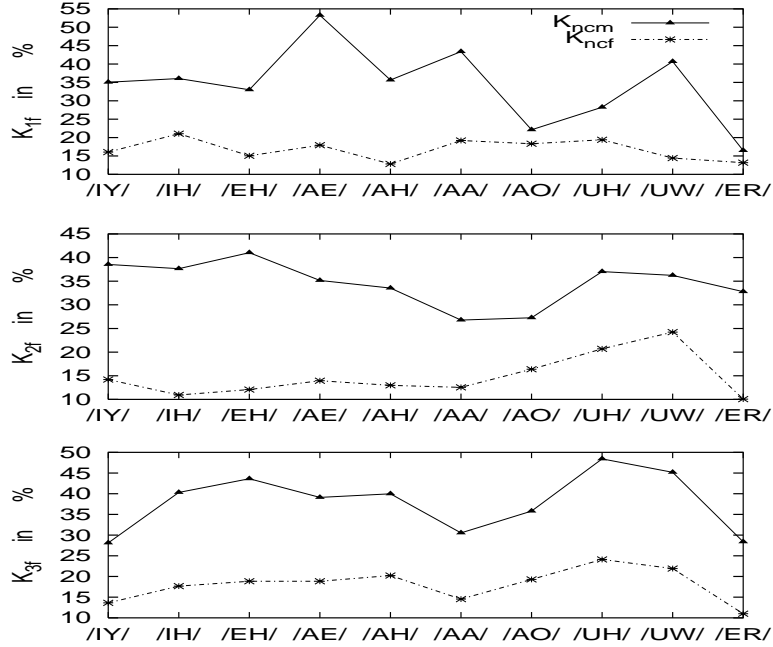


Figure 2.3: Weighting factors for children with reference male and female. Figure shows the weighting factors in percentage for children with female and male as the reference speakers. K_{ncf} plots (Eqn 2.4), K_{ncm} plots (Eqn 2.3). Child-Female K_{ncf} scaling variations are less vowel and formant specific, when compared to Child-Male K_{ncm} scaling variations. Note: K_{ncm} plot is also shown in Fig 2.2.

where $n =$ formant number. This represents the best prediction of the subject's scale factors for the particular vowel. For child subjects, in which case k values will be higher than 24%,

$$k_n = k_{nf} \cdot \frac{24}{17} + (k - 24)$$

The rationale behind this being the tendency of child-female difference to be less formant and vowel specific as it is evident from Fig 2.3.

Note: With male as the reference speaker, scale factor k for average female speaker is around 17 (i.e., $\alpha = 1.17$).

2.3 Proposed Non-Linear Scaling

The main motivation for the work proposed here, is to apply these methods to reduce inter-speaker difference for speech recognition/ classification application. While uniform scaling may be easily applied for such an application, it is preferable to use non-uniform normalization because of the improved separability. But Fant’s non-uniform normalization is not practically implementable, because firstly it requires the knowledge of vowel category and formant number, and secondly it is not directly applicable to continuous spectral patterns. In the proposed algorithm the idea is to model the weighting factor, k_{nf} , as a function of frequency alone and hence implement the non-linear scaling. This algorithm should do away with the need for *a priori* knowledge about the vowel category, while at the same time should do better than conventional linear scaling.

2.3.1 Weighting Factor as a Function of Frequency

To estimate the weighting factor k_{nf} as a function of frequency, the (F_1, F_2, F_3) triplets, of 33 men and 28 women covering all the 10 vowels [16] spoken in the same context were used. Though Fant[12] used a mixture of vowel databases from various languages like Swedish, American English, Danish, Estonian, Dutch, Japanese etc., in this work only American English vowel-formant database is used for both analysis and testing.

The weighting factor, k_{nf} , in Eqn 2.5 is a function of both formant number and vowel category, i.e., a vowel and formant specific scale factor. These weighting factors, k_{nf} , as obtained for American English database and for a mixture of six languages (as used by Fant), are tabulated adjacent to each other for comparison in Tab 2.1. This information is averaged over vowel category and formant number, to obtain a weighting function, γ , which is purely a function of frequency [17].

The plots of k_{nf} values for two vowel categories, for all female speakers with respect to frequency, are shown in Fig 2.4. Such discrete distributions of k_{nf} , for all vowel categories are accumulated and then averaged as a function of frequency. Here the formant frequencies of all the speakers in a vowel category are mapped on to the corresponding formant and vowel specific scale factor, k_{nf} . Hence we get an array

Formant scale factors in % Vowels	k_1		k_2		k_3	
	(A)	(F)	(A)	(F)	(A)	(F)
/IY/	16	7	21	21	12	13
/IH/	12	11	24	22	19	18
/EH/	15	19	25	18	20	20
/AE/	29	27	18	17	17	18
/AH/	20	18	18	18	16	16
/AA/	20	25	12	15	13	15
/AO/	03	11	09	06	13	13
/UH/	07	03	13	12	19	18
/UW/	22	06	09	01	19	23
/ER/	02	02	20	21	15	16

Table 2.1: Formant and vowel specific scale factors, K_{nfm} Eqn 2.2,
Here (A) denotes American English vowel database [16] used for analysis. (F) denotes mixture of six languages which Fant used for analysis [12]

of frequency specific scale factor values, which is not dependent on formant number and vowel category. Mathematically,

$$\gamma(f_n) \rightarrow k_{nf}$$

where k_{nf} is the n^{th} formant of the i^{th} vowel category, f_n denotes n^{th} formant of all speakers in the i^{th} vowel category. Note that for a given frequency f_o , corresponding k_{nf} can have different values depending upon the subject, vowel category and formant numbers, $\gamma(f_o)$ is computed as a simple average of all the values of k_{nf} , over a small band (100Hz width) in the vicinity of f_o to get a smooth frequency specific scale factor function.

$$\gamma(b_i) = \sum_{f_n \in b_i} \gamma(f_n) / \mathcal{N}(\{f_n \in b_n\}) \quad (2.6)$$

where $\mathcal{N}(\cdot) \rightarrow$ cardinality of the set and $n=1,2,3$. A plot of $\gamma(f)$ as a function of frequency is shown in Fig 2.5.

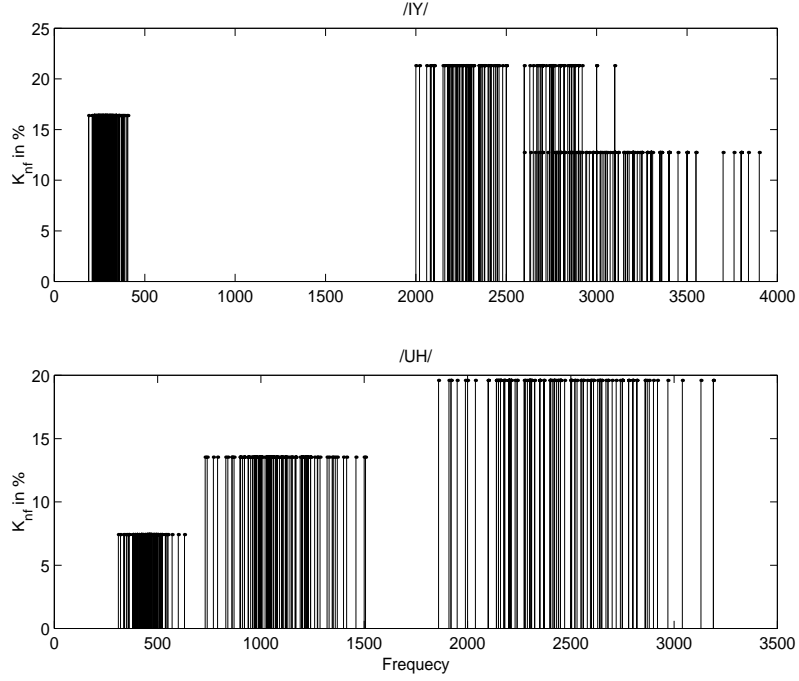


Figure 2.4: Weighting factor distribution for /IY/ and /UH/ vowel categories. Figure shows the weighting factors in percentage for female group with male as the reference speaker. The discrete distributions of K_{n_f} over formant frequencies (in Hz) for various vowel categories are later combined to get a smoother function (Eqn 2.6).

Proposed Normalization Procedure:

To normalize the subject's formant frequency F_n of any vowel, which lies in the m^{th} frequency band, b_m (i.e. $F_n \in b_m$), divide it by a scale factor given by, for adults:

$$\begin{aligned}
 k_{nonuniform} &= \frac{\gamma(b_m).k}{17} \text{ (for male reference speaker)} \\
 &= \frac{\gamma(b_m).k}{-14.53} \text{ (for female reference speaker)}
 \end{aligned} \tag{2.7}$$

where $F_n \in b_m$, k is the subject's uniform scale factor, for child ($k > 24$):

$$\begin{aligned}
 k_{nonuniform} &= \gamma(b_m). \frac{24}{17} + (k - 24) \text{ (for male reference speaker)} \\
 &= \frac{\gamma(b_m).k}{-14.53} \text{ (for female reference speaker)}
 \end{aligned} \tag{2.8}$$

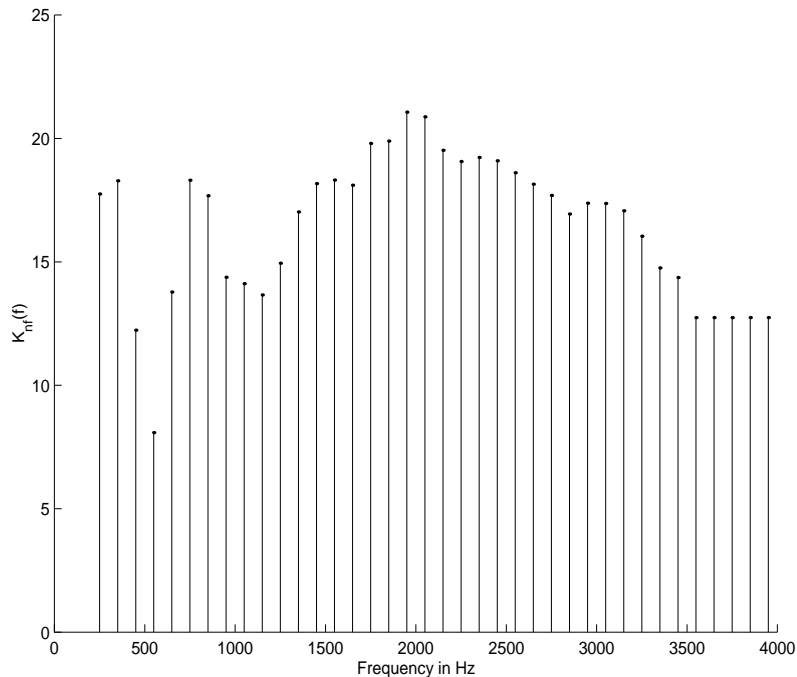


Figure 2.5: Proposed non-linear scaling function.

Figure shows the scaling function in percentage for female group with male as the reference speaker (Eqn 2.6).

2.3.2 Experiments and Results

The motivation for normalization is to improve the separability among vowel clusters in (F_1, F_2, F_3) space. Here the effectiveness (in reducing intracluster variance and retaining the separation between clusters) of the proposed normalization scheme is measured using F-ratio, residual variance after normalization [12] and scatter plots. Peterson and Barney [16] database is used, for both analysis and testing.

A Subject's Scale Factor Estimation

The determination of the average scale factor k is done differently for linear [8] and non-linear [12] scaling schemes. Nordstrom and Lindblom have used only open vowels (i.e., those vowels with $F_1 > 600\text{Hz}$, /AA/, /AE/ & /AA/). Subject's scale

factor is given by,

$$\alpha = (1 + k_{open}/100) = \frac{F_{3av}}{F_{3ref}} \quad (2.9)$$

where F_{3av} and F_{3ref} are the average 3rd formant measurements for open vowels, of the subject and of “all male speakers” (reference speaker is male) respectively. This scale factor has been used for all linear scaling experiments. Apart from using the scale factor from open vowels k_{open} , Fant also used the k value determined from $(F_2 F_3)^{1/2}$ of the front vowel /IY/, with 0.5 weighting. As noted earlier F_2 , F_3 are known to be related to half-wavelength resonance in the pharynx and the mouth cavity. Mouth-cavity length difference between females and males is more than pharynx cavity length. This would result in $k_2 /IY/ \geq k_3 /IY/$, such that $k = 0.5(k_2 /IY/ + k_3 /IY/)$, thus it is more balanced to include this factor also, in the estimation of actual k . Thus the scale factor used for non-linear scaling experiments is given by,

$$k = \frac{2k_{open} + \frac{1}{2}(k_2 /IY/ + k_3 /IY/)}{3} \quad (2.10)$$

B Results

Normalization Plots

The distances between average (F_1, F_2) points after applying various normalization procedures, to the single database, are shown in Fig 2.6. The weighting factors, k_{nf} , used for Fant’s non-uniform normalization scheme, are taken from Tab 2.1, while for the proposed scheme uses the smooth weighting function, $\gamma(f)$ obtained Eqn 2.6.

F-ratio

Since discriminability between vowel clusters is as important as reduction of variance with any given vowel clusters, a good measure for the usefulness of the normalization schemes, would be the F-ratio[18]

In deriving the F-ratio separability, let M_i and R_i denote the mean formant (F_1, F_2, F_3) vector and its covariance matrix, respectively, of the i^{th} vowel class. An equal probability of vowel classes is assumed. Let $M_o = (1/I) \sum_{i=1}^I M_i$, where I denotes the number of vowel classes being compared. Then between-class S_b and

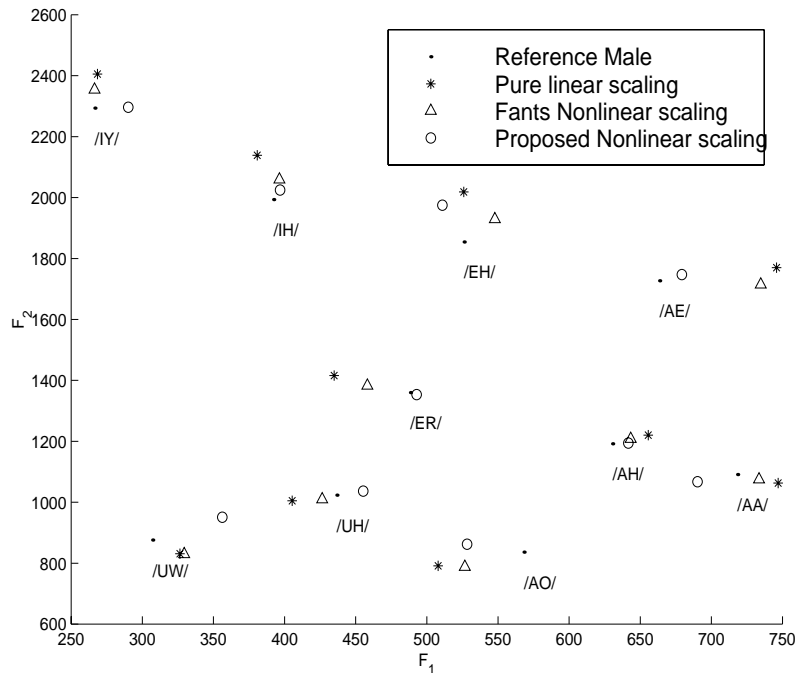


Figure 2.6: Effect of normalization on female cluster.

Figure shows the average (F_1, F_2) points of normalized female cluster. Normalization being done by various methods. With normalization distance between male and female clusters are reduced. F_1 and F_2 are in Hz

within class S_w scatter matrices, are computed by,

$$S_w = \frac{1}{I} \sum_{i=1}^I R_i$$

$$S_w = \frac{1}{I} \sum_{i=1}^I (M_i - M_o)(M_i - M_o)_T$$

The separability criterion is then given by,

$$J = \text{trace}\{(S_b + S_w)^{-1}S_b\} \quad (2.11)$$

The cluster discriminability (in terms of F-ratio, J, using first three formants) of intrinsic, linearly scaled, Fant's non-linearly scaled and proposed non-linearly scaled vowel clusters from Peterson and Barney database are tabulated in

<i>F-Ratio</i>	<i>Ref. Male</i>		<i>Ref. Female</i>	
	Ad. Only	Ad. & Ch.	Ad. Only	Ad. & Ch.
Unwarped	2.21	2.01	2.21	2.01
Uniform Scaling	2.45	2.43	2.45	2.43
Simple Non-uniform	2.46	2.43	2.52	2.48
Fant's Non-uniform	2.70	2.68	2.75	2.72

Table 2.2: Cluster discriminability in terms of F-Ratio.

Performance of the proposed non-linear scaling function as compared to other methods, applied on Peterson and Barney data [16]. Here Ad. stands for adult speakers and Ch. stands for child speakers.

Table 2.2. In Eqn 2.11 as the separability improves, J should approach the *ideal* value of 3. Here one should note that the F-Ratio in case of uniform scaling does not change with change in the reference speaker, as the variance measure is unaffected by a change in linear scaling.

Residual Variance

For each subject, Eqn 2.8 and Eqn 2.9 give the k_n to be used with respect to the reference speaker. This is our “prediction” of the formant specific scale factor value k_n . The actual or the observed scale factor k_{obs} of the subject may be quite different. The efficacy of the normalization scheme is reflected by how close the prediction k_n is to the observed k_n ($= k_{obs}$). This is given by the residual variance remaining after normalization [12]. Since with reference female speaker the cluster separability for the proposed method is higher than with reference male speaker (See Tab 2.2), the residual variances for the first three formants, V_n , between male-female clusters after normalization is tabulated only for this case in Table 2.3.

	V_1	V_2	V_3
Uniform	19.76	14.22	4.77
Simple Non-uniform	17.37	11.61	4.67
Fant's Non-uniform	15.93	11.22	4.51

Table 2.3: Residual variance after normalization.

Prediction error residue between male-female cluster after normalization, with reference female speaker. V_n denotes the residual variance for the n^{th} formant after normalization.

Scatter Plots

The F1-F2 scatter plots of Peterson-Barney data using the various normalization schemes. Better separability using the normalization method proposed can be seen in Fig 2.7.

2.4 Summary

The reasons for scaling of formant patterns of speech signals, and its nature was noted. The non-linearity of the scaling function, was modelled and used in the proposed non-uniform normalization procedure. The proposed method of approximating the non-linearity only as function of frequency was compared with both linear scaling and non-linear scaling with explicitly modelling the non-linear scaling as a function of both vowel category and formant number, as proposed by Fant. The proposed method needs no more information than the uniform scaling method, but performs better, both in terms of F-ratio and residual variance measures, while its performance is slightly less than that of Fant's non-uniform scaling method, which requires additional information.

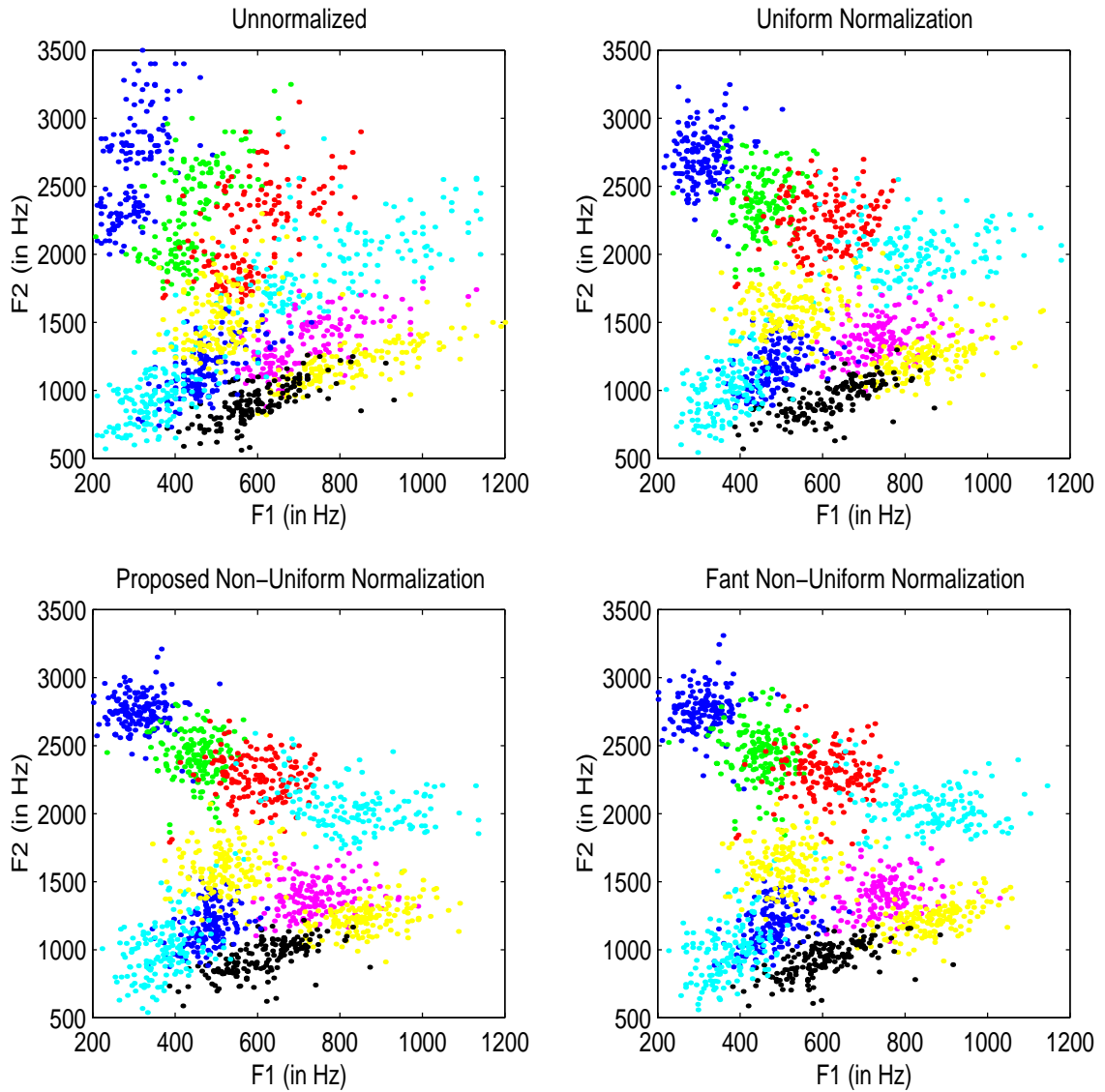


Figure 2.7: Scatter plots of $F_1 - F_2$ for 10 Vowels
Figure shows the Scatter plots of $F_1 - F_2$ for 10 Vowels from Peterson-barney data with and without normalization. As seen in the figure the proposed normalization scheme provides good separability among vowel clusters.

Chapter 3

A Warping Function for Non-uniform Vowel Normalization

It is a commonly accepted concept, that vocal tract length scaling results in scaling of the resultant spectra as a whole. This scaling is assumed to be of the same non-linear behaviour as observed in case of formant patterns. The proposed normalization procedure, in its present form, is applicable only to discrete formant patterns. But the state of the art speech recognizers today make use of continuous spectral patterns, hence for the proposed non-uniform normalization to be implemented on a recognizer/classifier, we need to extend the procedure for continuous spectral patterns [19].

3.1 Non-linear Scaling of Continuous Vowel Spectra

The proposed non-linear scaling was tested on some continuous spectra of vowel data, as a first step in applying it to a recognizer. Spectral pattern alignment for speakers of both the genders and of all ages were tested. For every speaker in the database, his/her Vocal Tract Length (VTL) and hence the scale factor is calculated. This gross scale factor, α , is used for linear scaling of spectra. The non-linear scaling operation uses both the scale factor, α , and the proposed scaling function, $\gamma(f)$ (Eqn 2.6).

3.1.1 Vocal Tract Length Estimation

For VTL estimation experiments, the steady vowel portions are extracted from Hillenbrand database, consisting upto four chosen vowel contexts, spoken in "hVd" context. Twenty speakers five each from adult males, adult females, boys and girls were analysed. Each vowel context chosen represented a vowel category, /IH/ in the front vowel category, /AE/ in open vowels and /UH/ in rounded vowels. Context /AH/ was made use of, for vocal tract length estimation, as it can be approximately modelled well by a uniform tube [7]. It can also be noticed from Fig 2.1 that the distance between the predicted and actual (F_1, F_2) points, is the least for the vowel /AH/, hence supporting the assumption.

All data, sampled at 16KHz, are sectioned with an overlapping window of 20ms frame size and with an overlap of 10ms. A first order backward difference of preemphasis, and Hamming windowing is done. Stationary part of the vowel data frame is subjected to Weighted Overlap Spectral Averaging (WOSA)[20] to compute the smooth spectrum (See Feature Extraction, Sec 4.1).

One idea that has been suggested for computing the VTL is to use higher order formants, since they are mostly affected only by the overall VTL [10]. Here the basic assumption is that, the higher formant frequencies do not deviate *much*, from those of a uniform tube, having the same length. The length l is estimated from the i^{th} formant frequency F_i as,

$$l = \frac{(2i - 1)c}{4F_i} \quad (3.1)$$

where c is the velocity of sound in air. Accuracy of this approach has been tested in [6]. Manually upto five formants are located for each of the twenty speakers, which are then used for their vocal tract length estimation. An average of these lengths, are taken as the actual VTL of the speaker. A scatter plot of the estimated VTL, of all speakers in the chosen database, used for analysis is shown in Fig 3.1.

3.1.2 Experiments and Results

The implementation details of linear and non-linear scaling of vowel spectra is presented in this section. The smooth spectrum of the frame considered, is obtained by calculating the Fourier transform of its autocorrelation function, at equi-spaced

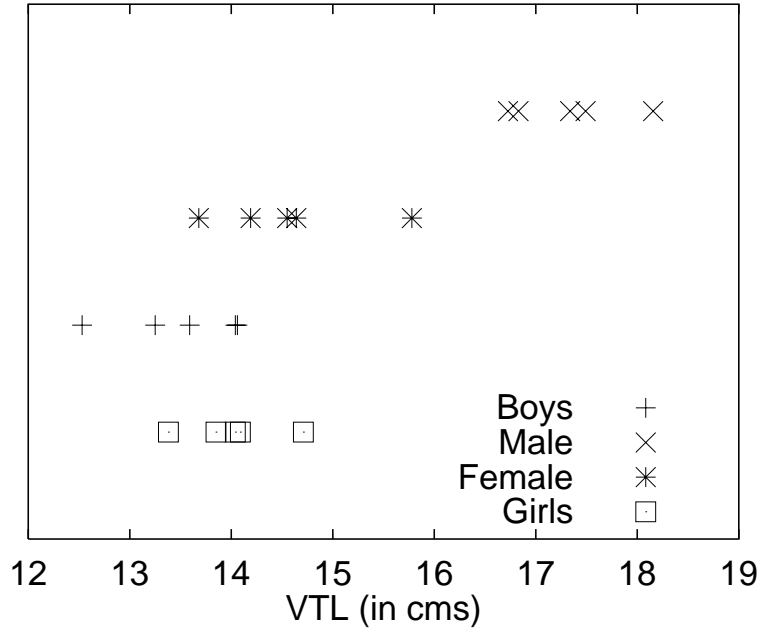


Figure 3.1: Scatter plot of estimated VTL for all twenty speakers used in the analysis. As can be noted on an average, VTL for men = 17.3cm, VTL for women = 14.6cm, VTL for boys = 13.4cm and VTL for girls=13.8cm.

frequencies given by the sampling grid \underline{F} . Essentially \underline{F} is a 500 point array of equi-spaced frequency values ranging from 100Hz to 7000Hz. The scale factor α for a given speaker, is computed by taking the ratio of the subjects estimated vocal tract length to that of the average male (reference) vocal tract length (See Eqn 2.1). Linear scaling, is now effected by computing the spectrum at the frequencies given by, the scaled sampling grid, $\alpha \underline{F}$. In the case of non-linear scaling, α is no longer a fixed number, but is an array, whose values are computed at frequencies given by \underline{F} . Non-linear scaling, is effected by computing the spectrum values at frequencies given by

$$\underline{F}' = \underline{\alpha} .* \underline{F} \quad (3.2)$$

where $\underline{\alpha} = 1 + (\gamma(\underline{F})/100)$

(.* denotes point to point array multiplication)

A Results

For no scaling, linear scaling and proposed non-linear scaling cases, the spectral patterns for vowels in different vowel categories are shown in Fig 3.2, Fig 3.3 & Fig 3.4. It can be noticed, that for all cases non-linear scaling aligns the spectral patterns, from different speakers, much more effectively than the linear scaling.

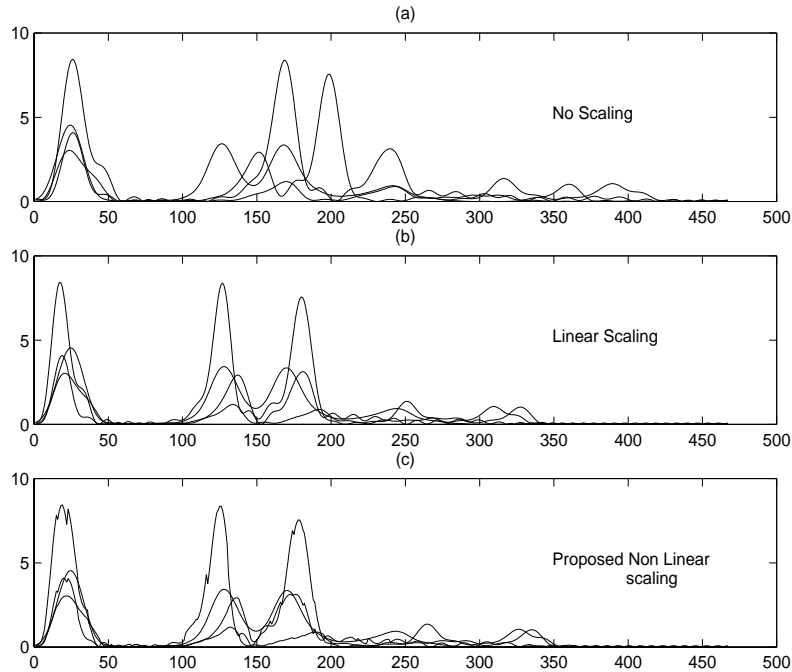


Figure 3.2: Various scaling functions as applied to the front vowel /IH/. Figure shows the spectral patterns for male#5, female#5, boy#2 and girl#1 with (a) no scaling (b) linear scaling by their respective α 's (c) proposed non-linear scaling. Abscissa represents equi-spaced samples.

Gross linear scaling accounts for most of the VTL induced variabilities. The proposed non-linear scaling function, Eqn 3.2, does a finer alignment upon this. This alignment is at its best, for the first formant region, in all the vowel categories. On the whole, for front vowels it performs better than, in other vowel categories. This statement is also supported by its better performance, when applied to /IY/ vowel, for the discrete formant patterns Fig 2.6.

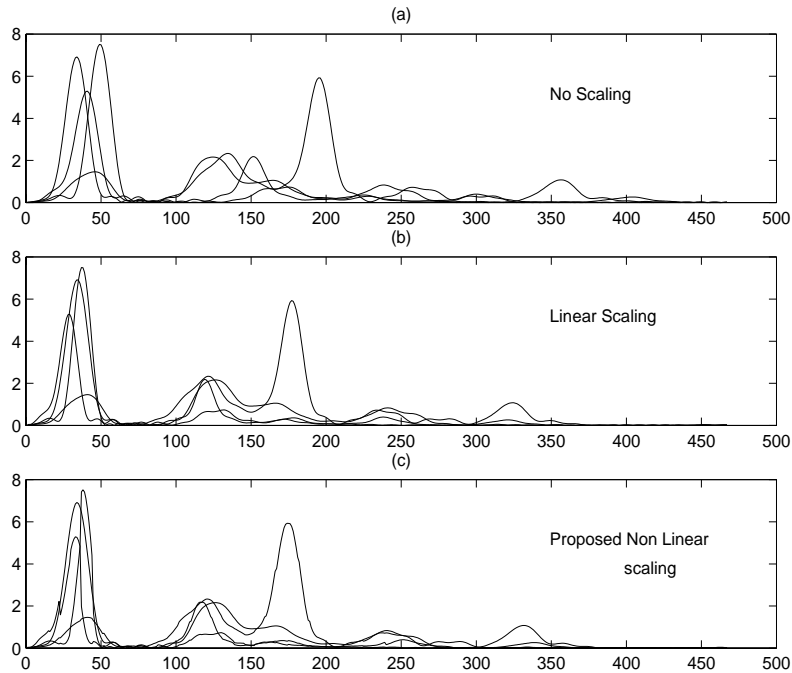


Figure 3.3: Various scaling functions as applied to the open vowel /AE/.
Conditions same as in Fig 3.2

3.2 Approximate Scaling Function

For the above experiments the scaling function used was the one from Eqn 2.6 [17]. A piecewise approximation to this function can be made, as shown in Fig 3.5. Since there are very few data points, at some of the low frequencies, only those frequency bands that have large number of data points are used, in deriving the piecewise approximation. The spectral alignment plots (for the vowel /IH/, for the same set of speakers), using this approximate scaling function, for the non-linear scaling, are shown in Fig 3.6. Once again it can be noticed that non-linear scaling with this approximate scaling function still performs better than the linear scaling.

3.3 Warping-Function

Given the superiority of non-linear scaling, it would be of interest to model this non-linearity [20][12][21][22]. Approximating the scaling function ($k_{nf} \rightarrow \gamma(f)$) as a

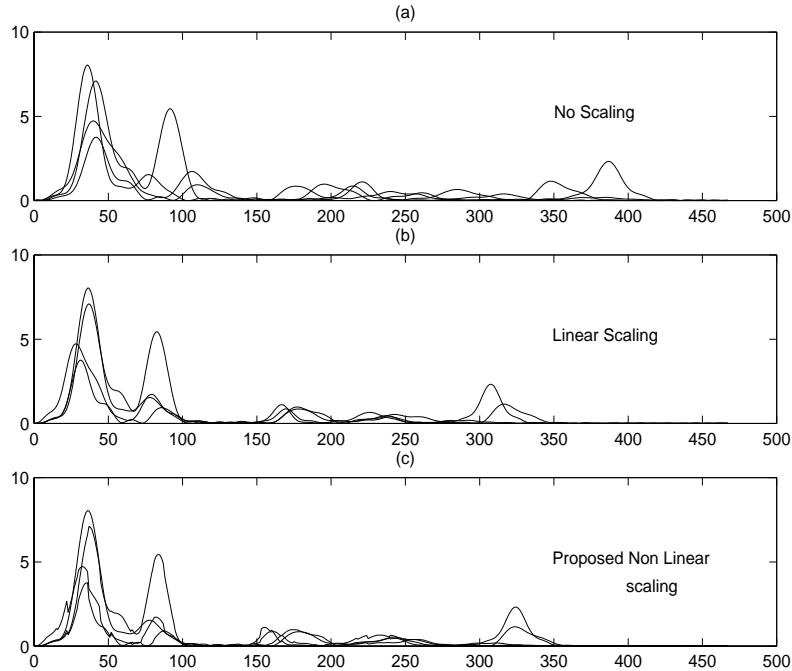


Figure 3.4: Various scaling functions as applied to the rounded vowel /UH/.
Conditions same as in Fig 3.2

function of frequency alone, is a step in this direction. With the proposed non-linear scaling function 2.6 normalization can be done without context dependence, but we still need to estimate the exact scale factor for the speaker. Here we need to address certain questions like: (1) Is there any method of utilizing this knowledge without estimating α ? (2) Now let us suppose we have the complete knowledge about the exact scaling function that *exists*, then can we come up with a universal warping function, which leads to scale invariance? [20]. Such a warping function would be of great value in deriving speaker independent robust features. Motivated by such questions a warping function for the proposed non-linear scaling function is derived. In the analysis, consider a function g , which transforms the frequency axis as $f' = g(\alpha, f)$ where α is the speaker's scale factor. The non-linearity in scaling is modelled similar to earlier works [20][21] as,

$$f' = g(\alpha, f) = \alpha(f) f = \alpha^{\beta(f)} f \quad (3.3)$$

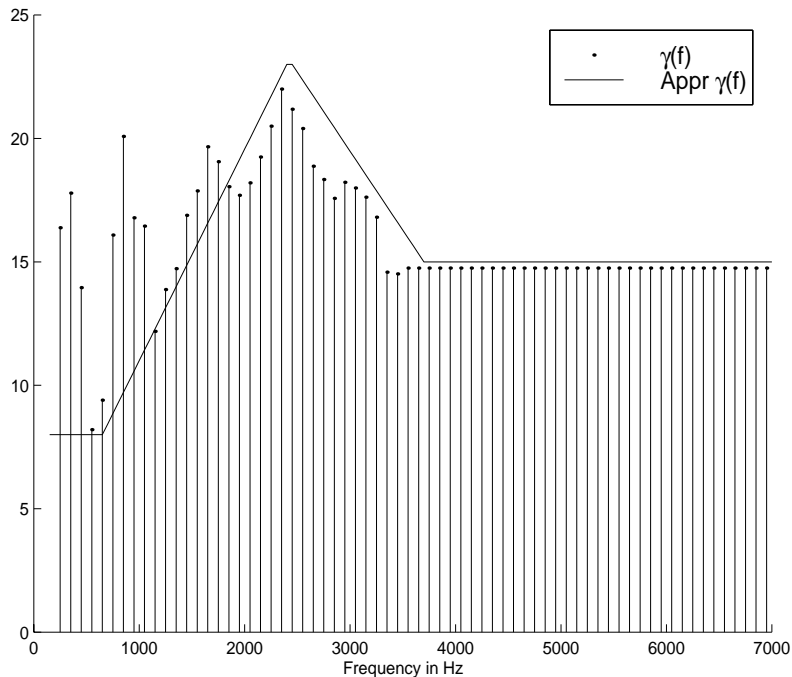


Figure 3.5: Approximate scaling function

Piecewise approximation of the weighting factor as a function of frequency alone.

where α is the subject's scale factor, with respect to a reference speaker, independent of frequency, while $\beta(f)$ depends only on the frequency and is independent of the speaker. The non-linearity of the scale factor is hence captured by $\beta(f)$. Given the model for non-linearity, Eqn 3.3 can be modified, to obtain scale invariance as,

$$\log(f') = \beta(f)\log(\alpha) + \log(f) \quad (3.4)$$

$$\frac{\log(f')}{\beta(f)} = \frac{\log(f)}{\beta(f)} + \log(\alpha) \quad (3.5)$$

$$\nu' = \nu + \text{Constant shift} \quad (3.6)$$

$$\text{where } \nu = W(f) = \frac{\log(f)}{\beta(f)} \quad (3.7)$$

In arriving at this warping function $W(f)$, it has been assumed that $\beta(f') \simeq \beta(f)$. Hence $W(f)$ is a warping function, such that in the warped domain, ν , the spectral patterns between different speakers are approximately translated versions of each

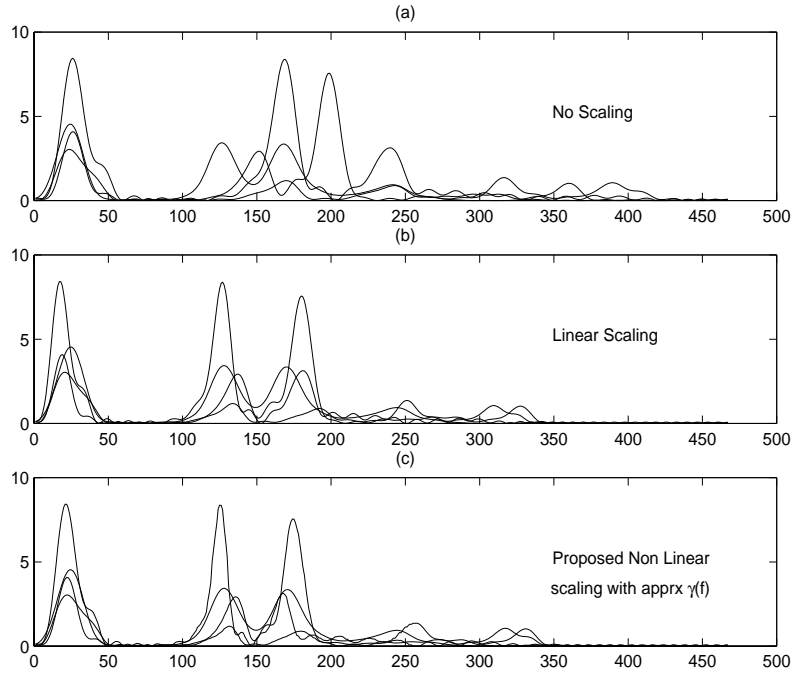


Figure 3.6: Approximate scaling function as applied to the front vowel /IH/ *The non-linear scaling uses the approximate scaling as given by Fig 3.5*

other. The magnitude of the Fourier transform of these warped spectral patterns are invariant to translations, leading to scale invariant features, of real speech signals. For the given model, we can derive the warping function as,

$$\alpha(f) = 1 + \frac{\gamma(f) * (\alpha - 1)}{17} = \alpha^{\beta(f)} \quad (3.8)$$

hence,

$$\beta(f) = \frac{\log(1 + \frac{\gamma(f)*(\alpha-1)}{17})}{\log(\alpha)} \quad (3.9)$$

Eqn 3.9 is valid for all values of α and is invariant to the choice of the reference speaker. The question now is *what value of α should be used?* If an average male speaker is considered as the reference speaker then for an average female speaker its

value is around 1.17. Substituting this value in Eqn 3.9 we have,

$$\beta(f) = \frac{\log(1 + \frac{\gamma(f)}{100})}{\log(1.17)} \quad (3.10)$$

$$W(f) = \frac{\log(f)}{\beta(f)} \quad (3.11)$$

$W(f)$ is the desired warping function. Let us try to find an insight, into the

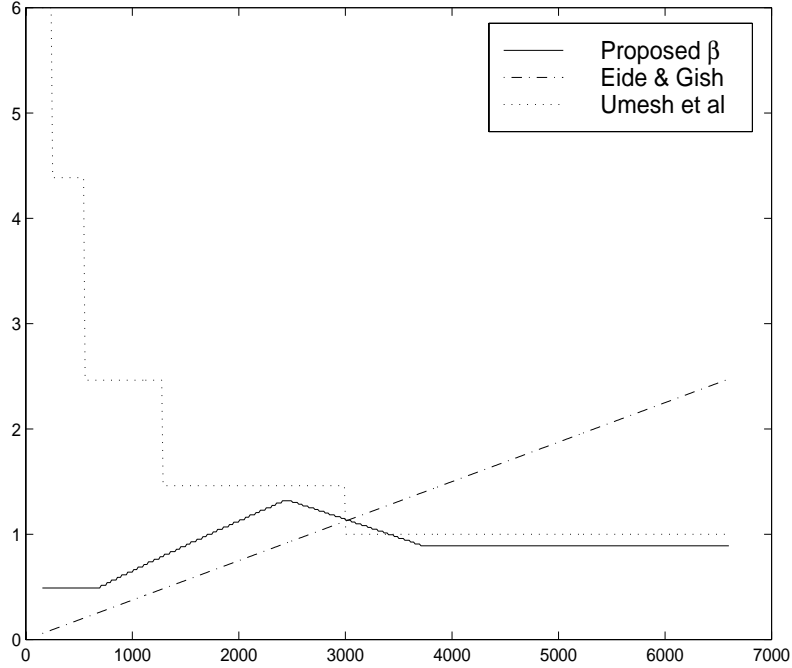


Figure 3.7: $\beta(f)$ for non-linear scaling schemes

$\beta(f)$ as given by Eqn 3.10 along with other scaling functions (See Sec 3.3) are plotted, f is in Hz.

physiological relevance, of the proposed $\beta(f)$ function. We know, from the studies of Fant [7], that the Helmholtz resonator (two tube), is a good approximation of the vocal tract, for first formant (low frequency) region, of the front vowels like /IY/. It has been noted that, due to a change in the vocal tract size by a factor of α , the first formant for such a vowel (with low F_1) gets scaled by a factor of $\sqrt{\alpha}$ instead of α . An interesting observation can be made from the Fig 3.7, in which the derived beta function value is around 0.5 in the low frequency region. Implying

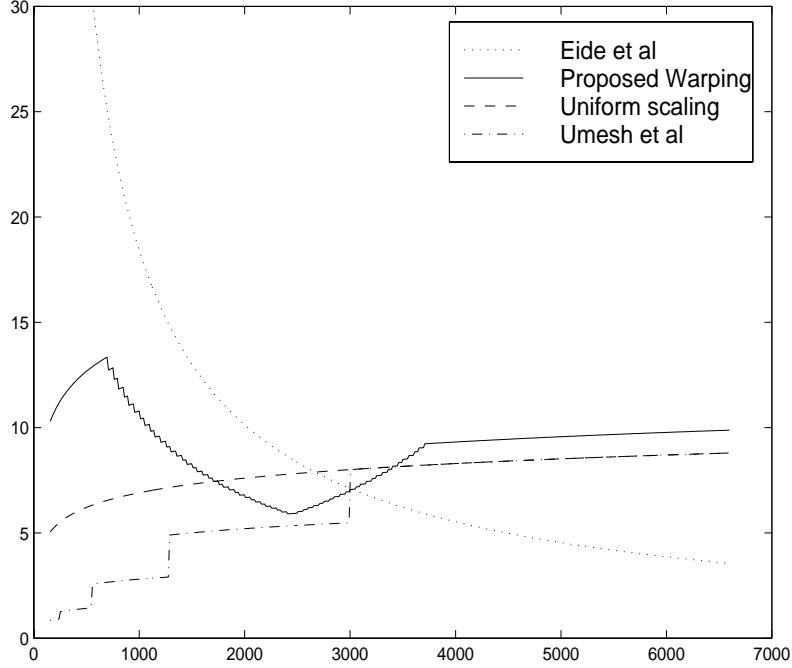


Figure 3.8: Warping functions for various non-linear scaling schemes $W(f)$ as given by Eqn 3.11 along with various other scaling functions (See Eqn 3.12) are plotted, f is in Hz.

that the effective scale factor would be $\sqrt{\alpha}$ as it should have been ideally. The scale factor α for a speaker is calculated from the measurements made in the high frequency (higher formant) region hence its intuitive that $\beta(f)$ should approach unity at high frequencies (beyond 3500Hz). In order to study its relation with the existing approaches, which use similar models for non-linear scaling, the warping functions from their proposed $\beta(f)$ functions are derived, which are shown in Fig 3.7. For these various scaling schemes, their equivalent warping functions are given below. A subjective comparison of these functions is shown in Fig 3.8.

$$\begin{aligned}
 \text{Uniform scaling} & : W(f) = \log(f) \\
 \text{Eide \& Gish} & : W(f) = \log(f) / \left(\frac{3f}{8000}\right) \\
 \text{Umesh et al} & : W(f) = \frac{\log(f)}{\beta_i}
 \end{aligned} \tag{3.12}$$

[100,240)	[240,550)	[550,1280)	[1280,3000)	[3000,7000)
6.0	4.3869	2.4629	1.4616	1

Table 3.1: β values for five logarithmically equi-spaced frequency regions, As used by Umesh et al [20]

where β_i is a piecewise function of frequency as given in Tab 3.1.

3.4 Summary

A non-uniform vowel normalization procedure, which accomplishes better alignment of spectral patterns for various vowels, between adult and child speakers, was presented. It was found to be more effective for front vowels than other vowel categories. A warping function aimed at removing inter speaker differences, due to VTL variations, was also derived from the proposed non-linear scaling procedure. Its physiological relevance and its relationship with other existing warping functions were presented.

Chapter 4

Recognizers With Scale Factor Estimation

Non-uniform normalization can be implemented on a HMM-based recognizer in many ways. All these approaches attempt to "normalize" the parametric representation of the speech signal, with the intention of reducing inter-speaker differences caused by vocal tract length variations. There are two broad approaches to feature based speaker normalization (1) The first approach is to directly estimate the "gross scale factor α " either by maximum likelihood (ML) method [5] or by formant estimations (physiological motivations) from the speech data [22]. (2) The second type of systems use a suitable scale invariant transformation, so that there is no need for explicit " α " estimation. Both of these systems can implement either linear scaling or non-linear scaling. Hence the speaker normalization techniques mainly vary in aspects like (1) need for estimating the "gross scale factor α ", (2) method of estimating α (if needed), (3) model for scaling function (linear/non-linear). In this thesis both the types of recognizers i.e., the one which estimate α explicitly in ML sense and the one which uses a scale-invariant transformation, incorporating the proposed non-linear scaling function have been implemented and analysed. In this chapter recognizers which estimate the scale factor in ML sense have been used to implement the proposed non-linear scaling function $\gamma(f)$ (Eqn 2.6).

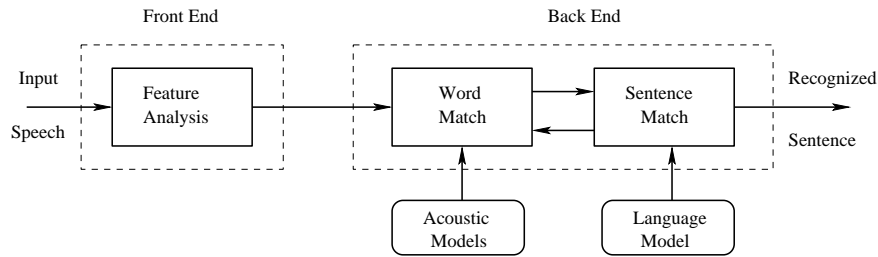


Figure 4.1: Block diagram of a continuous speech recognizer.

The block diagram depicts the various stages in a continuous speech automatic speech recognizer.

4.1 Hidden Markov Model Based Speech Recognizer

Automatic speech recognizer is a system which allows the computer to recognize the spoken words of a person. Automatic speech recognition (ASR) problem is to find a sequence of words to a given set of acoustic features. A block diagram of the various stages involved in an ASR is shown in Fig 4.1. It clearly involves two stages (1) Feature extraction (2) Pattern recognition.

4.1.1 Feature Extraction

Feature extraction stage is necessary to reduce the dimensionality of the problem and to get a parsimonious representation of the speech signal, where only phonetically relevant information is retained, eliminating unwanted distortions. The scheme for extraction of Mel Frequency Cepstral Coefficients (MFCC) and Weighted Overlap Spectral Average -Mel (WOSA-Mel) features, is depicted in Fig 4.2.

Preemphasis: The characteristics of the vocal tract determine the sound produced by a phoneme. Such characteristics are evidenced in the frequency domain by the location of peaks (formants). A roll-off of 20 dB/decade is observed in the spectral domain of a speech signal. A preemphasis of high frequencies is therefore required to obtain similar amplitude for all formants. A first order FIR filter with a transfer function $H(z) = 1 - 0.97z^{-1}$ is used for the purpose.

Windowing: Traditional methods for spectral estimation are valid only for stationary signals. For speech this is true only over short intervals of time. Hence a short time analysis can be performed by windowing the speech signal with a suitable window like Hamming window.

Spectral analysis: The spectrum of speech signal contains information about the spoken sound and the speaker. Spectral envelope reveals those speech signal features which are mainly due to the shape of the vocal tract. Here we discuss two methods of smoothing employed to remove the effects of pitch.

1. *Filter Bank:* In this Fourier analysis is done on each frame of the speech signal. A bank of filters placed uniformly spaced on the mel scale (motivated by human perceptual studies)[23] average the spectral energies over bands. These filter outputs are taken as samples of smooth spectral envelope.
2. *WOSA:* This method is similar to averaged periodogram technique [24] In this method each frame of speech is segmented into L overlapping subframes, and each subframe is Hamming windowed. In this work a subframe of 64 samples each with an overlap of 45 samples was used. An estimate of the autocorrelation for each subframe is obtained and averaged over L subframes. This averaged autocorrelation estimate is used to compute the smooth-spectrum, by Fourier analysis. This method effectively removes the effect of pitch, since the duration of each subframe is less than pitch-interval.

Cepstral features: Logarithm of the magnitude of these samples of smooth spectral envelope are subjected for inverse Discrete Cosine Transform (DCT). The DCT has the property to produce highly decorrelated features [25]. The zero order MFCC coefficient is approximately equivalent to the log energy of the frame. This is discarded and normalized energy of the frame is used in its place. Cepstral mean subtraction is found to compensate for channel variations [26].

Temporal Cepstral derivatives: The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given analysis frame. An improved representation can be obtained by extending the analysis to include information about the temporal cepstral derivative [27]. A simple first and second order difference is used as an approximate (and noisy) estimate of the cepstral derivative.

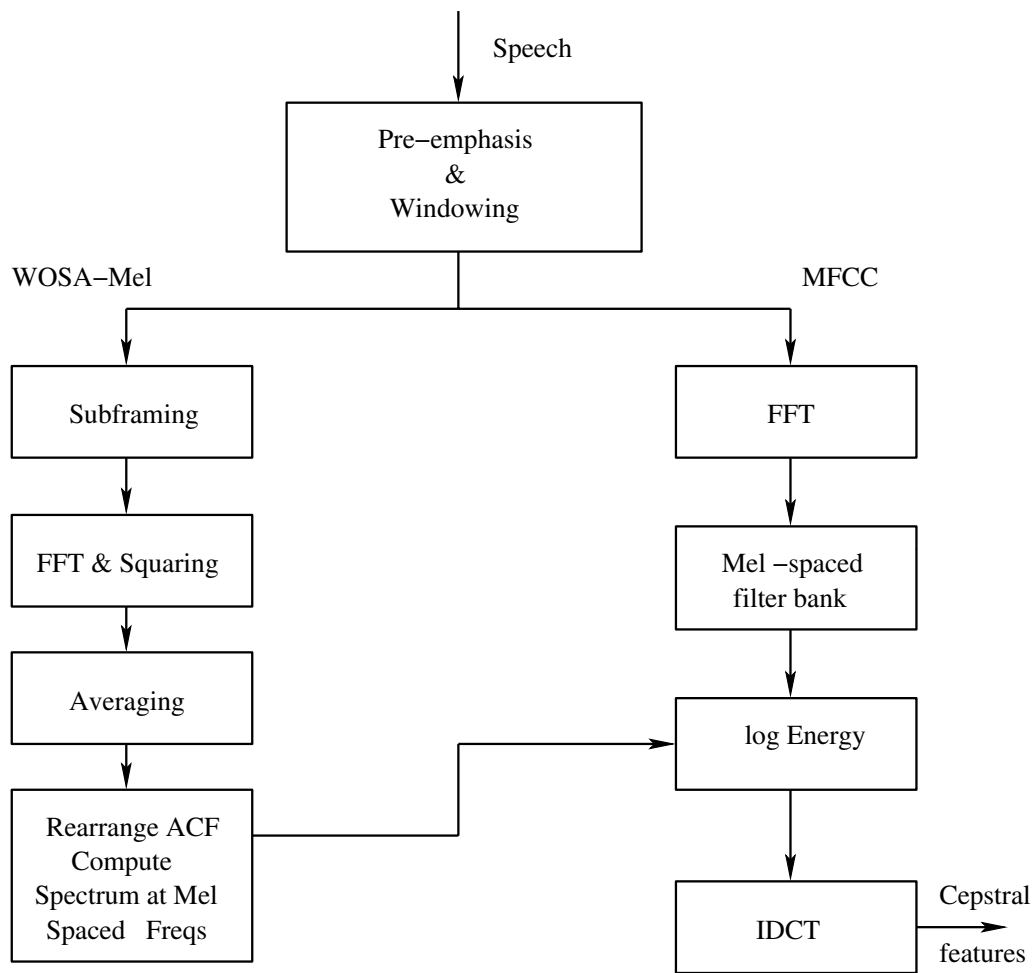


Figure 4.2: Feature Extraction stages.

The block diagram depicts clearly the various stages involved in spectral analysis methods. Note ACF denotes autocorrelation function

4.1.2 Pattern Recognition

The pattern recognition problem for ASR can be solved in any of the three paradigms (1) Vector quantization (2) Hidden Markov Modelling (3) Artificial Neural Network. Speech recognition is associated with lot of uncertainties due to different variabilities (like speaker, channel noise, etc.) Stochastic modelling is a flexible method for accounting such variabilities. One of the major advantages of using HMMs in speech recognition problem is their ability to provide a uniform framework for stochastic representation of both acoustic and lexicon rules, along with other sources of knowledge. A very brief introduction to the usage of HMMs in ASR is given below. Eminent works can be referred to for more fundamental details on their usage in ASR [28] [29].

HMMs constitute a “doubly stochastic process” in which the observed data is modelled having generated by a piecewise stationary process. HMMs can be used to model a specific unit of speech such as a sentence, a word, a subword or a phone. HMMs are characterized by the number of states, the transition probabilities among them and the output probability distribution associated with each state. These probability density functions could be either discrete or continuous. Discrete density functions are modelled by a set of discrete point probabilities for each of the vectors in the codebook obtained by Vector quantizing the observation. To avoid quantization errors often observations are modelled directly with a mixture of Gaussians resulting in Continuous density function.

Training procedure involves optimizing HMM parameters given an ensemble of training data. Using segmental k-means algorithm the HMM parameters are initialized suitably. Baum-Welch method is used to iteratively estimate the transition probabilities and state probability density function parameters [30].

With embedding of both acoustic and linguistic knowledge we get a huge extended HMM network of states. The goal of decoding process in HMMs is to determine the maximum likelihood sequence of states which has generated the observed signal. An exhaustive search through this network of states is unfeasible for any realistic recognition problem. Hence Viterbi-like sub-optimal algorithms are adopted for efficient searches. The basic idea in such algorithms is simple, i.e. at any discrete time t , all the probabilities of the “hypotheses being in any *admissible* state” are computed. At the end only the more likely sequences are selected. Hence

avoiding exhaustive search.

With the availability of such efficient training and search algorithms HMMs provide a unified platform for statistical modelling of the ASR problem.

4.2 Non-uniform Normalization on the Recognizer

As it has been noted earlier vocal tract size variations which result in scaling of the frequency spectrum of speech signals, account for a major portion of the inter-speaker variations. Hence it is intuitive to normalize the frequency spectrum of each speaker, with proper estimation of the scaling factor. Speaker's scale factor can be estimated by linking the articulatory variations to spectral parameters. Such estimated scale factor, can then be plugged-in to scale the frequency spectrum of the speech signal suitably by either linearly or non-linearly [22][6]. A second class of systems differ, in the way the scale factor is estimated for the speaker. In this class of recognizers, for every speaker in the training set an optimal scaling factor (in ML sense), $\hat{\alpha}$, is estimated which is then used for warping the utterance. All of the warped utterances are used to build a "normalized" HMM. Similarly, during recognition, $\hat{\alpha}$ is estimated for every input speech, which is then used to scale the speech utterance. Warped utterance is subjected to decoding on the normalized HMM.

4.2.1 Scaling Factor Estimation in ML Sense

The scaling factor, α , represents the ratio between a speaker's vocal tract length and some notion of reference vocal tract length. However it is very difficult to reliably estimate one such factor from the acoustic data, especially in the absence of that "golden reference speaker". In the maximum likelihood method of estimating the scaling factor, the reference speaker notion is served by a reference HMM-model. The goal is to choose an α for an utterance, such that its likelihood for the given model is maximized.

$$\hat{\alpha} = \arg \max_{\alpha} Pr(X_i^{\alpha} | \lambda, W_i) \quad (4.1)$$

where X_i^{α} : The warped feature set of the i^{th} utterance.

λ : Reference HMM model.

W_i : The transcription of the i^{th} utterance.

The optimum scaling factor is obtained by searching over a grid of 13 factors spaced evenly between $0.88 \leq \alpha \leq 1.12$ for adults, $0.76 \leq \alpha \leq 1.0$ for children. Reflecting the range of 25% variation of VTLs among males & females and about 36% between children and adults.

4.2.2 Training and Testing Procedure

It is clear from the algorithm that the scaling factor estimation process requires a pre-existing HMM model. Therefore, an iterative procedure is used to choose the best scaling factor for each speaker and then build a model using the warped training utterances.

Training Procedure:

1. Divide the whole training database into two halves.
2. Build an “unnormalized” model (λ_T) using one half of the database.
3. Now for every utterance in the second half choose an “ α ” such that $Pr(X_i^\alpha | \lambda_T, W_i)$ is maximized, where X_i^α is warped feature set using the linear/non-linear warping function.
4. Now these two sets are swapped and the above procedure is repeated iteratively. until there is no significant change in α values.
5. Build a “normalized” model (λ_N) using all the warped utterances in the training set.

Here it is assumed that phonetic transcription for the training set is available. If this is not true then a first pass of decoding is required to obtain the time alignment of the phonemes in the utterances, which are then used to stack the models and hence estimate α . This first pass of decoding is a one time process for every utterance in the training set. During recognition, the goal is to scale the frequency axis of each test utterance to "match" that of the normalized HMM model λ .

Testing Procedure:

1. The unwarped utterance X_i and the normalized model λ_N are used to obtain preliminary transcription of the utterance. Let the transcription obtained from the unwarped feature set for the i^{th} utterance be denoted as W_i
2. $\hat{\alpha} = \arg \max_{\alpha} Pr(X_i^{\alpha} | \lambda, W_i)$ with linear/non-linear warping function.
3. The utterance $X_i^{\hat{\alpha}}$ is decoded with the model λ_N to obtain the final recognition result.

A block diagram explaining the various stages involved in recognition with speaker normalization used is shown in Fig 4.3

4.2.3 Non-Linear Scaling With Filterbank Analysis

In the previous section the process of scale factor estimation and its usage in HMM training and testing has been explained. In this section an efficient method to implement both linear and non-linear scaling function in a recognition system whose features are extracted from a filter-bank is explained. The standard Davis-Mermelstein [23] filterbank frontend is used to derive Mel frequency complex cepstrum (MFCC) features, for the HMM-based recognizer. It works by first calculating the magnitude spectrum of the windowed speech by passing it through a mel-scale filterbank and finally taking the inverse cosine transform to arrive at the cepstrum. Though linear scaling can be implemented as a simple resampling of the speech signal in time domain, but it is more efficient to push the process onto the filterbank front end itself. Moreover time domain resampling is difficult to conceptualize for non-linear scaling implementation. Frequency scaling can be implemented by simply varying varying the spacing and width of the component filters of the filterbank, without changing the original speech signal [11]. This process is depicted in Fig 4.4. For example to compress the speech signal in the frequency domain, the frequency scale of the signal is kept the same but stretch the frequency scale of the filters. Warping is implemented by dividing the filter boundaries with a suitable alpha given by,

$$\begin{aligned} F_l^{(i)'} &= F_l(i)/\alpha(F_l(i)) \\ F_h^{(i)'} &= F_h(i)/\alpha(F_h(i)) \end{aligned} \tag{4.2}$$

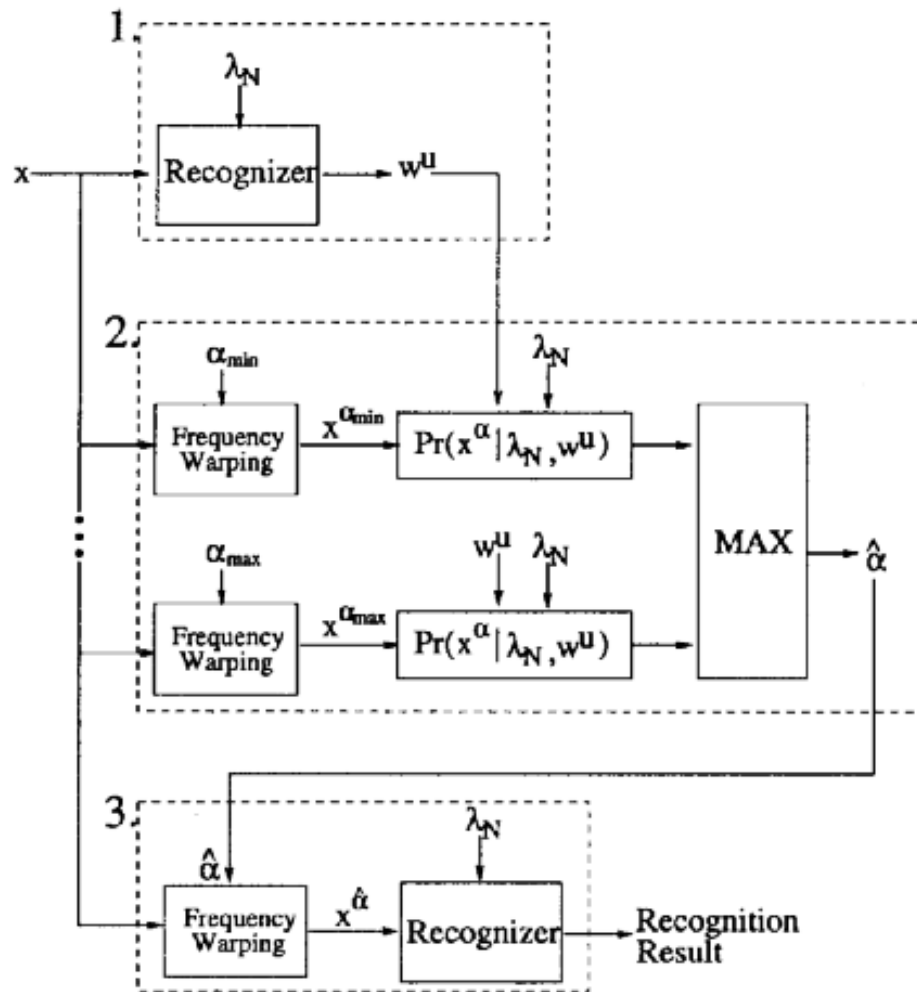


Figure 4.3: Recognition with speaker normalization.

The block diagram depicts clearly the various stages involved in recognition with the speaker normalization used [11]

For linear warping $\alpha(f) = \alpha$ for all $f \in [270, 3850]$

For non-linear warping $\alpha(f) = \frac{(\alpha-1)*\gamma(f)}{17} + 1$

where $F_l^{(i)}$ & $F_h^{(i)}$ are the low and high end boundaries of the i^{th} component filter in the filterbank respectively. α denotes the speaker's gross scale factor estimated and $\gamma(f_o)$ is the approximate weighting function value (Fig 3.5) in the vicinity of f_o . Tab 4.1 gives the Mel spaced triangular filter boundaries with no warping, and the corresponding $\gamma(f)/17$ values.

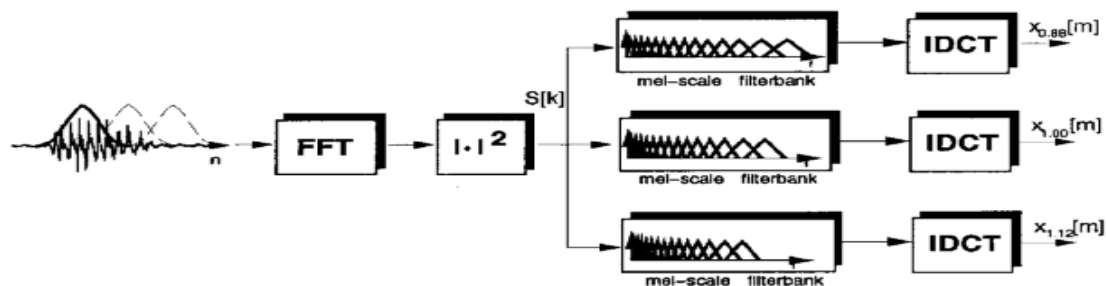


Figure 4.4: Mel filterbank analysis with frequency warping.

index	f in Hz	$\gamma(f)/17$	index	f in Hz	$\gamma(f)/17$	index	f in Hz	$\gamma(f)/17$
1	187.5	0.5294	9	1000	0.828	17	2156.3	1.47
2	312.5	0.5294	10	1093.8	0.909	18	2343.8	1.41
3	406.25	0.5294	11	1218.8	0.989	19	2593.8	1.29
4	500	0.5294	12	1312.5	1.069	20	2843.8	1.17
5	593.75	0.5294	13	1468.8	1.149	21	3125	1.058
6	687.5	0.588	14	1593.8	1.23	22	3437.5	0.941
7	812.5	0.668	15	1781.3	1.31	23	3781.3	0.7497
8	906.25	0.748	16	1937.5	1.39			

Table 4.1: Mel filter boundaries and $\gamma(f)$

Mel-spaced triangular filter boundaries and the weighting values used

4.2.4 Experiments and Results

This section presents an account of the experiments done to investigate the effectiveness of speaker normalization procedure proposed. Speech recognition accuracy is used a performance measure for both linear and non-linear scaling functions for speaker normalization.

A Tasks and Databases

Two telephone based continuous digit databases one from "30K Numbers Corpus" Oregon Graduate Institute (OGI) and the other from AT&T were used in these experiments. The size of the vocabulary was eleven words: "one to nine" "zero" and "oh". The training utterances were endpointed. All of the training dataset was hand labelled, with basic phoneme labels (ARPABET). Training set contained around 6640 utterances. Testing database was made up of two parts one had utterances purely from adults while the second part had utterances from children of age ranging from ten to seventeen. Testing database was never exposed to models during any part of the training. Testing dataset had around 726 utterances with a total of 3517 words from adults, and 800 utterances with a total of 2834 words from children. Number of words in each utterance varied from three to ten. After decoding the number of substitution errors(S), deletion errors (D) and insertion errors (I) can be calculated. Percent accuracy and percent correct defined as,

$$\text{Percent correct} = \frac{N - D - S}{N} \times 100\% \quad (4.3)$$

$$\text{Percent accuracy} = \frac{N - D - S - I}{N} \times 100\% \quad (4.4)$$

are used to evaluate the performance of various techniques.

B Baseline Speech Recognizer

The experiments are done using monophone HMM based speech recognition system adapted from RES [31] system. The peripheral modules like Mel-filterbank analysis, warping function implementation, training and testing modules, WOSA analysis, database interface module (OGI- NIST format), label interface module (Arpabet), feature file interface module, cepstral mean subtraction (CMS) module and finally

Testset	Adults	Children
Baseline	(93.54, 91.61)	(74.06, 70.89)
Linear	(95.28, 93.46)	(83.73, 80.95)
Proposed non-linear	(94.75, 93.00)	(81.23, 78.23)

Table 4.2: Recognition performance

Recognition performance before and after scaling for speaker normalization With scaling factor estimated in ML sense, Performance is given in terms of (% Correct,% Accuracy)

the complete two pass strategy: warping module are added into the basic skeletal recognizer from RES as a part of this work. All the modules were coded in C++.

Each phoneme, including silence, was modelled by three active state continuous density left-to-right HMM's. The observation densities were mixtures of five multivariate Gaussian distributions with diagonal covariance matrices.

All data were recorded over telephone set, sampled at 8KHz. Speech signals are sectioned with an overlapping window of 20ms frame size (160 samples) and with an overlap of 10ms. A first order backward difference of pre-emphasis, and Hamming window is done. For each data frame 256 point FFT is taken for Mel-filter bank analysis. Thirty-nine dimensional feature vectors were used: normalized energy, c[1]-c[12] cepstra derived from mel-spaced filterbank of twenty one filters and their first and second order differences.

C Speech Recognition Performance

Tab 4.2 shows the recognition performance on two test sets one for adults and the other for children. Results show that (1) the linear frequency scaling provides substantial improvement over the baseline, the improvement is pronounced for children case (2) Contrary to hypothesis, non-linear frequency scaling is not able to improve upon linear scaling and needs further investigation.

4.3 Summary

A brief overview of the HMM-based ASR was presented. The implementation of one such recognizer for digit recognition task was discussed. Methods to incorporate the proposed non-linear scaling function from previous chapters were presented. Performance of the recognizer over the baseline was compared and analysed. The failure of the non-linear scaling functions to provide substantial improvements over the performance of the recognizer with linear scaling can have many reasons. The proposed non-linear scaling function is motivated from the deviations of the scaling function behaviour from linear scaling in vowels. This background is not identical to the condition in which it is being used in a digit recognizer. Further investigation is needed to really get the advantages of non-linear scaling, in terms of improvements in recognition performance, on HMM-based recognizers with scale factor estimation.

Chapter 5

Recognizers With Non-Linear Scale Invariance

In the Chapter 4 a recognizer with explicit scale factor estimation was explained. The non-linear scale function model derived in Chapter 2 (Eqn: 2.6) could be incorporated in a straight forward manner. The main disadvantage of this method is the need to estimate one such scale factor for every utterance, increasing complexity and introducing estimation errors. There is an another class of recognizers which work on the principle of applying suitable scale-invariant transformation on the speech spectra [4]. The basic idea is to warp a pair of mutually scaled spectra such that in the warped domain they are shifted versions of each other. By taking the magnitude of the Fourier transform of these shifted functions we get identical coefficients. This method is explained in detail in the following sections. By suitably warping the spectra even non-linear scaling can be taken care off [20]. Hence the proposed non-linear scaling function is modified so that it can be incorporated in such a non-linear scale invariant transformation.

5.1 Non-linear Scale Invariance on a Recognizer

This class of recognizers aim to reduce inter-speaker differences by suitably transforming the smooth spectra. Here the transformations used are linear/non-linear scale invariant [20]. If the linear scaling hypothesis is true then the spectra of two

speakers are related by

$$S_A(\omega) = S_B(\alpha_{AB}\omega)$$

By exponentially sampling the spectra

$$S_A(e^\omega) = S_B(e^{\omega + \ln\alpha_{AB}})$$

They become shifted functions in the warped domain

$$S_A(\nu) = S_B(\nu + \ln\alpha_{AB})$$

The magnitude of their Fourier transform leads to scale invariance,

$$|\mathcal{F}(S_A(\nu))| = |\mathcal{F}(S_B(\nu + \ln\alpha_{AB}))|$$

Here exponential sampling denotes linear scaling of the frequency axis, which is realized as equal sampling in log domain. Fig 5.1 shows the above process pictorially. Two linearly scaled signals after exponential sampling (Log warped) become shifted versions of each other.

Non-linear warping is realized by sampling unequally in logarithmically equi-spaced bands. Non-linearly scaled functions should turn out to be shifted versions in such a non-linearly warped domain.

The main advantage of scale-invariant transformation is the reduced computation as compared to systems which explicitly scale estimates the scale factor and also provide ease of incorporating non-linearity in scaling function.

5.1.1 Non-Linear Scale Invariant Transformation

As noted earlier exponential sampling of the smooth spectrum leads to linear scaled functions being shifted versions of each other. Exponential sampling is nothing but equi-spaced sampling in log domain. By changing the number of samples in such bands we can implement the non-linear warping function. Given the model for non-linearity of the scaling function $\alpha(f\epsilon B_i) = \alpha^{(i)} = \alpha^{\beta_i}$ (Eqn 3.3), we have the warping function over N logarithmically equi-spaced frequency bands as, $W(f\epsilon B_i) = \frac{\log(f\epsilon B_i)}{\beta_i}$. Here β_i depends only on the i^{th} frequency band, B_i , but α depends on the pair of speakers considered. To linearly warp one speaker with respect to other, we

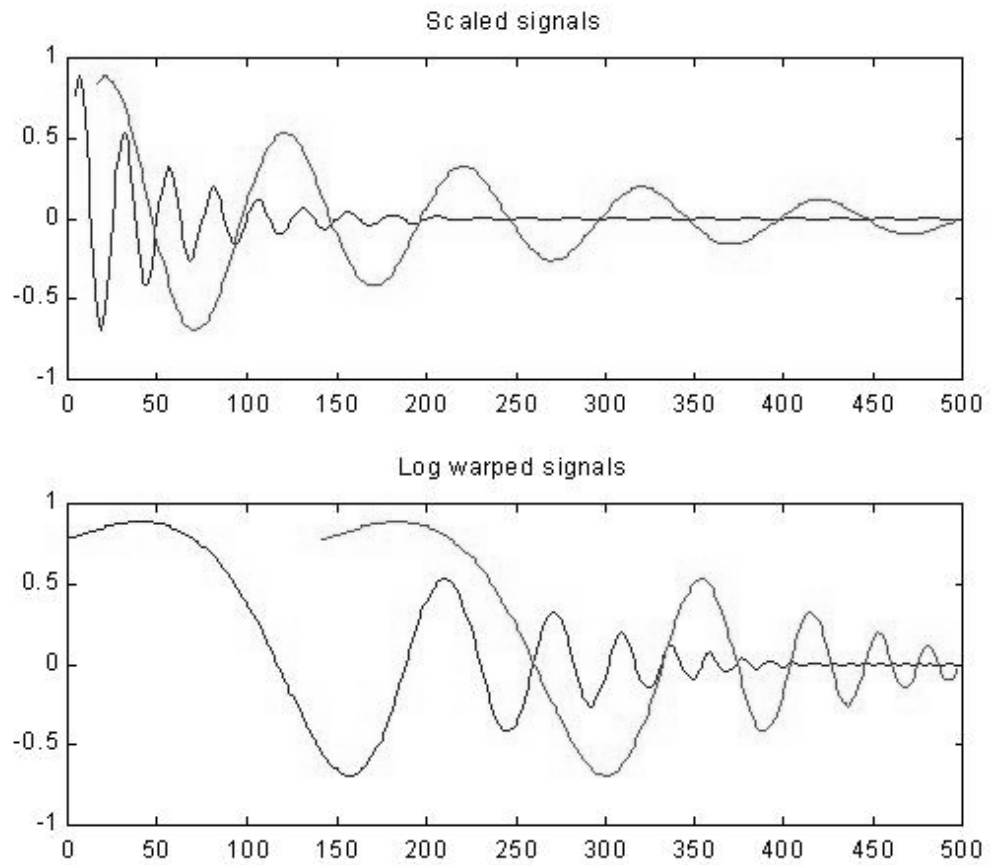


Figure 5.1: Log warping of two scaled signals.

Two linearly scaled signals after exponential sampling become shifted versions of each other. Magnitude of the Fourier transform these signals are identical (See Sec 5.1).

need to compute $B(e^f)$ for $e^f \in [U_i, L_i]$, where U_i and L_i are upper and lower frequency limits of i^{th} frequency band. In the discrete implementation of the warping function $B(e^f)$ is computed at M_i equally spaced intervals in the region $\log(L_i)$ to $\log(U_i)$. Now the problem is to find a suitable set of M_i 's for a given β_i set (See [20] for details). Let,

$$\Delta\nu_i = \frac{\log(U_i) - \log(L_i)}{M_i} \quad (5.1)$$

be the spacing in the i^{th} log-frequency band. Then, the exponentially spaced samples in the i^{th} frequency region are $B(e^{m_i\Delta\nu_i + \log(L_i)})$ for $m_i = 0, 1, 2, \dots, (M_i - 1)$. If the scaling between two spectra is given by $S_A(f) = S_B(\alpha_{AB}^{(i)} f)$ where $\alpha_{AB}^{(i)}$ be the scaling factor in the i^{th} frequency band. By exponentially sampling $A(f)$ we have,

$$A(e^{m_i\Delta\nu_i + \log(L_i)}) = B(e^{m_i\Delta\nu_i + \log(\alpha_{AB}^{(i)}) + \log(L_i)})$$

$$m_i = 0, 1, \dots, (M_i - 1) \quad (5.2)$$

Hence in the warped domain they form the piece-wise shifted versions of each other,

$$A[m_i] = B[m_i + \frac{\log(\alpha_{AB}^{(i)})}{\Delta\nu_i}] \quad (5.3)$$

With the shift in the i^{th} frequency band being,

$$\frac{\log(\alpha_{AB}^{(i)})}{\Delta\nu_i} = \frac{\log(\alpha_{AB}^{\beta_i})}{\Delta\nu_i} = \frac{\beta_i \log(\alpha_{AB})}{\Delta\nu_i} \quad (5.4)$$

The condition for the shift to be equal in all these N logarithmically equi-spaced regions is [20],

$$\beta_i M_i = \beta_j M_j$$

$$\text{for } i, j = 0, 1, \dots, (N - 1). \quad (5.5)$$

The total number of samples held constant, M_i 's are given by,

$$\sum_{i=0}^{N-1} M_i = M_{const} \quad (5.6)$$

$$M_i = \frac{M_{const}}{\beta_i} \quad (5.7)$$

$$\sum_{j=0}^{N-1} (1/\beta_j)$$

Band(Hz)	M_i	Band(Hz)	M_i
[270, 376)	7	[1019, 1421)	8
[376, 524)	11	[1421, 1981)	6
[524, 731)	9	[1981, 2761)	7
[731, 1019)	8	[2761, 3850)	7

Table 5.1: Implementation of proposed eight band non-linear frequency warping. M_i denotes the number of samples in logarithmically equi-spaced bands.

Band(Hz)	M_i	Band(Hz)	M_i
[270, 322)	3	[1114, 1330)	5
[322, 384)	4	[1330, 1587)	4
[384, 459)	5	[1587, 1895)	3
[459, 548)	8	[1895, 2262)	2
[548, 654)	5	[2262, 2701)	3
[654, 781)	4	[2701, 3225)	4
[781, 933)	4	[3225, 3850)	4
[933, 1114)	5		

Table 5.2: Implementation of proposed fifteen band non-linear frequency warping. M_i denotes the number of samples in logarithmically equi-spaced bands.

M_i values so obtained are to be suitably quantized to integer values, conforming to the constraint given by Eqn 5.6. The M_i values used for eight bands and fifteen bands for the proposed non-linear scaling function are derived on similar lines and are tabulated in tables Tab 5.1 and Tab 5.2.

5.1.2 Experiments and Results

The proposed non-linear scaling function is incorporated into the framework of non-linear scale-invariant transformation by deriving the β_i set and hence M_i set from Eqn 3.10 & Eqn 5.7. Effectiveness of the proposed warping function and other

warping functions (Eqn 3.12) are quantified in terms of the speech recognition performance for both matched(adults) and unmatched(children) cases.

A Tasks and Databases

The task and databases are identical to the one used in the previous Chapter (See Tasks and Databases). The training set has around 11000 utterances, with the testset having 846 utterances (3700 words) from adults and 800 utterances (2700 words) from children. A piecewise approximation to the proposed non-linear scaling function with eight and fifteen bands are used for analysis.

B Baseline Speech Recognizer

The experiments are done using monophone HMM based speech recognition system adapted from HTK [32]. Warping function implementation & WOSA analysis modules were coded in C, and plugged into the basic recognizer.

Each phoneme including silence, was modelled by four active states, continuous density left-to-right HMM's. The observation densities were (1) mixtures of five multivariate Gaussian distributions with diagonal covariance matrices (2) Single multivariate Gaussian distribution with full covariance matrix.

For WOSA analysis each data frame is again sectioned into subframes of 64 samples each with an overlap of 45 samples and Hamming windowing is done. 127 point autocorrelation function so obtained is subjected to Fourier analysis to obtain a smooth spectrum estimate. Thirty-nine dimensional feature vectors were used: normalized energy, $c[1]$ - $c[12]$ cepstra derived from 64 samples equi-spaced in the warped domain from WOSA analysis (See Feature Extraction 4.1), and their first and second order differences.

C Speech Recognition Performance

Tab 5.3 shows the recognition performance on two test sets one for adults and the other for children. Results show that (1)Log warping or exponential sampling (linear scaling) performs substantially better in matched case, while Mel warping is of better advantage in mismatched case. (2)Proposed non-linear warping fails to provide

Testset	<i>Adults</i>		<i>Children</i>	
Covariance	Diagonal	Full	Diagonal	Full
Mel warp	(87.89, 85.86)	(92.96, 92.35)	(73.23, 70.36)	(81.32, 79.05)
Log warp	(89.11, 86.23)	(94.37, 93.40)	(72.08, 67.66)	(80.63, 76.72)
Eide et al	(86.81, 83.62)	(91.13, 89.61)	(67.55, 63.03)	(72.08, 68.74)
Umesh et al	(88.39, 86.36)	(93.35, 92.71)	(71.94, 68.42)	(79.81, 77.11)
Proposed non-linear				
with 8 bands	(88.50, 86.36)	(92.90, 92.05)	(70.86, 67.01)	(78.48, 75.96)
with 15 bands	(88.11, 85.89)	(93.35, 92.57)	(71.79, 67.98)	(78.91, 76.39)

Table 5.3: Recognition performance of various non-linear scale invariant transformations *The warping functions (Eqn 3.12) are used to compute the M_i 's in logarithmically equi-spaced bands, Performance is given in terms of (% Correct, % Accuracy)*

substantial improvement over the linear scaling performance on the recognizer. (3) The Eide & Gish [21] falls apart for mismatched case.

5.2 Summary

The basic theory for scale invariant transformation was presented. A method for incorporating non-linear scaling in such a paradigm was also given. For the proposed non-linear scaling function one such scale-invariant transformation was derived and implemented on a HMM-based digit recognizer. Similar transformations were obtained for other non-linear scale functions, which assume similar model for non-linearity. These transformations were applied in parallel and results were tabulated. Results suggest that non-linear scaling needs more in-depth investigation to be meaningfully implemented on a HMM-based recognizer.

Chapter 6

Conclusions

In this thesis we have attempted to exploit the additional information available from classical speech analysis studies regarding the nature of scaling that exists among speakers of different age and gender. This information is suitably combined so that it can be readily incorporated into the state of the art recognizers. The proposed scaling function has been analysed using different methods like formant data analysis, spectral alignments and normalization schemes on a HMM-based recognizer.

When the proposed non-linear scaling function is applied to vowel formant database, we obtain encouraging performance in terms of both F-ratio and residual variance. The proposed vowel normalization procedure provides substantial improvement over linear scaling in reducing the variance of vowel clusters. Spectral alignment plots also show good alignment with the proposed normalization procedure, for the representative cases. Since these plots are subjective in nature we lack objective measures to comment on the effectiveness of the proposed procedure for continuous spectra.

The proposed non-linear scaling function was incorporated into a HMM-based recognizer. From the recognition performance results it can be inferred that even though linear scaling provides substantial improvements, the performance of the non-linear scaling is not much better than linear scaling (infact it slightly hurts the performance). This may be due to many reasons and one of them may be due to the coarse sampling of the frequency axis in filterbank analysis. The differential variations of the filter boundaries between linear and non-linear scaling is marginal

due to this coarse sampling.

In order to overcome this problem, the spectral analysis method was changed from filterbank method to WOSA method. By deriving a non-linear scale invariant transformation and applying it on the WOSA-smooth spectrum, we achieve slightly better results (Percent accuracy is comparable to linear scaling case).

In spite of very promising performance of the proposed non-linear scaling function in formant data analysis, its inability to provide improvements on a HMM-based recognizer suggests further investigations into the suitability of using the present HMM-based recognizers for such normalization methods.

Future Work:

As has been already noted, further investigations need to be done to fully exploit the advantages of non-linear scaling function in a HMM-based recognizer, in a more meaningful manner. The proposed scaling function was the result of combining the information from earlier vowel analysis study by Fant and others. More refined methods of suitably combining these informations can lead to a better scaling function. Though the physiological relevance of the warping function derived was just touched upon, further detail studies can be done in this aspect. With the performance of the proposed non-uniform normalization procedure being promising for vowel data, it may be worth building a vowel classifier incorporating the non-linear scaling.

References

- [1] C. J. Leggetter and P. C. Woodland. Flexible Speaker Adaptation for Large Vocabulary Speech Recognition. In *Proc. Eurospeech 95*, pages 1155–1158, 1995.
- [2] T. Anastasakos, J. McDonough, and J. Makhoul. Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker normalization. In *Proc. ICASSP-97*, pages 1043–1046, Munich, Germany, Apr. 1994.
- [3] T. Anastasakos, F. Kubala, J. Makhoul, and R Schwartz. Adaptation to New Microphones Using Tied-Mixture Normalization. In *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, pages 433–436, 1994.
- [4] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Scale Transform In Speech Analysis. *IEEE Transactions on Speech and Audio Processing*, January 1999.
- [5] T. Kamm, G. Andreou, and J. Cohen. Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. In *Proc. of the 15th Annual Speech Research Symposium*, pages 175–178, Johns Hopkins University, Baltimore, June 1995.
- [6] H. Wakita. Normalization of Vowels by Vocal-Tract Length and its Application to vowel identification. *IEEE Trans. Acoustic, Speech, Signal Processing*, ASSP-25(2):183–192, April 1977.
- [7] G. Fant. *Speech Sounds and Features*. M.I.T. Press, Cambridge, MA, 1973.
- [8] P. E. Nordstrom and B. Lindblom. A Normalization Procedure for Vowel Formant Data. In *Int. Cong. Phonetic Sc.*, Leeds, Aug. 1975.

- [9] James L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer Verlag, New York, 1972.
- [10] John R. Deller, John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, New York, 1993.
- [11] Li Lee and Richard C. Rose. Speaker Normalization Using Efficient Frequency Warping Procedures. In *Proc. IEEE ICASSP'96*, pages 353–356, Atlanta, USA, May 1996.
- [12] G. Fant. A Non-Uniform Vowel Normalization. Technical report, Speech Transmiss. Lab. Rep., Royal Inst. tech., Stockholm, Sweden, 1975.
- [13] James D. Miller. Auditory-Perceptual Interpretation of the Vowel. *Journal of Acoust. Soc. Am.*, 85(5):2114–2134, May 1989.
- [14] A. K. Syrdal and H. S. Gopal. A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels. *Journal of Acoust., Soc. Am.*, 79(4):1086–1100, Apr. 1986.
- [15] S. Umesh, L. Cohen, and D. Nelson. Frequency-Warping And Speaker-Normalization. In *Proc. IEEE International Conference in Acoustics , Speech, and Signal Proc.*, pages 983–986, Munich, Germany, April 1997.
- [16] G. E. Peterson and H. L. Barney. Control Methods Used in a Study of the Vowels. *J. Acoust. Soc. America*, 24(2):175–194, March 1952.
- [17] Vinay M. K., S. Umesh, and Rohit Sinha. A Simple Procedure For Non Uniform Vowel Normalization. Submitted at TENCON-2001.
- [18] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 1990.
- [19] Vinay M. K., S. Umesh, and Rohit Sinha. A Warping Function For Non Uniform Vowel Normalization. Submitted at SPCOM-2001.
- [20] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Frequency-Warping in Speech. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

- [21] Ellen Eide and Herbert Gish. A Parametric Approach to Vocal Tract Length Normalization. In *Proc. IEEE ICASSP'96*, pages 346–349, Atlanta, USA, May 1996.
- [22] Y. Ono, H. Wakita, and Y. Zhao. Speaker Normalization Using Constrained Spectra Shifts In Auditory Filter Domain. In *Proc. Eurospeech-93*, pages 117–124, Berlin, Germany, Sept. 1993.
- [23] S. B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monsyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoustic, Speech, Signal Processing*, ASSP-28:357–366, Aug. 1980.
- [24] A. H. Nuttall and G. C. Carter. Spectral Estimation using Combined Time and Lag Weighting. *Proceedings of the IEEE*, 70:1115–1125, Sept. 1982.
- [25] N. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice Hall, 1984.
- [26] A. Acero and Richard Stern. Environmental Robustness in Automatic Speech Recognition. In *Proc. ICASSP-90*, pages 849–952, Albuquerque, Apr 1990.
- [27] S. Furui. Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. Acoust., Speech, Signal Processing*, 34(1):52–59, Feb. 1986.
- [28] L. R. Rabiner, J. G. Wilpon, and F. K. Soong. High performance connected digit recognition using hidden markov models. In *Proc. ICASSP 88*, April 1988.
- [29] L. Rabiner and B. H. Juang. *Fundamentals Of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.
- [30] L. E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes. *Inequalities*, pages 1–8, 1972.
- [31] C. Becchetti and L. P. Ricotti. *Speech Recognition, Theory and C++ Implementation*. Jhon Wiley & Sons, England, 1999.
- [32] S. J. Young and P C Woodland. *HTK version 1.5 User Reference & Programmer Manual*. Cambridge University Engg. Dept. Speech Group, 1993.