A SIMPLE APPROACH TO NON-UNIFORM VOWEL NORMALIZATION

S. Umesh, S. V. Bharath Kumar, M. K. Vinay, Rajesh Sharma, and Rohit Sinha*

Department of Electrical Engineering
Indian Institute of Technology
Kanpur - 208016, INDIA
{sumesh, bharatsv, rsharma, srohit}@iitk.ac.in

ABSTRACT

In this paper, we present results of non-uniform vowel normalization and show that the frequency-warping necessary to do non-uniform vowel normalization is similar to the mel-scale. We compare our methods to Fant's non-uniform vowel normalization method and show that with proposed frequency warping approach we can achieve similar performance without any knowledge of the spoken vowel and the formant number. The proposed approach is motivated by a desire to perform non-uniform speaker normalization in automatic speech recognition systems. We also present results of a more comprehensive study of our earlier work on non-uniform scaling which again shows that mel-scale is the appropriate warping function. All the results in this paper are based on data from Peterson & Barney and Hillenbrand *et al.* vowel databases.

1. INTRODUCTION

There are wide differences in the formants (F_1, F_2, F_3) of vowels spoken by children, male and female speakers. These are usually attributed to the physiological differences in the vocal-tracts of the speakers. Nordström and Lindblom [1] proposed a normalization procedure in which the formants are scaled by a constant scale factor based on the estimate of the speaker's average vocal-tract length in open vowels as determined from measurement of F_3. This is usually referred to as uniform scaling. Fant [2] then proposed that the scale factor be made a function of both formant number and vowel category. With this approach, Fant claims to reduce the female-male variance to one-half of that remaining after simple uniform scaling of the type suggested by Nordström and Lindblom.

In this paper, we propose two methods of non-uniform vowel normalization, which unlike Fant's method assume no prior knowledge of the vowel category or formant number and yet achieve normalization performance comparable to that of Fant's. In the first method, the basic idea is to scale each formant by its frequency dependent scale factor (and not based on the formant number or vowel category as Fant's method does). In the second approach we propose a model for the relationship between formant frequencies of two speakers and use it for normalization. Corresponding to this model, we obtain a mapping from physical frequency to an alternate domain where the warped spectra are shifted versions of one another for similar enunciations. Interestingly, this mapping which is estimated from vowel data is similar to mel-curve. We also present results of a more comprehensive study of our earlier work, where we have computed a frequency-warping function, such that in the warped domain, spectral envelopes from different speakers are similar except for a possible translation factor. This numerically computed frequency-warping function is also similar to the mel-curve. The methods proposed in this paper are motivated by a desire to do non-uniform speaker normalization in Automatic Speech Recognition (ASR) systems [3].

The paper is organized as follows. In Section 2, we review the method of uniform scaling and Fant's method. In Section 3, we present our proposed methods of non-uniform scaling that do not assume any prior information. In Section 4, we compare the performance of our proposed methods with Fant's method in terms of residual variance after normalization and also using F-ratio to determine the separability of vowels after normalization. All the results in this paper are based on data from Peterson & Barney [4] and Hillenbrand *et al.* [5] databases. In Section 5, we present results of a comprehensive study using these vowel databases to derive a frequency warping function, such that in the warped domain spectral envelopes from different speakers for similar enunciations are similar except for a possible translation factor. We conclude by pointing out to the interesting nature of the frequency warping functions that are obtained by these two different methods.

2. REVIEW OF VOWEL NORMALIZATION METHODS

Nordström and Lindblom (N–L) have suggested a method for uniform scaling, where the formant frequencies of the subject to be normalized are simply to be divided by the factor

$$\alpha = \left(1 + \frac{k}{100}\right) = \frac{F_{3_{sub}}}{F_{3_{ref}}} \qquad (1)$$

where k is the scale factor expressed in percentage. $F_{3_{sub}}$ and $F_{3_{ref}}$ are the *average* F_3 of open vowels of the subject and the male reference speaker respectively.

In Fant's approach to non-uniform scaling, the correction factor k, is made a function of both formant number and vowel category. Briefly, Fant calculates the reference scale factor, $k^j_{n\mathcal{F}}$ between the average female and the average male (i.e. the reference speaker) for the n^{th} formant of j^{th} vowel class. He calculates the factor k for the average female to be 17 using a method slightly different from that in Eq. (1), and by using 6–8 vowel databases of different languages. Fant's non-uniform normalization for any particular adult subject speaker is given by the weighting of $k^j_{n\mathcal{F}}$ with the ratio of subject's particular k to the $k = 17$ of the average

[*] The author is presently with eMuzed India Pvt Ltd, 839, 2nd cross, 7th main, HAL II Stage, Bangalore-560008, India. Email: vinaymk@emuzed.com

female speaker, i.e.

$$k_n^j = k_{n\mathcal{F}}^j \left(\frac{k}{17}\right) \tag{2}$$

For the child speaker, Fant proposes the following non-uniform normalization scheme.

$$k_n^j = k_{n\mathcal{F}}^j \left(\frac{24}{17}\right) + (k-24) \quad \text{for} \quad k > 24 \tag{3}$$

The above formula represents the best prediction of the subject's scale factor for a particular formant of a particular vowel.

In our experiment with Fant's approach, we have observed that using average female as the reference speaker provides better normalization and also a common formula for both adult and child speakers, which is given by

$$k_n^j = k_{n\mathcal{M}}^j \left(\frac{k}{\varphi}\right) \tag{4}$$

where φ is the scale factor between average male and the reference speaker (i.e. average female). Based on the Fant's approach, we calculated φ to be -14.65 for Peterson & Barney (PnB) and -12.18 for Hillenbrand (HiL) databases respectively. We have used the subscript $\mathcal{M}$ in the notation $k_{n\mathcal{M}}^j$ to emphasize that the scaling is for the average male with respect to reference (average female) speaker.

The above non-uniform normalization scheme cannot be directly applied for speaker-normalization since it requires knowledge of the vowel category and formant number. We now propose different two methods for non-uniform normalization and show that their performance approaches to that of Fant's and yet do not make any assumptions and hence may be applied in ASR systems.

3. PROPOSED METHODS OF NON-UNIFORM NORMALIZATION

3.1. Frequency Dependent Scale Factor, $\gamma(f)$

In frequency dependent scale factor (FDSF) approach, we model the weighting factor $k_{n\mathcal{M}}$ as a function of frequency alone. This is essentially done by plotting $k_{n\mathcal{M}}$ for each formant number and vowel as a function of the subject's formant frequency. This is done for all speakers in the database and the averaging is done along the frequency axis over small bands of 100 Hz width. We denote this frequency dependent scale factor as $\gamma(f)$. A plot of $\gamma(f)$ is shown in Fig. 1 for both the databases, where each stem corresponds to the value of $\gamma(f)$ over a 100 Hz band. The non-uniform normalization scheme is given by

$$k_f = \gamma(f) \left(\frac{k}{\varphi}\right) \tag{5}$$

Note that in this approach we do not need to know the formant number, n and the vowel category, j, to do normalization.

3.2. Model Based Non-Uniform Normalization

In model based non-uniform normalization (MBN) approach, the relationship between formant frequencies of a subject speaker and the reference speaker is assumed to have the following form

$$F_R = \alpha_{RS} \left(1 + \frac{F_S}{b}\right)^c \tag{6}$$

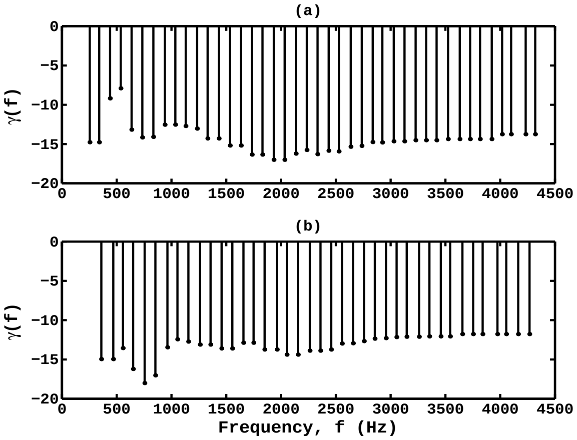

Fig. 1. Frequency dependent scale factor, $\gamma(f)$ for (a) Peterson & Barney database (b) Hillenbrand database.

Database	b	σ_b	c	σ_c
PnB	0.7710	0.3362	0.9756	0.0575
HiL	0.7369	0.2700	0.9761	0.0448

Table 1. Estimates of parameters b and c for model based non-uniform normalization for Peterson & Barney (PnB) and Hillenbrand (HiL) databases. σ_b and σ_c are the standard deviations of b and c respectively.

where F_R and F_S are the formant frequencies of the reference speaker and the subject speaker. α_{RS}, b and c are the model parameters to be estimated. In our analysis the reference speaker is taken to be the average female speaker of the database. The validity of the model was tested for all speakers and the average estimation error energy in fitting the data is less than 1.5% of the energy of the corresponding data. Note that while α_{RS} changes from speaker to speaker, b and c are assumed constant. Table 1 shows the estimate of b and c values of the model for the two databases. Hence the normalization scheme involves using Eq. (6) where b and c are chosen from the table for the appropriate databases. The speaker dependent parameter α_{RS} can be computed from the least-squares fit between the formant frequencies of the reference and subject speaker. This step can be considered to be equivalent to the estimation of k in other approaches.

4. COMPARISON OF NON-UNIFORM NORMALIZATION SCHEMES

In this section, we will compare the performance of the different normalization methods. One of the measures used by Fant to find the efficacy of the non-uniform normalization scheme is the percentage of variance remaining after non-uniform normalization when compared to the uniform normalization scheme of Nordström and Lindblom. The variance in each of the three formants F_1, F_2, F_3 after normalization is given by

$$V_n = \sum_{subject} \sum_{vowel} \left| k_{n,observed} - k_{n,predicted} \right|^2, \quad n = 1, 2, 3 \tag{7}$$

Residual Variance(%)		Ad. & Ch.		Adults		Children	
		PnB	HiL	PnB	HiL	PnB	HiL
Fant	R_1	90	103	80	75	108	151
	R_2	80	89	78	78	84	97
	R_3	93	78	92	83	96	74
FDSF	R_1	88	101	86	84	91	130
	R_2	78	81	82	81	72	81
	R_3	100	82	97	86	106	79
MBN	R_1	93	80	96	77	85	84
	R_2	72	79	79	74	62	83
	R_3	84	73	84	78	84	69

Table 2. Percentage variance remaining after different non-uniform normalization methods when compared to uniform normalization (N–L), for the three formants. Ad. stands for adult speakers and Ch. stands for child speakers. FDSF refers to frequency dependent scaling factor method and MBN refers to model based non-uniform normalization method. PnB and HiL refer to Peterson & Barney and Hillenbrand databases.

where $k_{n,observed}$ is calculated using the actual value of the n^{th} formant of each vowel of the subject and the reference speaker. $k_{n,predicted}$ is the predicted value of scale factor for the n^{th} formant of each vowel of the subject using the different normalization schemes. We compute the percentage residual variance after non-uniform normalization compared to uniform normalization (N–L) for the n^{th} formant as

$$R_n = \frac{V_{n,non-uniform}}{V_{n,uniform}} \times 100 \tag{8}$$

Table 2 shows that the performance of the proposed non-uniform normalization schemes are comparable to Fant's method even though they assume no a priori information about the vowel category and formant number unlike Fant's method. Further, for Hillenbrand data, it can be seen that the proposed model-based non-uniform normalization outperforms Fant's method especially for children.

Since discriminability between vowel clusters is as important as reduction of variance within any given vowel cluster, a good measure for the usefulness of the normalization schemes would be the F-ratio. In deriving F-ratio separability measure, 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 $M_o = \frac{1}{I}\sum_{i=1}^{I} M_i$, where I denotes the number of vowel classes being compared. Then the within-class S_w and between-class S_b scatter matrices, are computed by,

$$S_w = \frac{1}{I}\sum_{i=1}^{I} R_i \quad \text{and} \quad S_b = \frac{1}{I}\sum_{i=1}^{I}(M_i - M_o)(M_i - M_o)^T$$

The separability criterion is then given by,

$$J = trace\{(S_b + S_w)^{-1} S_b\} \tag{9}$$

The vowel cluster discriminability in terms of F-ratio, J, for unnormalized (unwarped), uniform normalization and non-uniform normalization methods are shown in Table 3. From the table it can be seen that even in terms of vowel discriminability the proposed methods perform similar to Fant's method. In Eq. (9) as the separability improves, J should approach the ideal value of 3.

F-Ratio (J)	Ad. & Ch.		Adults		Children	
	PnB	HiL	PnB	HiL	PnB	HiL
Unwarped	2.01	2.13	2.21	2.28	2.31	2.31
N–L	2.42	2.47	2.45	2.56	2.43	2.37
Fant	2.49	2.52	2.52	2.63	2.41	2.40
FDSF	2.47	2.53	2.50	2.61	2.47	2.44
MBN	2.49	2.56	2.51	2.62	2.50	2.46

Table 3. Vowel cluster discriminability in terms of F-Ratio, J. Ad. stands for adult speakers and Ch. stands for child speakers.

5. NUMERICAL COMPUTATION OF WARPING FUNCTION FOR NON-UNIFORM SCALING

In uniform scaling, the formant frequencies are assumed to be scaled versions of one another, or more commonly we assume spectral envelopes between two speakers are scaled versions of each other, i.e. $S_1(f) = S_2(\alpha_{12} f)$. It can be easily seen that in the log-warped domain, i.e. $\lambda = \log(f)$ the spectral envelopes are shifted versions of each other i.e.,

$$\mathbf{s}_1(\lambda) = S_1(f = e^{\lambda}) = S_2(\alpha_{12} e^{\lambda}) = \mathbf{s}_2(\lambda + \ln \alpha_{12}) \tag{10}$$

In our previous work [6, 7], we have shown that the ratio of the formants between any two speakers is not a constant (i.e. uniform scaling is not true), and the trend is to have larger values (compression/dilation) at lower frequencies. We then proceeded to numerically compute the frequency warping function for the non-uniform scaling such that in the warped domain the spectral envelopes for similar enunciation are translated versions of one another. Interestingly this warping function is similar to mel-scale. Our results in the previous work were based on a small sample of vowels extracted from TIMIT database. In this paper, we have re-estimated the frequency-warping function using larger vowel databases from PnB and HiL. We show that the warping function is indeed very similar to mel-scale and we also provide standard deviations to indicate the reliability of the estimates.

In brief, we describe the numerical computation of the warping function as follows. We divide the frequency axis into logarithmically equal regions. In each region, we assume that the spectral envelopes of any two speakers are scaled versions of each other. So for the i^{th} frequency region, $f \in (L_i,\ U_i)$, we have

$$S_1(f) = S_2(\alpha_{12}^{\beta_i} f), \quad L_i \leq f \leq U_i \tag{11}$$

where α_{12} is a constant independent of i (the frequency region) and is dependent on the pair of speakers while β_i depends only on i and is independent of the speakers. We estimate α_{12} from ratio of formants in the last frequency region (i.e. formants above 2400 Hz) assuming $\beta_i = 1$ for the last band. This is reasonable since the higher formants are mostly affected by the length of pharyngeal cavity. Using this estimated value of α_{12} we compute the values of β_i in other frequency regions. Table 4 shows the estimate of β_i for PnB and HiL data obtained by averaging over all pairs of speakers. We have also shown the standard deviations of the estimates in Table 4. If we use the discrete implementation of the warping function $\lambda = \log f$ (exponential sampling of the frequency axis), then in each band the spectral envelopes are shifted versions of one another as shown below.

$$\mathcal{S}_1\left(m_{i_k}\Delta\lambda_i + \ln L_i\right) = \mathcal{S}_2\left(\left(m_{i_k} + \frac{\ln \alpha_{12}^{\beta_i}}{\Delta\lambda_i}\right)\Delta\lambda_i + \ln L_i\right) \tag{12}$$

Peterson & Barney			Hillenbrand		
Band (Hz)	β_i	σ_{β_i}	Band (Hz)	β_i	σ_{β_i}
[190,356)	2.13	0.13	[310,524)	1.50	0.03
[356,667)	1.22	0.05	[524,893)	1.55	0.03
[667,1249)	1.51	0.05	[893,1523)	1.46	0.03
[1249,2339)	1.27	0.04	[1523,2598)	1.40	0.02
[2339,4381)	1.00	0.00	[2598,4431)	1.00	0.00

Table 4. Average estimates of β_i obtained in each frequency region. σ_{β_i} denotes the standard deviation of β_i for i^{th} frequency band. In this case, $1 \leq i \leq 5$

where $k = 0, 1, 2, ..., M_{i-1}$, M_i is the number of spectral samples in i^{th} frequency band and $\Delta\lambda_i = \frac{\ln U_i - \ln L_i}{M_i}$. It can be seen that the two functions differ by a translation factor $\frac{\beta_i \ln \alpha_{12}}{\Delta\lambda_i}$ in the i^{th} frequency band. Since we desire the warped envelopes to be translated versions of one another over the entire range of interest, we require the following condition to be satisfied between any two frequency regions i and j: $\frac{\beta_i \ln \alpha_{12}}{\Delta\lambda_i} = \frac{\beta_j \ln \alpha_{12}}{\Delta\lambda_j}$. Equivalently we require $\beta_i M_i = \beta_j M_j$ by making use of the fact that we have chosen logarithmically equal regions which result in $\ln \frac{U_i}{L_i} = \ln \frac{U_j}{L_j}$. We can therefore choose M_i for different frequency regions (i.e. the spacing of the samples in λ domain) such that $\beta_i M_i = \beta_j M_j$ using estimates of β_i given in Table 4. With this choice, the non-uniformly spaced samples in λ domain are represented as uniformly spaced samples in another domain (say ν domain). Since the scale is arbitrary in ν domain, we can choose the spacing of samples and origin to some convenient values. Fig. 2 shows the relationship between physical frequency, f, and the warped domain, ν, at a discrete set of points determined from above constraint. The curve has been obtained by connecting the discrete points. Note the remarkable similarity between the warping function $\nu = g\left(f\right)$ and the mel-scale obtained using Shaughnessy formula, given by $\nu = 2595 \log\left(1 + \frac{f}{700}\right)$. We have also plotted the actual mel frequency data points obtained by Stevens & Volkman [8].

Finally, we would like to point out that for our proposed non-uniform normalization scheme based on model described in Eq. (6), the following warping function

$$\nu = c \log\left(1 + \frac{f}{b}\right) \qquad (13)$$

results in the spectral envelopes being translated versions of each other in the ν domain. It is interesting to note that this is also similar to mel-scale.

6. DISCUSSION & CONCLUSION

We have proposed two different ways of implementing nonuniform vowel normalization that require no prior knowledge about the vowel category and formant number unlike Fant's method. Using residual variance and F-ratio test for vowel discriminability as performance measures, we show that our proposed methods are comparable (or even better as in HiL data) to Fant's method of non-uniform normalization without requiring the knowledge of vowel category and formant number. The motivation for the proposed methods is to apply them for non-uniform speaker normalization in ASR systems [3]. We also present results using Peterson & Barney and Hillenbrand data verifying our previous work [6, 7], where we have shown that a mel-like warping function is necessary to separate speaker dependent terms as a translation factor.

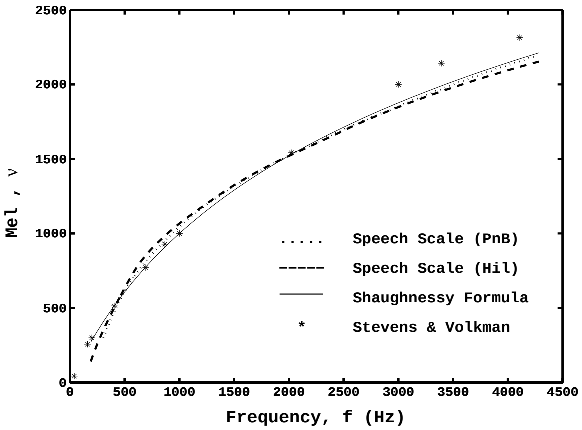

Fig. 2. $\nu = g(f)$ numerically computed from β_i estimates. We have also shown the Shaughnessy formula for mel-scale and the actual Stevens & Volkman data points from their mel-scale experiments.

7. REFERENCES

[1] P. E. Nordström and B. Lindblom, "A Normalization Procedure for Vowel Formant Data," in *Int. Cong. Phonetic Sci.*, Leeds, England, August, 1975.

[2] G. Fant, "A Non-Uniform Vowel Normalization," Technical Report, Speech Transmiss. Lab. Rep., Royal Inst. Tech., Stockholm, Sweden, 1975.

[3] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," Submitted to ICASSP'2002.

[4] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. America*, vol. 24, pp. 175–184, March 1952.

[5] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels," *J. Acoust. Soc. Am.*, vol. 97, pp. 3099–3111, May 1995.

[6] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-Warping in Speech," in *Proc. International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

[7] S. Umesh, L. Cohen, and D. Nelson, "Frequency-Warping and Speaker-Normalization," in *Proc. IEEE ICASSP'97*, Munich, Germany, April 1997, pp. 983–986.

[8] S. S. Stevens and J. Volkman, "The Relation of Pitch to Frequency," *American Journal of Psychology*, vol. 53, pp. 329, 1940.