NON-UNIFORM SPEAKER NORMALIZATION USING AFFINE-TRANSFORMATION

*S. V. Bharath Kumar**

Imaging Technologies Lab
General Electric - Global Research
JFWTC, Bangalore - 560086, INDIA
bharath.sv@geind.ge.com

S. Umesh, Rohit Sinha

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

ABSTRACT

In this paper, we propose a mathematical model to describe the relation between the formant frequencies of speakers and show that with the proposed affine model, speaker differences separate out as translation factors when a "mel-like" warping is performed. Using speech data we estimate the parameters of this warping function and show that it is close to the usual mel-formula. This model is motivated by Rohit *et al.*'s [1] shift-based non-uniform speaker-normalization method, which provides improvement over the conventional maximum-likelihood based speaker normalization methods. We therefore provide a unified framework that relates the relationship between formants of speakers and method of removing speakers difference (which involves mel-warping) in a neat mathematical framework which is substantiated by our recognition experiments.

1. INTRODUCTION

A major source of variability in similar enunciations by different speakers is attributed to the physiological differences in the vocal-tract of the speakers. As an approximation, the vocal tract is assumed to be a tube of uniform cross-section, in which case the speaker variability is directly related to the vocal-tract length (VTL). It has been found that VTL variation causes scaling in the spectral domain [2] since the formant frequencies are inversely proportional to length of the tube. Fant [3] and Umesh *et al.* [4, 5] have shown that uniform/linear scaling of formant frequencies is a very crude approximation and that the formant scaling is non-linear and is phoneme dependent. The non-linearity of the scaling factor has been modelled in various parametric forms [6, 7, 8] but there is no specific motivation to choose a specific parametric form.

In this paper, we propose an affine model to relate the formant frequencies between speakers enunciating the same sound and derive the corresponding frequency-warping function to perform speaker-normalization. The proposed affine model is motivated by the desire to determine whether a "mel-like" frequency-warping function can be derived from speech data alone, which then shows an interesting relation between hearing and speech. Then, corresponding to the proposed affine model relation between speakers, we obtain a mapping from physical frequency to an alternate domain such that in the alternate domain the warped spectra are shifted versions of one and another for similar enunciations. We show that the frequency-warping function for affine model is similar to the mel-warp function. The shift-based non-uniform speaker normalization method proposed by Rohit *et al.* [1] can be used to perform speaker normalization thus removing the speaker differences, which appear as shift factors in warped domain. The basic idea in [1] is to reformulate the linear frequency-warping operation as a shift operation in an alternate domain to perform speaker normalization.

The paper is organized as follows. In Section 2, we present our study on determining the relationship between the formant frequencies of speakers and derive the frequency-warping function for the proposed affine-transformation. We numerically compute the parameters of this frequency-warping function in Section 3. In Section 4, we compare the warping function obtained from the proposed method with log-warp and mel-warp functions. The digit recognition accuracy before and after normalization is used as the measure in evaluating the efficacy of normalization. The frequency-warping functions computed in this paper are based on vowel formant data from Peterson & Barney [9] and Hillenbrand *et al.* [10] databases. We conclude by pointing out to the interesting nature of the frequency-warping function associated with the proposed model, which behaves like mel-warp function.

2. AFFINE MODEL TO DESCRIBE THE RELATION BETWEEN FORMANT FREQUENCIES

We propose the following affine-transformation model relating formant frequencies of the subject speaker and the reference speaker as

$$(F_{\mathcal{R}} + \mathtt{A}) = \alpha_{\mathcal{RS}} (F_{\mathcal{S}} + \mathtt{A}) \quad (1)$$

where $F_{\mathcal{R}}, F_{\mathcal{S}}$ are formant frequencies of the reference speaker, $\mathcal{R}$ and the subject speaker, $\mathcal{S}$ respectively. $\alpha_{\mathcal{RS}}$ and $\mathtt{A}$ are the parameters of the model defined in Eq. (1), which are to be estimated from the speech data. Eq. (1) is similar to linear scaling model, $F_{\mathcal{R}} = \alpha_{\mathcal{RS}} F_{\mathcal{S}}$ except for factor of $\mathtt{A}$. The scaling factor according to affine-transformation model is defined as $\alpha_{\mathcal{RS}} = \frac{F_{\mathcal{R}}+\mathtt{A}}{F_{\mathcal{S}}+\mathtt{A}}$. We assume $\mathtt{A}$ to be speaker-independent parameter and $\alpha_{\mathcal{RS}}$ to be speaker-dependent parameter. In our analysis, the reference speaker is taken to be the average female speaker of the database.

The chief motivating factor in choosing the model in Eq. (1) is to study whether a "mel-like" frequency-warping function can be obtained from speech data alone. Suppose if there exists a "mel-like" warping function obtained from speech data alone, then this shows certain connection between the speech production process and the hearing mechanism. This justifies the use of mel-warp function in speech recognition, not only from psychoacoustic point

*The author performed the work while at Department of Electrical Engineering, Indian Institute of Technology, Kanpur-208016, India.

of view but also from the point of view of speaker normalization. It also provides a neat mathematical framework that relates the relationship between formant frequencies of speakers and shift-based non-uniform speaker normalization method [1].

Consider the affine-transformation model in Eq. (1). It can be rewritten as

$$\left(1+\frac{F_{\mathcal{R}}}{\mathtt{A}}\right)=\alpha_{\mathcal{RS}}\left(1+\frac{F_{\mathcal{S}}}{\mathtt{A}}\right),\ \mathtt{A}\neq 0 \tag{2}$$

Taking logarithms on both sides of Eq. (2), we have

$$\log\left(1+\frac{F_{\mathcal{R}}}{\mathtt{A}}\right)=\log\alpha_{\mathcal{RS}}+\log\left(1+\frac{F_{\mathcal{S}}}{\mathtt{A}}\right) \tag{3}$$

Define $\nu=\log\left(1+\frac{f}{\mathtt{A}}\right)$, then $\nu_{\mathcal{R}}=\nu_{\mathcal{S}}+\log\alpha_{\mathcal{RS}}$, where $\nu_{\mathcal{R}}$ and $\nu_{\mathcal{S}}$ are the warped frequencies of $f=F_{\mathcal{R}}$ and $f=F_{\mathcal{S}}$ respectively. Hence, the warped frequencies appear as shifted or translated versions in ν-domain and the translation factor is speaker-dependent. The warping function to do speaker normalization using affine-transformation model is given by,

$$\nu=\log\left(1+\frac{f}{\mathtt{A}}\right) \tag{4}$$

which is interestingly similar to mel-warp function. For example, Fant's technical mel-formula is given by

$$\eta_{fant}=\frac{1000}{\log 2}\log_{10}\left(1+\frac{f}{1000}\right)$$

while another formula that is commonly used for mel-scale is

$$\eta_{mel}=2595\log_{10}\left(1+\frac{f}{700}\right) \tag{5}$$

As an illustration of the above idea, let us consider uniform scaling wherein the formant frequencies are assumed to be scaled versions of one another, or more commonly we assume spectral envelopes of two speakers, say $\mathcal{A}$ and $\mathcal{B}$ are scaled versions of each other, i.e. $S_{\mathcal{A}}(f)=S_{\mathcal{B}}(f')$ and $f'=\alpha_{\mathcal{AB}}f$. 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.,

$$\mathbf{s}_{\mathcal{A}}(\lambda)=S_{\mathcal{A}}(f=e^{\lambda})=S_{\mathcal{B}}(\alpha_{\mathcal{AB}}e^{\lambda})=\mathbf{s}_{\mathcal{B}}(\lambda+\log\alpha_{\mathcal{AB}}) \tag{6}$$

We now consider the proposed affine-transformation in Eq. (1), i.e., $f'=\alpha_{\mathcal{AB}}f+\mathtt{A}\left(\alpha_{\mathcal{AB}}-1\right)$. It is easy to see that, in the warped domain, $\nu=\log\left(1+\frac{f}{\mathtt{A}}\right)$, the spectral envelopes are shifted versions of each other i.e.,

$$\begin{aligned}\mathbf{s}_{\mathcal{A}}(\nu) &= S_{\mathcal{A}}(f=\mathtt{A}\left(e^{\nu}-1\right))\\ &= S_{\mathcal{B}}(f'=\alpha_{\mathcal{AB}}f+\mathtt{A}\left(\alpha_{\mathcal{AB}}-1\right))\\ &= S_{\mathcal{B}}(\mathtt{A}\alpha_{\mathcal{AB}}\left(e^{\nu}-1\right)+\mathtt{A}\left(\alpha_{\mathcal{AB}}-1\right))\\ &= S_{\mathcal{B}}\left(\mathtt{A}\left(e^{\log\alpha_{\mathcal{AB}}+\nu}-1\right)\right)\\ &= \mathbf{s}_{\mathcal{B}}(\nu+\log\alpha_{\mathcal{AB}})\end{aligned} \tag{7}$$

The warped spectra appear as shifted versions of one and another in the warped domain, $\nu=\log\left(1+\frac{f}{\mathtt{A}}\right)$. This idea is exploited to do speaker-normalization in [1].

The cepstral coefficients, which are the defacto features used in state-of-the-art ASRs, are computed on warped spectra. Since, the spectra appear as translated versions in the warped domain, the cepstral coefficients computed on the warped spectra for a given speaker will be modulated version of that of the reference speaker. To make the theory complete, let us define $W:[0,\pi]\rightarrow[0,\pi]$ such that $\omega\rightarrow\hat{\nu}=W(\omega)$ and $\omega,\hat{\nu}\in[0,\pi]$. Let f_s be the sampling frequency in f-domain. So, $\omega=\frac{2\pi f}{f_s}$. ω is the digital frequency in rad/s and f is the analog frequency in Hz. Eq. (1) can be rewritten as

$$\left(\frac{\omega_{\mathcal{R}}f_s}{2\pi}+\mathtt{A}\right)=\alpha_{\mathcal{RS}}\left(\frac{\omega_{\mathcal{S}}f_s}{2\pi}+\mathtt{A}\right)$$

$$\log\left(1+\frac{\omega_{\mathcal{R}}f_s}{2\pi\mathtt{A}}\right)=\log\alpha_{\mathcal{RS}}+\log\left(1+\frac{\omega_{\mathcal{S}}f_s}{2\pi\mathtt{A}}\right)$$

Let $W'(\omega)=\log\left(1+\frac{\omega f_s}{2\pi\mathtt{A}}\right)$. It is clear that $W'(0)=0$ and $W'(\pi)=\log\left(1+\frac{f_s}{2\mathtt{A}}\right)$. Define $\hat{\nu}=W(\omega)=\frac{\pi}{W'(\pi)}W'(\omega)$. Now $W(0)=0$ and $W(\pi)=\pi$. Hence the frequency-warping function is given as

$$\hat{\nu}=W(\omega)=\frac{\pi}{\log\left(1+\frac{f_s}{2\mathtt{A}}\right)}\log\left(1+\frac{\omega f_s}{2\pi\mathtt{A}}\right) \tag{8}$$

Since f_s and $\mathtt{A}$ are constants, define $\mathtt{K}_1=\frac{\pi}{\log\left(1+\frac{f_s}{2\mathtt{A}}\right)}$ and $\mathtt{K}_2=\frac{2\pi\mathtt{A}}{f_s}$. The warping function can be expressed as

$$\hat{\nu}=W(\omega)=\mathtt{K}_1\log\left(1+\frac{\omega}{\mathtt{K}_2}\right) \tag{9}$$

$\mathtt{K}_1$ and $\mathtt{K}_2$ are dependent on $\mathtt{A}$ and the determination of $\mathtt{A}$ defines the warping function $W(\omega)$.

The model in Eq. (1), as mentioned earlier, is motivated by the desire to determine whether any connection exists between the speech production process and the hearing mechanism. So, we made a comprehensive study on the relationship between speakers, by finding out different models, that normalize the speakers. The analysis is carried out on the formant data of vowels collected from 14 "representative" speakers for both Peterson & Barney (PnB) and Hillenbrand (HiL) databases. 5 male, 5 female and 4 child "representative" speakers for both the databases are obtained for experimentation. A representative speaker is computed as the mean of formant frequencies of a set of speakers. TableCurve2D curve-fitting software package is used to fit relationships between different combinations of subject and reference speakers. Table 1 shows the best simple curvefits obtained by TableCurve2D for the vowel data of PnB and HiL databases, which are ranked according to the accuracy of fit.

3. NUMERICAL COMPUTATION OF '$\mathtt{A}$'

The numerical computation of $\mathtt{A}$ involves fitting the affine-transformation model in Eq. (1) for the data points involving the formant frequencies of the reference and subject speaker. The experiment to compute $\mathtt{A}$ is carried out on the formant data from PnB and HiL databases. Each speaker in both of these databases is characterized by formant vector (F_1,F_2,F_3). PnB and HiL databases have 10 and 12 vowels uttered respectively by each subject. Hence the number of data points involving the formant frequencies of a given speaker of PnB and HiL is 30 and 36 respectively. The average female speaker of the respective databases is chosen to be the reference speaker. For each subject speaker of a given database, α and $\mathtt{A}$ are computed with respect to the reference speaker and the mean estimate of $\mathtt{A}$ is computed for each of the databases. It has

Rank	Model Equations	
	Peterson & Barney	Hillenbrand
1	$y = ax + b(a-1)$	$y = ax + b(a-1)$
2	$y = a + bx^c$	$y = a + be^{-x/c}$
3	$y = ax^b$	$y = ax^b$
4	$y = a + be^{-x/c}$	$y^{0.5} = a + bx^{0.5}$
5	$y^{0.5} = a + bx^{0.5}$	$\log(y) = a + b\log(x)$
6	$y = a + bx$	$y = a + bx$
7	$\log(y) = a + b\log(x)$	$y = a + bx^c$
8	$y^{-1} = a + bx^{-1}$	$y^{-1} = a + bx^{-1}$
9	$y^2 = a + bx^2$	$y^2 = a + bx^2$
10	$y = a + \frac{bx}{\log(x)}$	$y = a + \frac{bx}{\log(x)}$
11	$y^{0.5} = a + bx^{0.5}\log(x)$	$y^{0.5} = a + bx^{0.5}\log(x)$
12	$y^2 = a + bx^2\log(x)$	$y = a + bx\log(x)$
13	$y = a + bx\log(x)$	$y^2 = a + bx^2\log(x)$
14	$\log(y) = a + b(\log(x))^2$	$\log(y) = a + b(\log(x))^2$
15	$y^{-1} = a + bx^{-1}\log(x)$	$y^{-1} = a + bx^{-1}\log(x)$

Table 1. Best simple curvefits for vowel formant data of Peterson & Barney and Hillenbrand databases.

to be noted that $\mathtt{A}$ should be positive. But, while performing the global optimization, there are lot of subjects for whom the value of $\mathtt{A}$ saturated to the lower bound of $\mathtt{A}$ (i.e. 0) and actually for such speakers, the global optimum occurs at some negative value of $\mathtt{A}$. If we neglect the speakers who have $\mathtt{A} < 0$, then the mean estimate of $\mathtt{A}$ computed is not a good estimate. Hence, a different method is used to compute the values of $\mathtt{A}$ and α.

Let us consider that a given database has a total of M subjects of which K are female subjects. The estimates of α and $\mathtt{A}$ for a given subject is computed as follows.

$$F_{\mathcal{R}_i} = \alpha_{ij} F_{\mathcal{S}_j} + \mathtt{A}_{ij}\,(\alpha_{ij} - 1) \tag{10}$$

and $i = 1, 2, \cdots, K;\ j = 1, 2, \cdots, M$. $F_{\mathcal{R}_i}$ and $F_{\mathcal{S}_j}$ are the data points of formant frequencies (of size 30 or 36 depending upon the database) of i^{th} reference speaker (average female speaker) and j^{th} subject speaker respectively. The average female speaker i.e. reference speaker is given by

$$\begin{aligned} F_{\mathcal{R}} &= \frac{1}{K}\sum_{i=1}^{K} F_{\mathcal{R}_i} \\ &= \frac{1}{K}\sum_{i=1}^{K}\left(\alpha_{ij} F_{\mathcal{S}_j} + \mathtt{A}_{ij}\,(\alpha_{ij} - 1)\right) \\ &= F_{\mathcal{S}_j}\frac{1}{K}\sum_{i=1}^{K}\alpha_{ij} + \frac{1}{K}\sum_{i=1}^{K}\mathtt{A}_{ij}\,(\alpha_{ij} - 1) \end{aligned} \tag{11}$$

But from our model in Eq. (1), we require

$$F_{\mathcal{R}} = \alpha_j F_{\mathcal{S}_j} + \mathtt{A}_j\,(\alpha_j - 1),\ j = 1, 2, \cdots, M \tag{12}$$

It is clear from Eq. (11) and Eq. (12) that

$$\alpha_j = \frac{1}{K}\sum_{i=1}^{K}\alpha_{ij} \tag{13}$$

$$\mathtt{A}_j = \frac{\sum_{i=1}^{K}\mathtt{A}_{ij}\,(\alpha_{ij} - 1)}{\sum_{i=1}^{K}(\alpha_{ij} - 1)},\ j = 1, 2, \cdots, M \tag{14}$$

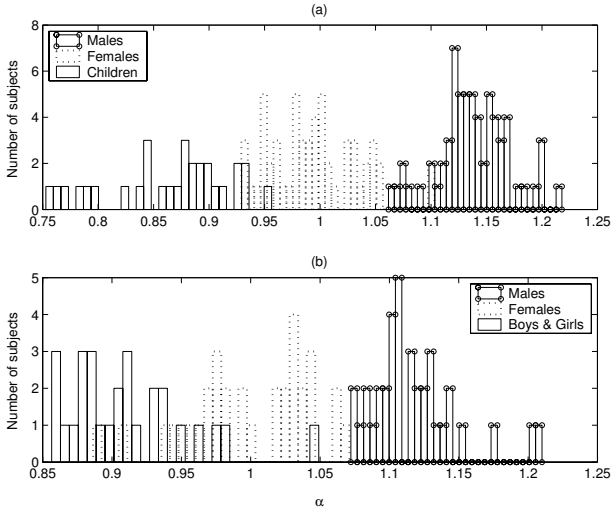

Fig. 1. Histogram of the speaker-dependent parameter, α in speaker normalization using affine-transformation for (a) Peterson & Barney database (b) Hillenbrand database.

The mean estimate of $\mathtt{A}$ for a given database is computed as

$$\mathtt{A} = \frac{1}{M}\sum_{j=1}^{M}\mathtt{A}_j \tag{15}$$

The value of $\mathtt{A}$ has been estimated to be 508.04 for PnB database and 495.67 for HiL database. The warping function for PnB and HiL databases is given by

$$\nu = \begin{cases} \log\left(1 + \frac{f}{508.04}\right) & \text{for PnB database,} \\ \log\left(1 + \frac{f}{495.67}\right) & \text{for HiL database.} \end{cases} \tag{16}$$

Figure 1 shows the histograms of $\alpha_{\mathcal{RS}}$ for male, female and child speakers of PnB and HiL databases. The trend in the estimates of $\alpha_{\mathcal{RS}}$ across the genders shows the existence of gender separability. Also, since average female is considered as the reference subject, the female speakers are centered around $\alpha = 1$ warping factor.

4. COMPARISON OF LOG-WARP, MEL-WARP AND AFFINE-WARP FUNCTIONS

Figure 2 shows the plot of log-warp, mel-warp and affine-warp functions. Since the value of $\mathtt{A}$ is almost same for both PnB and HiL databases, the affine-warp function for both of these databases appear same. We would like to remind the reader that $\mathtt{A}$ has been obtained from the study of vowels only. It is very interesting to note that the affine-warp function is almost same as mel-warp function. The mel-warp function in Eq. (5) is actually obtained by fitting a curve to Stevens & Volkman [11] data points. A model similar to Eq. (4) is fitted in [12], which computes the value of $\mathtt{A}$ to be 657.6. The Stevens & Volkman data is obtained by conducting experiments related to the human auditory response with "human perception". It has been found from this data that human ear behaves on mel-scale. Now, our experiments on speaker normalization, which is conducted on speech data alone shows the required frequency-warping to be close to mel-scale. This is indeed

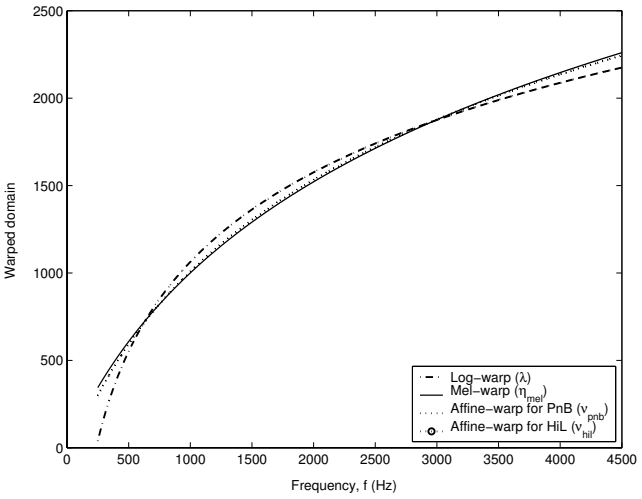

Fig. 2. Comparison of warping function, ν derived using affine-transformation model with log-warp and mel-warp functions. The affine-warp functions for Peterson & Barney (PnB) and Hillenbrand (HiL) databases overlap, which is obvious from Eq. (16).

very interesting and may explain mel-scale not only from the psychoacoustic point of view but also from the view point of speaker normalization.

The normalization performance of different warping functions is evaluated by computing the recognition accuracies on a telephone based connected word recognition task. The data for our digit recognition experiment is collected from *Numbers v1.0cd* corpus of OGI containing 2169 utterances from male and female speakers. A *mismatched* test set derived from other than *Numbers* corpus consists of 2798 utterances from children. Eleven word models are generated for 1 to 9, *zero*, *oh* along with one silence model. The word models and silence model are modelled as 16 and 3 states respectively. Word models have 5 diagonal Gaussian mixtures per state and silence model has 6 Gaussian mixtures per state. Speech signals are sectioned with an overlapping window of 20 ms frame size and with an overlap of 10 ms. A first-order backward difference pre-emphasis with factor 0.97 is computed. The spectral features are computed using Weighted Overlap Segment Averaging (WOSA) technique [1] with each frame being sectioned into hamming windowed sub-frames of 64 samples with an overlap of 45 samples. The cepstral features are then computed for recognition task. Table 2 shows the recognition performance of the digit recognizer, before and after normalization for different warping functions. The proposed affine model-based warping functions perform better than log-warp function and approach the performance of mel-warp function.

5. DISCUSSION & CONCLUSION

We have proposed an affine-model to describe the relationship between formant frequencies of any two speakers enunciating the same sound. The motivation for proposing the above model is based on the fact that the warping function necessary to do normalization is similar to mel-warp function. This study, therefore, provides an interesting model to use the mel-warp function in automatic speech recognition, not only from the psychoacoustic point of view but also from the view point of speaker normalization. We also provide a unified mathematical framework relating the proposed affine-transformation and shift-based non-uniform speaker normalization method. Using digit recognition results as a performance measure, we conclude that our proposed method performs similar to the shift-based speaker-normalization method of [1], which clearly showed improvement over the conventional speaker-normalization method for a digit recognition task.

% Recognition accuracy	Adults		Children	
	R_b	R_n	R_b	R_n
Mel-warp	96.98	97.48	86.27	92.04
Log-warp	96.76	97.15	86.96	90.80
Affine-PnB	97.03	97.44	86.51	92.00
Affine-HiL	96.96	97.43	86.58	91.98

Table 2. Recognition performance of various frequency warping functions on a digit recognizer before and after normalization. R_b and R_n represent the percentage recognition accuracies before (baseline) and after normalization respectively.

6. REFERENCES

[1] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," in *Proc. IEEE ICASSP*, Orlando, USA, May 2002.

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

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

[4] S. Umesh, L. Cohen, and D. Nelson, "Frequency Warping and the Mel Scale," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 104–107, March 2001.

[5] S. Umesh, S. V. Bharath Kumar, M. K. Vinay, Rajesh Sharma, and Rohit Sinha, "A Simple Approach to Non-Uniform Vowel Normalization," in *Proc. IEEE ICASSP*, Orlando, USA, May 2002, pp. 517–520.

[6] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *Proc. IEEE ICASSP*, Atlanta, USA, May 1996, pp. 346–348.

[7] P. Zhan and M. Westphal, "Speaker Normalization Based on Frequency Warping," in *Proc. IEEE ICASSP*, Munich, Germany, April 1997, pp. 1039–1042.

[8] J. McDonough, W. Bryne, and X. Luo, "Speaker Normalization with All-Pass Transforms," in *Proc. ICSLP*, Sydney, Australia, November 1998.

[9] 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.

[10] 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.

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

[12] S. Umesh, L. Cohen, and D. Nelson, "Fitting the Mel Scale," in *Proc. IEEE ICASSP*, 1999, pp. 217–220.