NON-UNIFORM SPEAKER NORMALIZATION USING FREQUENCY-DEPENDENT SCALING FUNCTION

*S. V. Bharath Kumar**

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

S. Umesh

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

ABSTRACT

In this paper, we present frequency-dependent scaling method of non-uniform speaker normalization and show that the frequency-warping necessary to perform non-uniform speaker normalization is similar to mel-scale. The proposed method is motivated by the desire to apply Fant's non-uniform vowel normalization scheme in automatic speech recognition systems. Using word error rate as performance measure, we show that the proposed method performs similar or even better than mel-warp function. Since, the proposed warping function is similar to mel-scale, we argue that our study "justifies" the usage of mel-scale in speech recognition, not only from the point of view of psychoacoustics but also from the view point of speaker normalization.

1. INTRODUCTION

The wide differences in the formants (F_1, F_2, F_3) of vowels spoken by children, male and female speakers are usually attributed to the physiological differences in the vocal-tracts of speakers. The vocal-tract is usually approximated to a tube of uniform cross-section so that the speaker variability is attributed directly to the vocal-tract length. Nordström and Lindblom [1] proposed a *uniform normalization* procedure in which the formants are scaled by a constant scale factor based on the estimate of speaker's average vocal-tract length in open vowels as determined from the measurement of F_3. Fant [2] proposed a *non-uniform normalization* procedure as an improvement over Nordström and Lindblom method, by modelling the scale factor as a function of both formant number and vowel category. Since Fant's method assumes the availability of vowel category and formant number, it cannot be applied for speaker normalization on automatic speech recognition (ASR) systems, as there is nothing to recognize once we know the vowel category. Umesh *et al.* [3] proposed a *non-uniform vowel normalization scheme* based on frequency-dependent scaling factor, which unlike Fant's method assumes no prior knowledge of the vowel category or formant number and yet achieves normalization performance comparable to that of Fant's.

In this paper, we present a more comprehensive study on the non-uniform vowel normalization of [3] using frequency dependent scaling method. The method in [3] is motivated by the desire to apply Fant's non-uniform normalization scheme to speech recognition by removing the dependence of scaling factor on vowel category and formant number, which we achieve by computing the frequency dependent scaling function. Based on the frequency-dependent scaling function, we derive a frequency-warping function, which maps the physical frequency to an alternate domain where the warped spectra are shifted versions of one another for similar enunciations. The frequency-warping function for the proposed method is similar to mel-warp function, which "justifies" the usage of mel-scale in speech recognition. The shift-based non-uniform speaker normalization method proposed by Rohit *et al.* [4] can be used to perform speaker normalization thus removing the speaker differences, which appear as shift factors in the warped domain.

The paper is organized as follows. In Section 2, we briefly present the non-uniform speaker normalization method of Umesh *et al.* [3] based on the frequency-dependent scaling factor. Extending on [3], we also present our comprehensive study to determine frequency-dependent scaling function, $\gamma(f)$ in Section 2. The method of computation of frequency-warping function based on $\gamma(f)$ is presented in Section 3. We also compare the behavior of the derived warping function with log-warp and mel-warp functions. In Section 4, we compare the performance of the proposed method with other normalization schemes by the word error rates before and after normalization, on a telephone based connected digit recognition task. The frequency-warping functions computed in this paper are based on the formant

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

data from Peterson & Barney [5] and Hillenbrand *et al.* [6] vowel databases. We conclude by pointing out to the interesting nature of the proposed frequency-warping function, which is similar to mel-warp function, thus showing a connection between speech production mechanism and human auditory response.

2. FREQUENCY-DEPENDENT SCALING METHOD

Umesh *et al.* [3] reviews the uniform and non-uniform vowel normalization methods of Nordström and Lindblom and Fant respectively. [3] also presents a modified version of Fant's non-uniform normalization scheme for both adult and child speakers, which is given by

$$k_n^j = k_{n\mathcal{M}}^j \left(\frac{k}{\varphi}\right) \tag{1}$$

where $k_{n\mathcal{M}}^j$ is the reference scale factor between the average female and the average male for n^{th} formant of j^{th} vowel class. The subscript $\mathcal{M}$ in the notation $k_{n\mathcal{M}}^j$ is to emphasize that the scaling is for the average male with respect to the reference (average female) speaker. k is the percentage scale factor defined by Nordström and Lindblom as $k = 100(\alpha - 1)$, where $\alpha = \frac{F_{3_{sub}}}{F_{3_{ref}}}$. $F_{3_{sub}}$ and $F_{3_{ref}}$ are the *average* F_3 of open vowels of the subject and the reference speaker respectively. φ is the scale factor between the average male and the reference speaker and is calculated to be -14.65 for Peterson & Barney (PnB) and -12.18 for Hillenbrand (HiL) databases respectively. k_n^j represents the best prediction of the subject's scale factor for n^{th} formant of j^{th} vowel class. The non-uniform normalization scheme in Eq. (1) cannot be applied directly for speaker normalization on ASR systems since it requires the knowledge of vowel category and formant number.

Umesh *et al.* proposed frequency-dependent scale (FDS) factor method as a solution to the above problem, which does not require any information about the vowel category and formant number but performs better than Fant's normalization method. The basic idea behind this method is to model the weighting factor $k_{n\mathcal{M}}$, as a function of frequency alone, thus making it both vowel and formant independent.

2.1. Frequency-Dependent Scaling Function, $\gamma(f)$

In frequency-dependent scale factor approach, we model the weighting factor $k_{n\mathcal{M}}$ as a function of frequency alone. The $k_{n\mathcal{M}}$ values are averaged over vowel category and formant number to obtain a frequency-dependent scaling function, $\gamma(f)$, which is purely a function of frequency.

Let us consider a vowel database of L vowels being uttered by M speakers, where each vowel is characterized by the formant vector (F_1, F_2, F_3), F_i being the i^{th} formant frequency. Hence, any subject, m is characterized by

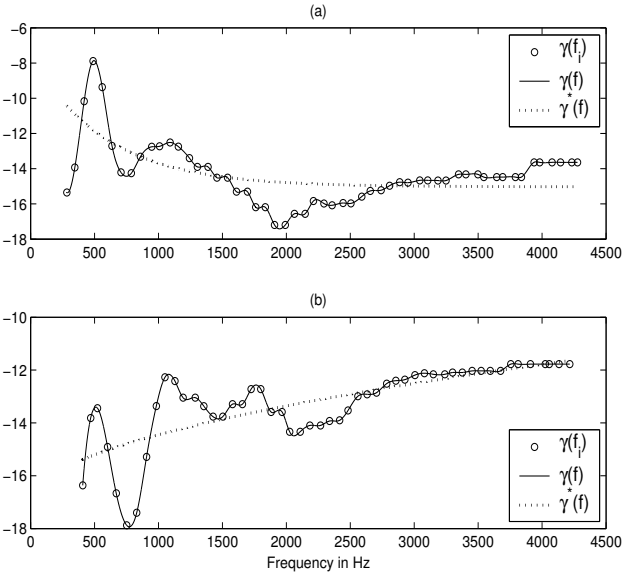

Fig. 1. Frequency-dependent scale factors, $\gamma(f_i)$ and frequency-dependent scaling function, $\gamma(f)$ for (a) Peterson & Barney and (b) Hillenbrand databases. $\gamma^*(f)$ represents the simple curvefit to $\gamma(f_i)$ against f_i. Refer to Eq. (2).

the frequency vector, $\mathsf{f}^m = \{F_{m,n}^j, j = 1, \cdots, L; n = 1, 2, 3\}$ and the frequency-dependent scale vector, $\mathsf{k}^m = \{k_{m,n\mathcal{M}}^j, j = 1, \cdots, L; n = 1, 2, 3\}$. From Eq. (1), it is clear that $k_{n\mathcal{M}}^j$ is speaker independent but is dependent on vowel category and formant number. Let us define $\mathsf{f} = \cup_{m=1}^{M} \mathsf{f}^m$ and $\mathsf{k} = \cup_{m=1}^{M} \mathsf{k}^m$. Since, f and k are discrete, we are interested in finding a continuous function $h(.)$, such that $h(\mathsf{f}) = \mathsf{k}$. To remove the dependence of k on formant number and vowel category, the elements of k are averaged over small frequency bands of 150 Hz width with 75 Hz overlap. The weighted mean of k values is computed for a given frequency band and is assigned to the mean frequency of that band. Let us assume that there are N frequency bands in $[\inf \mathsf{f}, \sup \mathsf{f}]$. The frequency-dependent scale factor for i^{th} frequency band, B_i, is given as

$$\gamma(f_i) = \frac{\sum_{\forall x \in A_i} x h(x)}{\sum_{\forall x \in A_i} x} \tag{2}$$

$$f_i = \frac{1}{\mathcal{C}(A_i)} \sum_{\forall x \in A_i} x$$

$$A_i = \{x | x \in B_i \subset \mathsf{f}\}, \forall\, i = 1, 2, \cdots, N.$$

$\gamma(f)$ is the frequency-dependent scaling function, which is independent of vowel class and formant number. $\mathcal{C}(.)$ is the cardinality operator. The above method is a small variation over the one discussed in [3] and provides better separation in terms of Fischer discriminant between the vowel clusters than the method discussed in [3]. Figure 1 shows the plot of $\gamma(f)$ for PnB and HiL databases, which is a cubic

spline fit to $\gamma(f_i)$, which are depicted as circles. Extending from Eq. (1), the non-uniform normalization scheme using frequency-dependent scaling function is given by

$$\kappa(f) = \gamma(f)\left(\frac{k}{\varphi}\right) \tag{3}$$

where $\kappa(f)$ is the speaker and frequency-dependent scaling function and is independent of vowel category and formant number. Hence, the normalization scheme in Eq. (3) can be used for speaker normalization on a speech recognizer unlike the Fant's method.

Figure 1 shows that $\gamma(f)$ is highly non-linear and so it would be interesting to find a closed form expression to $\gamma(f)$. Hence, a simple curve is fitted using TableCurve2D software package to $\gamma(f_i)$ against f_i, which is also shown in Figure 1 as $\gamma^*(f)$. It is clear that $\gamma^*(f)$ is not a good estimate of $\gamma(f)$, which indicates the difficulty in finding a simple closed form expression to $\gamma(f)$. Since the scale factor defined in Eq. (3) is frequency-dependent, replacing k in $\alpha = 1 + \frac{k}{100}$ by $\kappa(f)$, the frequency-dependent scale factor is obtained as

$$\alpha(f) = 1 + \frac{\kappa(f)}{100} = 1 + \frac{\gamma(f)(\alpha - 1)}{\varphi} \tag{4}$$

Eq. (4) shows that $\alpha(f)$ is not only frequency-dependent but also speaker-dependent because of the presence of factor α. One way to use the knowledge of scaling function, $\gamma(f)$ is to develop a *universal* frequency-warping function, leading to scale invariance.

3. FREQUENCY-WARPING FUNCTION BASED ON FREQUENCY-DEPENDENT SCALING METHOD

Consider two speakers $\mathcal{A}$ and $\mathcal{B}$, whose spectra are related by

$$S_{\mathcal{A}}(f) = S_{\mathcal{B}}(g(\alpha_{\mathcal{AB}}, f)) \tag{5}$$

where $g(\alpha_{\mathcal{AB}}, f)$ is some function that involves speaker dependencies through the first argument. Let $f' = g(\alpha, f) = \alpha(f)f = \alpha^{\beta(f)}f$, where α is the subject's scale factor with respect to the reference speaker and is frequency-independent. $\beta(f)$ is dependent only on frequency and is independent of speaker. $\beta(f)$ captures the non-linearity in scale factor, $\alpha(f)$. The motivation in modelling $\alpha(f) = \alpha^{\beta(f)}$ is to separate the speaker dependency of scale factor from frequency dependence. From Eq. (4), we have

$$\alpha(f) = 1 + \frac{\gamma(f)(\alpha - 1)}{\varphi} = \alpha^{\beta(f)}$$

$$\Rightarrow \beta(f) = \frac{\log(1 + \frac{\gamma(f)(\alpha-1)}{\varphi})}{\log(\alpha)} \tag{6}$$

The closed form expression to $\beta(f)$ is difficult to compute because of its dependence on $\gamma(f)$. Hence, instead of

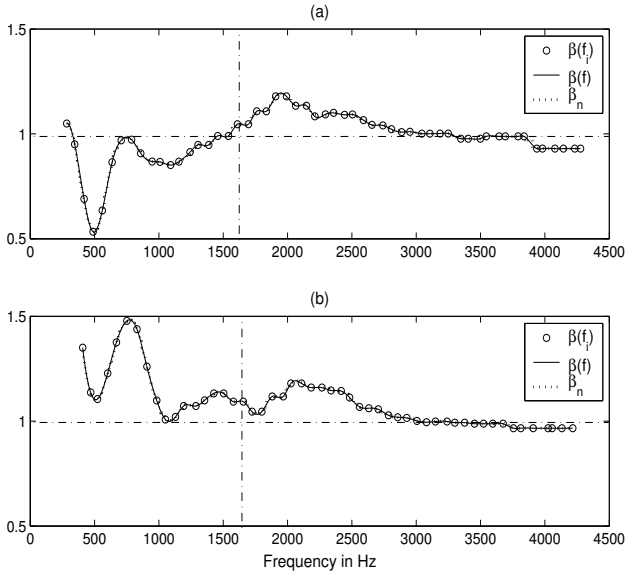

Fig. 2. Average values of $\beta(f_i)$ and $\beta(f)$ for (a) Peterson & Barney and (b) Hillenbrand databases. β_n is the piecewise approximation to $\beta(f)$. Refer to Eq. (7) and Eq. (11).

computing $\beta(f)$, we compute $\beta(f_i)$ from $\gamma(f_i)$, which is given as

$$\beta(f_i) = \frac{\log(1 + \frac{\gamma(f_i)(\alpha-1)}{\varphi})}{\log(\alpha)} \tag{7}$$

$\beta(f_i)$ is assumed to be independent of speaker variation, but practically, there might be a slight dependence because of the existence of α in Eq. (7). In a database of M speakers, different $\beta^m(f_i), m = 1, 2, \cdots M$, can be computed for speakers with corresponding α_m, by replacing α with α_m in Eq. (7). The assumption on $\beta(f_i)$ would be practically valid if the variance of $\beta^m(f_i)$ is small, which is the case in our experiments. Then, $\beta(f_i)$ is computed as $\beta(f_i) = \frac{1}{M}\sum_m \beta^m(f_i)$. Similar to $\gamma(f)$, $\beta(f)$ can be obtained as a cubic spline fit to $\beta(f_i)$. Figure 2 shows the plot of $\beta(f)$ and $\beta(f_i)$ for PnB and HiL databases. Having computed $\beta(f)$, the frequency warping function can be derived as follows.

Consider $g(\alpha, f)$, which can be modified as

$$\frac{\log(f')}{\beta(f)} = \log(\alpha) + \frac{\log(f)}{\beta(f)} \tag{8}$$

Assuming $\beta(f') \simeq \beta(f)$ (usually the case at higher frequencies), we get

$$\nu' = \nu + \log(\alpha) = \nu + \text{constant shift} \tag{9}$$

where

$$\nu = \psi(f) = \frac{\log(f)}{\beta(f)} \tag{10}$$

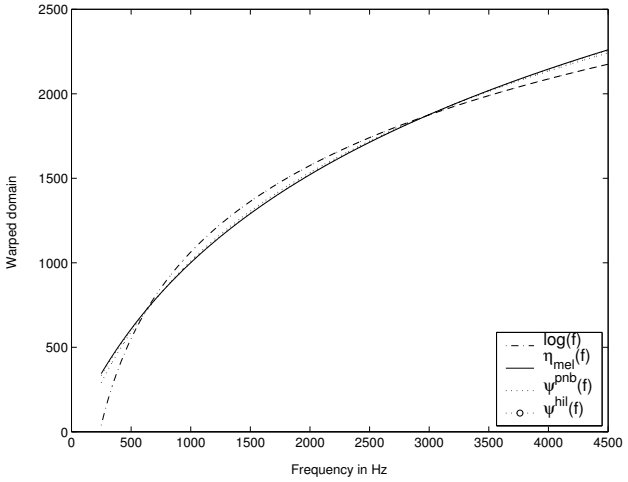

Fig. 3. Comparison of warping function, $\psi(f)$ derived using frequency-dependent scaling method with log-warp and mel-warp functions. The proposed-warp functions for Peterson & Barney (PnB) and Hillenbrand (HiL) databases overlap, which is obvious from Eq. (13).

is the frequency warping function. Eq. (9) shows that the spectra in the warped domain are translated versions of one another, so that the magnitude of Fourier transform of these warped spectral patterns are invariant to translations, leading to scale invariant features, of real speech signals. Since the assumption on $\beta(f)$ is not valid everywhere and the computation of closed form expression to $\beta(f)$ is difficult, we borrow a technique from [7] where a piecewise warping function is computed by modelling the frequency dependent scale factor as piecewise exponentials. This method provides a warping function, which satisfies the shift property of Eq. (9) in piecewise manner, thus leading to scale invariance.

The method in [7] involves the computation of piecewise exponential factors over logarithmically equi-spaced frequency bands, which provides piecewise warping function such that, in the warped domain, the warped spectra appear as shifted versions of one and another. A similar procedure can be used to compute piecewise warping function by discretizing $\beta(f)$ over logarithmically equi-spaced frequency bands. To get a better piecewise approximation to $\beta(f)$, the discretization is done over 150 logarithmically equi-spaced bands and β_n is computed as the mean $\beta(f)$ for n^{th} frequency band, i.e.,

$$\beta_n = \frac{1}{\mathcal{C}(Z_n)} \sum_{\forall x \in Z_n} x \qquad (11)$$
$$Z_n = \{\beta(f) | f \in B'_n\}$$

where B'_n is the n^{th} frequency band. It can be seen from Figure 2 that β_n is a good piecewise approximation to $\beta(f)$.

It is of interest to note the similar behavior of $\beta(f)$ over $f > 1600$ Hz for PnB and HiL databases. Based on [7], the piecewise frequency warping function to perform non-uniform speaker normalization using frequency-dependent scaling method is given as

$$\psi_n(f) = \frac{\log(f)}{\beta_n} \qquad (12)$$

Since the warping function obtained is discrete, we fit a parametric model $\psi(f) = a\log(1 + \frac{f}{b})$, similar to mel-warp, $\eta_{mel}(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right)$, using TableCurve-2D to obtain continuous warping function $\psi(f)$. a and b are the parameters of the model to be computed. The motivation to fit a mel-like model is to study the functional similarity of the warping function derived from speech data alone, with mel-warp function, which is obtained from psychoacoustic studies. If there exists a similarity between these warping functions, it would show a connection between speaker normalization and psychoacoustics, thus justifying the usage of mel-warp function in speech recognition from the point of view of speaker normalization. Based on our study, the warping function, $\psi(f)$ for PnB (97% fitting accuracy) and HiL (99.6% fitting accuracy) databases are computed as

$$\psi^{pnb}(f) = 2096.40\log\left(1 + \frac{f}{475.34}\right) \qquad (13a)$$

$$\psi^{hil}(f) = 2484.95\log\left(1 + \frac{f}{646.00}\right) \qquad (13b)$$

It is interesting to note from Eq. (13) that the warping function, $\psi(f)$ is functionally closer to $\eta_{mel}(f)$ than to $\log(f)$ function. Figure 3 shows the plot of log-warp, mel-warp and the proposed warping functions.

Mel-warp function is actually derived from psychoacoustic studies, by fitting a curve to Stevens & Volkman [8] data points, which gives a perceptual measure of pitch. By contrast, the proposed frequency warping function is a speech derived scale that maps physical frequency to an alternate domain, such that in the warped domain the speaker dependencies separate out as translation factors and it turns out to be very similar to mel scale. This indeed is very interesting and draws some relation between hearing mechanism and speech production. It also "justifies" the use of mel-scale in speech recognition, not only from the point of view of psychoacoustics but also from the view point of non-uniform speaker normalization.

4. COMPARISON OF VARIOUS SPEAKER NORMALIZATION METHODS

In this section, we present an account of the experiments that are being carried to investigate the effectiveness of various speaker normalization procedures. The normalization

% Word error rate	Adults		Children	
	E_b	E_n	E_b	E_n
$\eta_{mel}(f)$	3.02	2.52	13.73	7.96
$\log(f)$	3.24	2.85	13.04	9.20
$\psi^{pnb}(f)$	2.80	2.57	13.51	8.05
$\psi^{hil}(f)$	3.03	2.49	13.60	7.77

Table 1. Word error rate of various frequency warping functions on a digit recognizer before and after normalization. E_b and E_n represent the word error rates before (baseline) and after normalization respectively.

performance of different warping functions is evaluated by computing the word error rates on a telephone based connected word recognition task. The data for our digit recognition experiment is collected from *Numbers v1.0cd* corpus of OGI. The training set consists of 6078 utterances from adult male and female speakers. The performance of different normalization procedures are evaluated on two different test sets. The matched test set, dubbed as "Adults", is derived from Numbers corpus and consists of 2169 utterances from adult male and female speakers, while the mismatched test set, dubbed as "Children" is derived from other than Numbers corpus and consists of 2798 utterances from speakers having age between 6 to 18 years.

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 [4], with each frame being sectioned into Hamming windowed sub-frames of 64 samples and with an overlap of 45 samples. In our implementation, 64 points spaced non-uniformly depending on the warping function, are considered between $270-3850$ Hz. The non-uniformly sampled spectrum is then log compressed (in amplitude) and discrete cosine transform (DCT) is computed to derive cepstral features. 39 dimension feature vector comprising normalized log-energy, C_1 to C_{12} (excluding C_0) base cepstral coefficients and their first and second order derivatives are used. Finally, cepstral features are liftered and cepstral mean subtraction is performed.

Table 1 shows the word error rates of the digit recognizer, before and after normalization for different warping functions. The proposed warping functions based on the frequency-dependent scaling method perform better than log warp function and approach the performance of mel-warp function.

5. DISCUSSION & CONCLUSION

We have proposed a non-uniform speaker normalization method based on frequency-dependent scaling that requires no prior knowledge about the vowel category and formant number unlike Fant's method. The motivation for the proposed method is to apply Fant's method for non-uniform speaker normalization in ASR systems. The frequency warping function to perform non-uniform speaker normalization is similar to mel-scale, which justifies the use of mel-warp function in speech recognition not only from the point of psychoacoustics but also from the view point of speaker normalization. Interestingly, the warping function derived from the proposed method matches with the results of other independent studies [3, 9].

6. 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] 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.

[4] Rohit Sinha and S. Umesh, "Non-Uniform Scaling Based Speaker Normalization," in *Proc. IEEE ICASSP*, Orlando, USA, May 2002, pp. 589–592.

[5] 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.

[6] 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.

[7] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-Warping in Speech," in *Proc. ICSLP*, Philadelphia, USA, 1996.

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

[9] S. V. Bharath Kumar, S. Umesh, and Rohit Sinha, "Non-Uniform Speaker Normalization Using Affine-Transformation," in *Proc. IEEE ICASSP*, Montreal, Canada, May 2004.