STUDY OF NON-LINEAR FREQUENCY WARPING FUNCTIONS FOR SPEAKER NORMALIZATION

S. V. Bharath Kumar, S. Umesh and R. Sinha*

Department of ECE, University of California, San Diego, La Jolla, CA 92093-0407, USA.
Department of Electrical Engineering, Indian Institute of Technology, Kanpur - 208016, INDIA.
Cambridge University Engineering Department, Cambridge CB2 1PZ, UK.
{*bharathsv@ucsd.edu, sumesh@iitk.ac.in, rs460@eng.cam.ac.uk*}

ABSTRACT

In this paper, we study non-linear frequency-warping functions that are commonly used in speaker normalization. This study is motivated by our recently proposed affine transformation model for speaker normalization [1] which has provided improved recognition performance when compared to uniform scaling model [1, 2]. In this work, using formant data from Peterson & Barney and Hillenbrand vowel databases, we analyze the behavior of scale factor as a function of frequency. The empirical observation [3, 4] shows that while uniform scaling assumption may be valid at higher frequencies, there are significant deviations at low frequencies. We show that while our recently proposed model has behavior similar to the empirical result, the behavior of many of the commonly used non-linear models (including that of Eide-Gish, power law and bilinear transformation) differ significantly from the empirical result. This difference in behavior from the empirical observation may explain the limited improvement in recognition performance provided by these non-linear models when compared to conventional uniform-scaling model. We also show that our proposed model does better fitting to the formant data than these non-linear models. We, therefore, conclude that the affine-transformation model may be a more appropriate non-linear model for speaker normalization.

1. INTRODUCTION

One of the major factors affecting the performance of speaker-independent speech recognition is the variability in speech signal arising due to the physiological differences of vocal tract of speakers. Generally, as a first-order approximation, the vocal tract is assumed to be a tube of uniform cross-section. For this model, the formants of speakers would be frequency-scaled versions of one and another, i.e. $F_{\mathcal{R}} = \alpha_{\mathcal{RS}} F_{\mathcal{S}}$, where $\alpha_{\mathcal{RS}}$ is the ratio of vocal-tract lengths of reference speaker $\mathcal{R}$ and subject speaker $\mathcal{S}$. However, there have been numerous studies that show significant deviations from the uniform scaling assumption [5, 6].

Using actual speech data from Peterson & Barney (PnB) [7] and Hillenbrand (HiL) [8] vowel databases, we have empirically estimated the scale factor as a function of frequency in [3, 4]. It is to be noted that if the uniform scaling assumption were indeed true, the scale factor would be a constant independent of frequency. However, the empirical result shows that the speaker-specific scale factor changes as a function of frequency, or equivalently that non-linear frequency-warping is required for speaker normalization. In this paper, we are interested in comparing this empirical result obtained from actual speech data with the various non-linear models for speaker normalization that have been proposed in literature [1, 9, 10, 11].

We show that the non-linear warping function based on our recently proposed affine transformation method of [1] behaves similarly to the empirical observation, whereas the frequency-warping functions of [9, 10, 11] behave entirely different over all frequencies from the empirical observation. Therefore, we argue that our proposed non-linear model for speaker normalization is more appropriate than previously proposed ad-hoc models. This may also explain the limited success of these ad-hoc models in speaker normalization when compared to uniform scaling [9, 12], while the proposed model performs significantly better than uniform scaling and also approaches the performance of mel-warp function [1].

The paper is organized as follows. In Section 2, we review our empirically determined frequency-dependent scaling function, of Umesh *et al.* [3] and analyze its behavior for PnB and HiL. We briefly discuss the non-uniform normalization method of Bharath *et al.* [1] using affine transformation in Section 3 and show the corresponding frequency-warping function. In Section 4, we compare the non-linear frequency-warping functions based on our affine model, Eide-Gish model, power-law model and the bilinear transformation model with the empirical observation. We conclude based on above experiments that our proposed non-linear model is more appropriate for speaker normalization than these previously proposed ad-hoc models.

2. FREQUENCY-DEPENDENT SCALING METHOD

In [3], we reviewed the uniform and non-uniform vowel normalization methods of Nordström-Lindblom [13] and Fant [6] respectively. We also presented a modified version of Fant's non-uniform normalization scheme for both adult and child speakers. The original Fant's idea of non-uniform normalization is given by

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

where n is the formant number and j is the vowel class. This is in contrast to uniform scaling which only depends on the speaker-specific constant, k (or equivalently on the constant scale factor α, where $\alpha = (1 + \frac{k}{100})$). The non-uniform scaling arises due to $k_{n\mathcal{M}}^j$ which is the reference scale factor between the average female and

*A part of this work was done by Umesh under the Humboldt Research Fellowship and he gratefully acknowledges the foundation's support.

the average male for n^{th} formant of j^{th} vowel class. The database dependent constant φ is the scale factor between the average male and the reference speaker and is calculated to be -14.65 for PnB and -12.18 for HiL databases respectively. The non-uniform normalization scheme in (1) cannot be applied directly for speaker normalization on automatic speech recognition systems since it requires the knowledge of vowel category and formant number.

Umesh *et al.* [3] proposed the idea of frequency-dependent scale factor as a solution to the above problem. 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. Note that all non-linear frequency-warping methods [9, 10, 11] also model the scale factor as a function of frequency. The $k_{n\mathcal{M}}$ values are averaged over vowel category and formant number and over small frequency intervals to obtain discrete $\gamma(f_i)$ for each frequency interval. $\gamma(f)$ is obtained by a cubic spline fit to $\gamma(f_i)$ and is purely a function of frequency. The details of the method can be found in [4]. Extending from (1), the non-uniform normalization scheme using frequency-dependent scaling function is given by

$$\rho(k, f) = \gamma(f) \left(\frac{k}{\varphi}\right) \tag{2}$$

where $\rho(k, f)$ is the speaker and frequency-dependent scaling function but is independent of vowel category and formant number. The frequency-dependence comes from the *weighting* of $\gamma(f)$ on k. Hence, the normalization scheme in (2) can be used directly in a speech recognizer unlike Fant's method. Alternatively, because of the way we have defined α in the experiments (which is different from Fant's notation where $\alpha_{fant} = \frac{F_{\mathcal{S}}}{F_{\mathcal{R}}} = \alpha_{\mathcal{RS}}^{-1}$), we can re-write (2) as

$$\tilde{\alpha}_f(f) = \left(1 + \frac{\rho(k, f)}{100}\right)^{-1} = \left(1 + \frac{\gamma(f)(\alpha - 1)}{\varphi}\right)^{-1} \tag{3}$$

Figure 1 shows the plot of $\gamma(f)$ for PnB and HiL databases. Although we will use $\gamma(f)$ for most of the discussion, it might be easier to understand the weighting function in Figure 1 if the reader considers $\frac{\gamma(f)}{\varphi}$. For both databases, $\gamma(f)$ behaves similarly for $f \geq 1600$ Hz and tends to be constant at higher frequencies. This agrees with the uniform scaling (i.e. constant scale factor) assumption made for higher formants. However, there seems to be an anomaly in behavior between the two databases at low-frequencies.[1] This is particularly interesting since $\gamma(f)$ for PnB seems to suggest that scale factor weighting decreases at low frequencies, while the more recent HiL database suggests exactly the opposite.

In the subsequent sections we will compare this empirically obtained $\gamma(f)$ (seen above) and corresponding $\tilde{\alpha}_f(f)$ with those obtained from non-linear models proposed in literature.

3. AFFINE TRANSFORMATION METHOD

In [1], we have proposed a non-uniform speaker normalization scheme using the following affine transformation model relating

[1] As [8] points out, despite the widespread use of the PnB measurements there are several well recognized limitations to the database. For example, there is no indication that subjects were screened for dialect or there is information about the age and gender of child speakers in PnB. Hence, for our experiments, it is important that we model the HiL data properly.

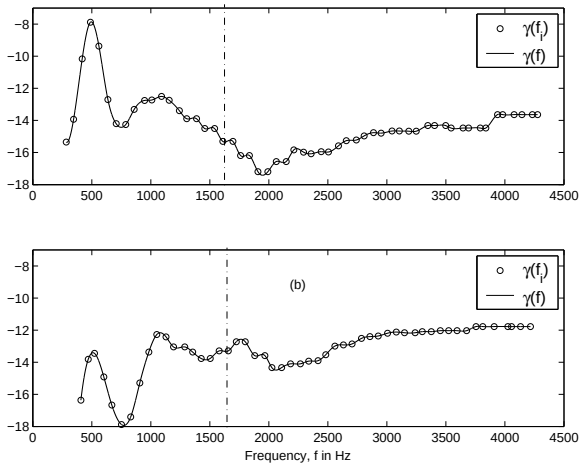

Fig. 1. Frequency-dependent scale factors, $\gamma(f_i)$ and frequency-dependent scaling function, $\gamma(f)$ for (a) Peterson & Barney and (b) Hillenbrand databases. It might be easier to understand the weighting function if the reader considers $\frac{\gamma(f)}{\varphi}$, where $\varphi = -14.65$ for Peterson & Barney and $\varphi = -12.18$ for Hillenbrand databases.

formant frequencies of the subject and the reference speakers as

$$(F_{\mathcal{R}} + \mathtt{A}) = \alpha_{\mathcal{RS}} (F_{\mathcal{S}} + \mathtt{A}) \tag{4}$$

where $F_{\mathcal{R}}$, $F_{\mathcal{S}}$ are formant frequencies of the reference speaker, $\mathcal{R}$ and the subject speaker, $\mathcal{S}$ respectively. The average female speaker of the database is chosen to be the reference speaker. $\mathtt{A}$ and $\alpha_{\mathcal{RS}}$ are assumed to be speaker-independent and speaker dependent parameters respectively and are estimated from speech data. The value of speaker-independent $\mathtt{A}$ has been estimated to be 508.04 for PnB database and 495.67 for HiL database. If we consider the average male and average female as two speakers in the database, then the corresponding $\alpha_{\mathcal{RS}}$ of the average male speaker is computed as 1.14 and 1.12 for PnB and HiL databases respectively.

One of the main motivations for choosing this model is that the transformation (similar to Bark-scale) $\nu = \log\left(1 + \frac{f}{\mathtt{A}}\right)$ makes the speaker-dependent scale factor to separate out as translation factor in ν domain. This is interestingly similar to one approach of speaker normalization which is based on applying different offsets (translations) in Bark scale for different speakers [14]. Therefore, this model provides a unifying theory between the frequency-warping and bark-scale-shift approaches to speaker normalization. In general, the frequency-warping function for the affine transformation model using (4) is given as

$$f_r = \alpha_s f_s + \mathtt{A}(\alpha_s - 1) \tag{5}$$

and the corresponding frequency-scaling relation can be written as $\tilde{\alpha}_a(f) = \alpha_s + \frac{\mathtt{A}(\alpha_s - 1)}{f}$, where α_s is the speaker-dependent scale factor.

4. COMPARISON OF NON-UNIFORM SPEAKER NORMALIZATION METHODS

In the following discussion we will formulate all non-uniform speaker normalization schemes in the following framework,

$$f_r = g(\alpha_s, f_s) = \tilde{\alpha}(f_s) f_s \tag{6}$$

PnB	Empirical	$\tilde{\alpha}_f(f_s) = \left(1 + \frac{\gamma(f_s)}{100}\right)^{-1}$
	Affine	$\tilde{\alpha}_a(f_s) = 1.14 + 71.13 f_s^{-1}$
	Eide-Gish	$\tilde{\alpha}_e(f_s) = 1.2^{\frac{3f_s}{8000}}$
	Power	$\tilde{\alpha}_p(f_s) = \left(\frac{8000}{f_s}\right)^{0.11}$
	Bilinear	$\tilde{\alpha}_b(f_s) = \frac{4000}{\pi f_s} \tan^{-1} \frac{0.98 \sin\left(\frac{\pi f_s}{4000}\right)}{1.02 \cos\left(\frac{\pi f_s}{4000}\right) - 0.28}$
HiL	Empirical	$\tilde{\alpha}_f(f_s) = \left(1 + \frac{\gamma(f_s)}{100}\right)^{-1}$
	Affine	$\tilde{\alpha}_a(f_s) = 1.12 + 59.48 f_s^{-1}$
	Eide-Gish	$\tilde{\alpha}_e(f_s) = 1.16^{\frac{3f_s}{8000}}$
	Power	$\tilde{\alpha}_p(f_s) = \left(\frac{8000}{f_s}\right)^{0.1}$
	Bilinear	$\tilde{\alpha}_b(f_s) = \frac{4000}{\pi f_s} \tan^{-1} \frac{0.98 \sin\left(\frac{\pi f_s}{4000}\right)}{1.02 \cos\left(\frac{\pi f_s}{4000}\right) - 0.25}$

Table 1. Frequency-dependent scale factor, $\tilde{\alpha}(f_s)$ for various non-uniform speaker normalization schemes with average male and average female to be the subject and reference speakers from Peterson & Barney (PnB) and Hillenbrand (HiL) databases.

where $g(\alpha_s, f_s)$ is the frequency-warping function and $\tilde{\alpha}(f_s)$ is the frequency-dependent scale factor of the subject with respect to reference speaker. Both $g(.)$ and $\tilde{\alpha}(.)$ depend on the speaker-dependent term α_s, though for the simplicity of notation, we do not explicitly show α_s as a parameter of $\tilde{\alpha}(.)$. Further, we will consider the specific case of average male and average female speaker and find the corresponding α_s for each case.

For the empirically determined frequency-dependent scaling function of (3), we have $k = \varphi$ since we are considering the average male and female speakers. The corresponding frequency-dependent scaling function is

$$\tilde{\alpha}_f(f_s) = \left(1 + \frac{\gamma(f_s)}{100}\right)^{-1} \tag{7}$$

Similarly, using (5), $\tilde{\alpha}(f_s)$ for affine transformation method is given as

$$\tilde{\alpha}_a(f_s) = \alpha_s + \frac{\mathtt{A}(\alpha_s - 1)}{f_s} \tag{8}$$

For the case of average male and average female speakers, $\alpha_s = 1.14$ for PnB and $\alpha_s = 1.12$ for HiL.

Eide *et al.* [9] proposed the following parametric form for non-uniform normalization, given as

$$f_r = g(k_s, f_s) = k_s^{\frac{3f_s}{8000}} f_s \tag{9}$$

$$\Rightarrow \tilde{\alpha}_e(f_s) = k_s^{\frac{3f_s}{8000}} \tag{10}$$

where k_s is the subject's scale factor and $\tilde{\alpha}_e(f_s)$ is the frequency-dependent scale factor for Eide-Gish normalization. The motivation for the choice of this model is based on Fant's observation that for the first formant of some vowels like $/IY/$, it might be better to use $\sqrt{k_s}$ as normalization factor. This observation seems to be consistent with PnB data, but the behavior of HiL data is quite different. Further, in the Eide-Gish model the frequency-dependent scale factor is monotonic (does not saturate at high frequencies as observed empirically) and has the value of $\sqrt{k_s}$ only at $f = \frac{4000}{3}$ Hz.

The bilinear transformation [10, 11] and power law are two other commonly used non-linear models that are motivated by the fact that they can "reasonably" approximate piece-wise linear warping of uniform scaling. The frequency-dependent scaling function for the power law, $\tilde{\alpha}_p(f_s)$ and the bilinear transformation, $\tilde{\alpha}_b(f_s)$ are given as

$$\tilde{\alpha}_p(f_s) = \left(\frac{f_s}{f_N}\right)^{\beta_s - 1} \tag{11}$$

$$\tilde{\alpha}_b(f_s) = \frac{f_N}{2\pi f_s} \tan^{-1} \frac{(1 - a_s^2) \sin\left(\frac{2\pi f_s}{f_N}\right)}{(1 + a_s^2) \cos\left(\frac{2\pi f_s}{f_N}\right) - 2a_s} \tag{12}$$

respectively, where f_N is the Nyquist frequency (assumed 8 KHz in this paper). β_s and a_s are the speaker-dependent parameters of the models in (11) and (12) respectively. Also, $|a_s| < 1$. Similar to the affine model, k_s, β_s and a_s are computed by fitting $f_r = \tilde{\alpha}(f_s) f_s$ using respective $\tilde{\alpha}(f_s)$ from (10-12) to the data points involving the formant frequencies of the average female and average male speakers. We found that $k_s = 1.20$, $\beta_s = 0.89$, $a_s = 0.14$ for PnB and $k_s = 1.16$, $\beta_s = 0.9$, $a_s = 0.13$ for HiL databases.

Table 1 shows the frequency-dependent scaling function, $\tilde{\alpha}(f_s)$ for various non-uniform speaker normalization schemes using the formant data of average male and average female from PnB and HiL databases. Figure 2 and Figure 3 show the plot of $\tilde{\alpha}(f_s)$ for aforementioned normalization schemes based on Table 1 for PnB and HiL databases. It is clear from Figure 2 and Figure 3 that the empirical scaling function, $\tilde{\alpha}_f(f_s)$ derived from both PnB and HiL databases tend to be constant at higher frequencies. In the case of affine model, $\tilde{\alpha}_a(f_s) \rightarrow \alpha_s$, for $f_s \gg \mathtt{A}$ and exhibits similar behavior w.r.t. empirical $\tilde{\alpha}_f(f_s)$ over $f_s \geq 1600$ Hz for both PnB and HiL databases. The Eide-Gish model, power law and bilinear transformation exhibit entirely different behavior from $\tilde{\alpha}_f(f_s)$ for all f_s.

While work on the use of other non-linear warping functions have not reported any significant improvement in recognition, we have obtained about 8% relative improvement when compared to the linear-scaling model as reported in [1, 2]. The affine method models the empirical data quite well as shown in Figure 2 and Figure 3 whereas Eide-Gish, power law and bilinear transformation models, which are proposed in an ad-hoc manner, behave entirely different from empirically computed scaling function. This may also explain the limited success of these ad-hoc models in terms of recognition performance when compared to uniform scaling.

We also made a study on the goodness of fit of $f_r = \tilde{\alpha}(f_s) f_s$ to the formant data of PnB and HiL using the appropriate scaling functions mentioned in Table 1. Let $\tilde{\epsilon}$ be the model-fitting error involved while using $\tilde{\alpha}(f_s)$ as the frequency-dependent scaling function, which is defined as $\tilde{\epsilon} = \|f_r - \tilde{\alpha}(f_s) f_s\|_2$. Table 2 shows that the affine model fits the formant data better than other proposed ad-hoc models for both the databases, as $\tilde{\epsilon}$ is minimum for affine model compared to other models. Hence, this experi-

$\tilde{\epsilon}$	Peterson & Barney	Hillenbrand
Affine	437.58	297.49
Eide-Gish	572.77	566.10
Power	595.86	438.83
Bilinear	617.53	486.25

Table 2. Model-fitting error, $\tilde{\epsilon}$ for various non-uniform speaker normalization schemes with average male and average female to be the subject and reference speakers from Peterson & Barney and Hillenbrand databases.

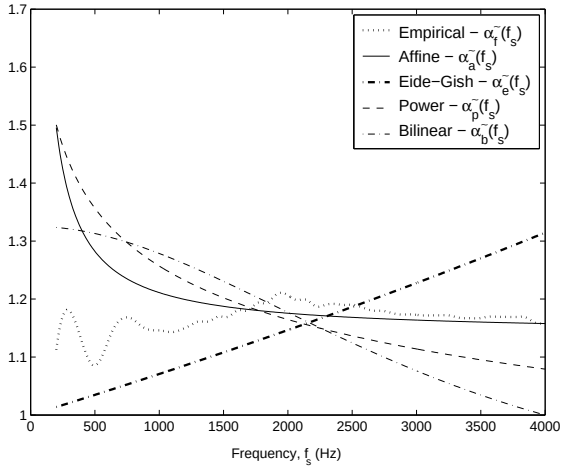

Fig. 2. Frequency-dependent scale factor, $\tilde{\alpha}(f_s)$ for various non-uniform speaker normalization schemes for Peterson & Barney database. Note that except for affine-model all the other models behave quite differently from the empirical result. The affine model has different behavior at low frequencies but matches the empirical result at high frequencies.

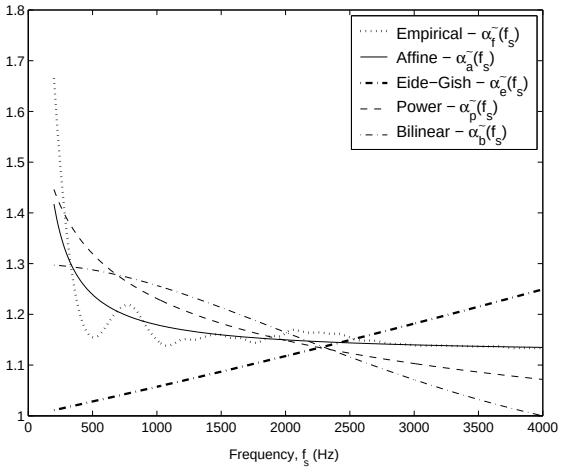

Fig. 3. Frequency-dependent scale factor, $\tilde{\alpha}(f_s)$ for various non-uniform speaker normalization schemes for Hillenbrand database. The affine-model matches the empirical result very closely.

ment again confirms that affine model is a better model for normalization when compared to other ad-hoc normalization models.

5. DISCUSSION & CONCLUSION

We analyze empirically computed frequency-dependent weighting function for Peterson & Barney (PnB) and Hillenbrand (HiL) vowel databases, which shows that while uniform scaling assumption might be true at high frequencies, there are significant deviations at low frequencies. Further, the behavior of PnB and HiL differ quite significantly at low frequencies. We compared our earlier proposed affine model with the empirical scaling function along with other ad-hoc non-linear normalization models of Eide-Gish, power-law and bilinear transformation. Our experiments indicate that affine model behaves similar to the empirical method at higher frequencies, whereas other models behave entirely different at all frequencies. Further, for HiL data our model matches the empirical result quite closely at all frequencies. In terms of recognition performance, while the affine model provides 8% relative improvement over uniform scaling [1, 2], the other non-linear models have not provided any significant improvement over uniform scaling model [9, 12]. Therefore, the affine model may be a better model for speaker normalization than the other previously proposed non-linear models.

6. REFERENCES

[1] S. V. Bharath Kumar, S. Umesh, and Rohit Sinha, "Non-Uniform Speaker Normalization Using Affine-Transformation," in *Proc. IEEE ICASSP*, Montreal, Canada, May 2004, pp. 121–124.

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

[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] S. V. Bharath Kumar and S. Umesh, "Non-Uniform Speaker Normalization Using Frequency-Dependent Scaling Function," in *Proc. 2004 Intl. Conf. on Signal Proc. and Communications*, Bangalore, India, December 2004, pp. 305–309.

[5] 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 2002.

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

[7] 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.

[8] 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.

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

[10] A. Acero and R. M. Stern, "Robust Speech Recognition by Normalization of the Acoustic Space," in *Proc. IEEE ICASSP*, Toronto, Canada, May 1991, pp. 893–896.

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

[12] S. Molau, S. Kanthak, and H. Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition," in *Proc. ESSV*, Cottbus, Germany, 2000.

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

[14] D. C. Burnett and M. Fanty, "Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task," in *Proc. IEEE ICSLP*, Philadelphia, USA, 1996.