

A New Spectral Transformation for Speaker Normalization

Pierre L. Dognin, Amro El-Jaroudi

Department of Electrical Engineering
University of Pittsburgh. Pittsburgh, PA. 15261

dognin@siglab.ee.pitt.edu, amro@ee.pitt.edu

Abstract

This paper proposes a new spectral transformation for speaker normalization. We use the Bilinear Transformation (BLT) to introduce a new frequency warping resulting from a mapping of a prototype Band-Pass (BP) filter into a general BP filter. This new transformation called “Band-Pass Transform” (BPT) offers two degrees of freedom enabling complex warpings of the frequency axis and different from previous works with BLT. A procedure based on the Nelder-Mead algorithm is proposed to estimate the BPT parameters. Our experimental results include a detailed study of the performance of the BPT compared to other VTLN methods for a subset of speakers and results on large test sets. BPT performs better than other VTLN methods and offers a gain of 1.13% absolute on Hub-5 English Eval01 set.

1. Introduction

One of the major challenges for Automatic Speech Recognition (ASR) is to handle speech variability. Inter-speaker variability is partly due to differences in speakers’ anatomy and especially in their Vocal Tract (VT) geometry. Dissimilarities in Vocal Tract Length (VTL) is a known source of speech variation. Therefore, VTL Normalization (VTLN) became a popular Speaker Normalization technique and it can be implemented as a transformation of a spectrum frequency axis. A simple linear warping of the frequency axis could be used in VTLN if our VT were a lossless tube of varying length. Unfortunately, the physical shape of our VT is a lot more complex and formants in speech are not linearly warped across voiced phonemes when the VTL varies. To address this fact, researchers started to look into nonlinear parametric spectral transformations for VTLN[1, 2]. The transformation proposed in [1] was chosen to accommodate most of the voiced phonemes:

$$f_w = k_s \left(\frac{3f}{8,000} \right) \times f \quad (1)$$

where f is the original frequency axis and f_w is the warped frequency axis. The frequency warping is controlled by one Speaker Dependent (SD) parameter k_s called *warping factor*.

The frequency warping properties of the Bilinear Transformation (or BLT) has not gone unnoticed and its application to VTLN has been the focus of some research effort. BLT properties have been primarily used in filter design and discrete-time series transformation[3]. For VTLN, the main interest of BLT is its ability to offer an efficient technique to perform frequency warping. Indeed, BLT offers a means to transform a discrete-time sequence into another sequence. The transformed sequence has the same Fourier Transform as the original sequence but with a warped frequency axis. If the sequence to

This research was funded by BBN Technologies, Cambridge, MA.

transform is chosen to be the cepstral speech features, then VTLN can be performed using BLT[2, 4]. In [2], a “First Order BLT” (or 1st-BLT) is used to perform a mapping in \mathbb{C} identical to the one that transforms a prototype Low-Pass (LP) filter into a desired LP filter. Let $w^{-1} = e^{-j\psi}$ be the complex variable on the unit circle associated with the prototype LP filter and $z^{-1} = e^{-j\varphi}$ the variable associated with the desired filter. In order to transform the prototype filter into the desired filter, we need to replace the variable w^{-1} by $G(z^{-1})$ in the Z-Transform representation of the prototype filter. This mapping of z^{-1} to w^{-1} is given by

$$z^{-1} \mapsto w^{-1} = M_a(z^{-1}) = \frac{z^{-1} - a}{1 - az^{-1}}, \quad a \in \mathbb{R}, |a| < 1 \quad (2)$$

that results in the frequency warping

$$\varphi = \psi - 2 \arctan \left(\frac{a \sin \psi}{1 + a \cos \psi} \right) \quad (3)$$

which is controlled by the real parameter a as presented in [5]. A natural extension of this approach is the All-Pass Transform (APT) allowing mappings with more than one parameter[4]. However, in regards to filter design BLT and APT provide mappings that correspond to the transformation of a Low-Pass (LP) prototype filter into another filter being LP, BP, High-Pass or of more complex transfer function (for APT). In this paper, we present a new approach that defines a new mapping which is not restricted to start from a LP filter. We are especially interested in the frequency warping that results from the transformation of a prototype BP filter into a general BP filter.

2. A new frequency warping for VTLN

We know from [5] that the mapping that transforms a prototype LP filter with cutoff frequency θ_p into a desired BP filter with cutoff frequencies ω_1 and ω_2 such that $\omega_1 < \omega_2$ is given by

$$z^{-1} \mapsto w^{-1} = G(z^{-1}) = -\frac{z^{-2} - \frac{2\alpha k}{k+1}z^{-1} + \frac{k-1}{k+1}}{\frac{k-1}{k+1}z^{-2} - \frac{2\alpha k}{k+1}z^{-1} + 1} \quad (4)$$

where the parameters α and k are defined by

$$\alpha = \frac{\cos \left(\frac{\omega_2 + \omega_1}{2} \right)}{\cos \left(\frac{\omega_2 - \omega_1}{2} \right)} \quad \text{and} \quad k = \frac{\tan \left(\frac{\theta_p}{2} \right)}{\tan \left(\frac{\omega_2 - \omega_1}{2} \right)}. \quad (5)$$

Interestingly, if we define $\gamma = \frac{k-1}{k+1}$, then (4) can be rewritten into

$$w^{-1} = -\frac{z^{-2} - \alpha(1 + \gamma)z^{-1} + \gamma}{1 - \alpha(1 + \gamma)z^{-1} + \gamma z^{-2}} \quad (6)$$

$$\Leftrightarrow w^{-1} = -\left\{ \frac{z^{-1} \left(\frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \right) + \gamma}{1 + \gamma z^{-1} \left(\frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \right)} \right\} \quad (7)$$

that offers an intuitive understanding of the different steps necessary to transform a prototype LP filter into a BP filter. The transformation in (7) is a combination of functions such that:

$$w^{-1} = -M_{-\gamma}(z^{-1}M_{\alpha}(z^{-1})) \quad \alpha, \gamma \in \mathbb{R}, |\alpha|, |\gamma| < 1 \quad (8)$$

where $M(z^{-1})$ comes from (2). If $|\gamma| < 1$, we notice that $k > 0$.

2.1. Definition of a new mapping

The new frequency warping we propose for Speaker Normalization is based on a new mapping that transforms a prototype BP filter into a general BP filter. The mapping's definition depends entirely on the cutoff frequencies of the two BP filters. The prototype BP filter is obtained from a prototype LP filter with cutoff frequency $\theta_p = \frac{\pi}{2}$ associated with the variable Z^{-1} . The prototype BP filter associated with the variable z^{-1} is chosen to have $\omega_{p1} = \frac{\pi}{4}$ and $\omega_{p2} = \frac{3\pi}{4}$ for cutoff frequencies. In order to transform the prototype LP filter into our prototype BP filter, we know from (4) and (5) that we need to replace Z^{-1} with the mapping

$$Z^{-1} = -z^{-2}. \quad (9)$$

Similarly, if we want to transform the prototype LP filter into a general BP filter of variable \hat{z}^{-1} with no assumptions on the cutoff frequencies, we use the mapping

$$Z^{-1} = -\frac{\hat{z}^{-2} - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma}{1 - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma\hat{z}^{-2}} \quad (10)$$

where the parameters α and γ (and therefore k) depend on the cutoff frequencies of both LP and BP filters. Finally, the transformation of our prototype BP into a general BP filter is achieved by the following mapping:

$$\begin{aligned} -z^{-2} &= -\frac{\hat{z}^{-2} - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma}{1 - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma\hat{z}^{-2}} \\ \Leftrightarrow z^{-2} &= \frac{\hat{z}^{-2} - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma}{1 - \alpha(1 + \gamma)\hat{z}^{-1} + \gamma\hat{z}^{-2}} \end{aligned} \quad (11)$$

where $z^{-1} = e^{-j\psi}$ is the old variable replaced by $\hat{z}^{-1} = e^{-j\varphi}$, the new variable. The fact that we obtain the old variable as a function of the new variable is not an issue in regards to deriving the resulting frequency warping, which is discussed in the next section. Even if the Right Hand Side (RHS) of the mapping in (11) can be expressed as a product of 1st-BLT terms, this mapping is different from other mappings defined from APT because the Left Hand Side (LHS) is z^{-2} instead of z^{-1} .

2.2. Frequency warping

In order to derive the frequency warping that results from (11), we need to know the relation between the arguments of the equation's LHS and RHS. Since $z^{-1} = e^{-j\psi}$ and $\hat{z}^{-1} = e^{-j\varphi}$, we can rewrite (11) into

$$\begin{aligned} e^{-j2\psi} &= \frac{e^{-j2\varphi} - \alpha(1 + \gamma)e^{-j\varphi} + \gamma}{1 - \alpha(1 + \gamma)e^{-j\varphi} + \gamma e^{-j2\varphi}} \\ &= e^{-j2\varphi} \frac{1 - \alpha(1 + \gamma)e^{j\varphi} + \gamma e^{j2\varphi}}{1 - \alpha(1 + \gamma)e^{-j\varphi} + \gamma e^{-j2\varphi}} \\ &= e^{-j2\varphi} \frac{1 - \alpha(1 + \gamma)e^{j\varphi} + \gamma e^{j2\varphi}}{(1 - \alpha(1 + \gamma)e^{j\varphi} + \gamma e^{j2\varphi})^*} \end{aligned} \quad (12)$$

where the notation z^* is used for the complex conjugate of z . Since $\arg\left(\frac{z}{z^*}\right) = 2\arg(z)$, we can derive the relation between

old and new frequency by taking the argument of the LHS and RHS of (12):

$$\begin{aligned} -2\psi &= -2\varphi + 2\arg(1 - \alpha(1 + \gamma)e^{j\varphi} + \gamma e^{j2\varphi}) \\ \Leftrightarrow \psi &= \varphi - \arg(1 - \alpha(1 + \gamma)e^{j\varphi} + \gamma e^{j2\varphi}) \\ \Leftrightarrow \psi &= \varphi - \underbrace{\arctan\left(\frac{-\alpha(1 + \gamma)\sin\varphi + \gamma\sin 2\varphi}{1 - \alpha(1 + \gamma)\cos\varphi + \gamma\cos 2\varphi}\right)}_{\rho(\varphi; \alpha, \gamma)} \end{aligned} \quad (13)$$

where ψ is the old frequency, φ is the transformed or new frequency and $\rho(\varphi; \alpha, \gamma)$ can be considered as a ‘‘correction term’’. With $\gamma = \frac{k-1}{k+1}$, (13) is in reality $\psi = f^{-1}(\varphi; \alpha, k)$ and is referred to as the ‘‘Band-Pass Transform’’ or BPT. The fact that we obtain an expression of the old frequency ψ as a function of the new frequency φ is not an issue *per se*. If the analytic form for the inverse frequency warping is known as $\psi = f^{-1}(\varphi; \alpha, k)$, it is possible to implement $\varphi = f(\psi; \alpha, k)$ by means of a search algorithm. It is a simple solution to find the values of f when only the analytic form f^{-1} is known as long as f^{-1} is one-to-one and monotonous, which is the case for the BPT.

This new frequency warping depends on the parameters α and k as seen in Figure (1). From (13), it can be noticed that for $\alpha = 0$ and $k = 1$ ($\gamma = 0$), no warping is performed as the equation reduces to $\psi = \varphi$. In order to analyze the properties of

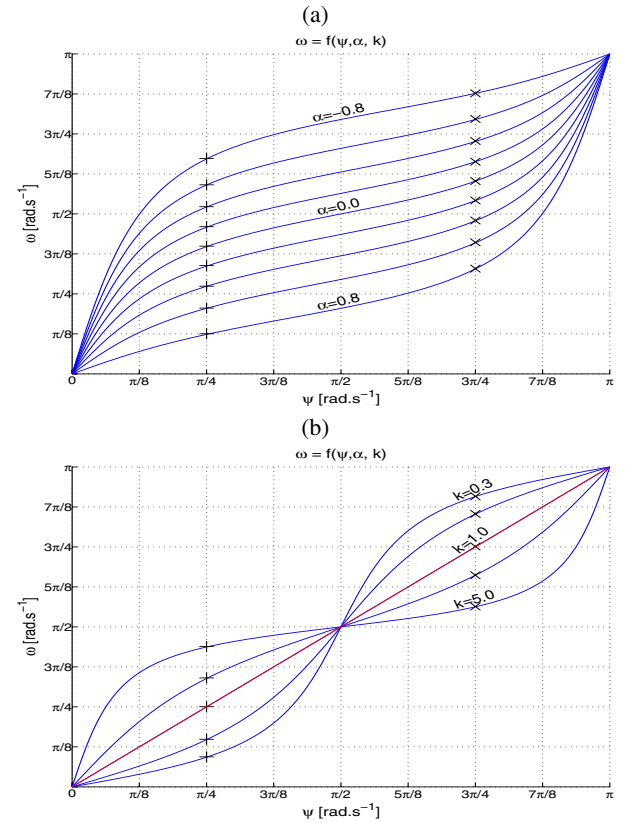


Figure 1: New frequency warping (a) for $k = 3$ and $\alpha = -0.8$ to 0.8 with a 0.2 step. (b) for $\alpha = 0$ and $k = \{0.3, 0.5, 1.0, 1.8, 5.0\}$.

this new frequency transformation, we study the behavior of its ‘‘control points’’. By control points we mean special points that condition the shape of the transformation's curve. In our case,

those control points are the DC-frequency ($\varphi=0$), the Nyquist frequency ($\varphi=\pi$) and the two cutoff frequencies $\omega_{p1}=\frac{\pi}{4}$ and $\omega_{p2}=\frac{3\pi}{4}$. The DC and Nyquist frequencies are mapped to themselves $\forall\alpha, k$ since $f(0; \alpha, k)=0$ and $f(\pi; \alpha, k)=\pi$. The cutoff frequencies are moved on the frequency axis accordingly to (13). On each curve in Figure (1) the marks “+” and “x” correspond respectively to the transforms of ω_{p1} and ω_{p2} , the two cutoff frequencies of our prototype BP filter. For a sampling frequency of 8000 Hz, ω_{p1} corresponds to $F_{p1}=1000$ Hz located in the middle of the first and second formant region while ω_{p2} corresponds to $F_{p2}=3000$ Hz located near the third formant region[6]. Therefore, the mapping of ω_{p1} and ω_{p2} illustrates the manner the formants’ locations change along the frequency axis. In Figure (1) we can see how the transformation changes with α and k . If we keep k constant and have α vary as in Figure (1.a), the ω_{p1} and ω_{p2} transforms slide up or down the frequency axis; and so will the formants. If we keep α constant and have k vary as in Figure (1.b), the larger k , the closer together the cutoff frequencies’ images move; and so will the formants. The combination of varying α and k provides a wide range of possible frequency axis transformations with interesting properties for VTLN. Each transformation has the physical meaning of transforming a BP filter into another BP filter.

3. Experimental setup and results

Previously, a new frequency transformation for VTLN has been defined and we need to establish the possible improvements it can offer compared to other VTLN methods. The experimental setup is similar to the one described in details in [7] as we utilize the BYBLOS system developed by BBN Technologies for the 2001 Hub-5 Evaluation Benchmark organized by NIST. The gender dependent acoustic models are built from a 40hours gender-balanced set. For all experimental results, VTLN is performed in Testing *only* and not in Training. Testing is performed on two different available sets: The Hub-5 English Dev01 set composed of 48 speakers (2hour long) and the Hub-5 English Eval01 set with 120 speakers (6hour long). For each set, a baseline experiment corresponds to a decoding without any VTLN.

The BPT is controlled by two parameters and there is a need to define a procedure to estimate a set of parameters $\beta_s = [\alpha_s, k_s]$ for each speaker s . As proposed in [8], the parameters can be chosen to maximize the likelihood of the sequence of observed transformed features \mathbf{X}_s^β given an acoustic model Λ and the transcripts \mathbf{W}_s of all the spoken utterances.

$$\hat{\beta}_s = \arg \max_{\beta} Pr(\mathbf{X}_s^\beta | \Lambda, \mathbf{W}_s) \quad (14)$$

A closed-form solution is not easy to find for (14) due to the non-linear transformation to the frequency warping performs from unwarped features \mathbf{X}_s to \mathbf{X}_s^β as mentioned in [8]. Any solution based on an optimization method requires numerical evaluations of the function $Pr(\mathbf{X}_s^\beta | \Lambda, \mathbf{W}_s)$ in (14). In our system, such evaluation is performed in two steps. First, a forward-backward decoding provides a n-best list consisting of the 100 best transcript hypotheses for the observed sequence \mathbf{X}_s^β . Second, this n-best list is “rescored” with a more detailed acoustic model and provides an acoustic and language model score for each of the 100 hypotheses. The best hypothesis for each utterance has the maximum global score. In our case, only the acoustic score will be used. The most time consuming task here is the creation of the n-best list. For a large set of speakers, this task can become computationally overwhelming. One way to decrease the computation time is to get the n-best lists

from a previous baseline experiment. Only a rescaling of the n-best lists (referred to as “n-best rescaling”) is performed. If the speed up is not sufficient, another more radical step is to get the best hypothesis from the baseline decoding. Then, only the best hypothesis for each utterance is rescored instead of the 100 in the n-best list. It is referred to as a “1-best rescaling”. Both techniques are used in our experimental results.

3.1. Nelder-Mead optimization method

The Nelder-Mead (NM) unconstrained optimization method is utilized to estimate our parameter set β_s . A detailed presentation of the NM algorithm and of its convergence properties is given in [9]. It is usually used to find a function’s minimum but can be easily modified for the search of a maximum. The NM algorithm is based on updating an initial *simplex*. In the 2-dimensional case, a simplex is composed of 3 points evaluated at different values of β . After each algorithm iteration, the simplex point with the worst score is replaced by a point of better score. In the special case where after few function evaluations no better point is found, the algorithm “shrinks” the simplex, keeping only the best score point. The algorithm first tries to find better points based on the assumption that a point geometrically further from the worst point could be an improvement, but modifies its strategy if it’s not the case. As a consequence, for a 2-dimensional search, each iteration consists of 1 to 4 function evaluations. After several iterations, it is expected that the simplex would move towards a region with a function’s maximum (in our case). One advantage of the NM approach is to require no evaluation of first or second derivative of our function which could be computationally costly. For our experimental results, we chose to stop the NM algorithm after 50 function evaluations which is approximately 20 iterations.

3.2. Experimental results for the BPT

The first step for our experimental results was to compare the BPT to other VTLN methods. Those methods perform a linear frequency warping (Linear VTLN), a “Eide” warping (Eide VTLN) as described in (1) or a 1st-BLT warping (1st-BLT VTLN) as presented in (3). All require the estimation of a SD parameter. In order to avoid having our results biased by the parameter estimation procedure, we opted in this case for the solution of *Oracle* experiments. An Oracle experiment (or simply Oracle) is a grid search of the parameter value that minimizes the system’s Word Error Rate (WER). The advantage of an Oracle is to provide the *best* WER attainable for the parameter values *on the grid*. However, since the WER is the objective function, the main disadvantages reside in having to run a decoding for each parameter value and to know the true transcription for the observed speech which makes an Oracle impractical in many cases. For the estimation of one parameter it is still computationally acceptable. For the BPT VTLN, the estimation of two SD parameters is required. While an Oracle on all test speakers is not imaginable, we can reduce the number of speakers dramatically to retain the Oracle solution. This will allow us to investigate thoroughly the BPT properties by allowing a fine sampling for our grid search. More importantly, it will also give a target value for β_s and a WER performance to attain when we estimate our parameter with the NM algorithm. In this case, results from Oracle experiments inform us on what to expect in term of possible performance for the BPT.

Two female speakers SW04537A and SW20316A from the Dev01 set are selected based on several criteria such as number of uttered words, improvement from VTLN methods, etc.

Oracle experiments were run for all VTLN methods. For the BPT, a fine grid search consisted of 41 values for α on the $[-0.20, 0.20]$ interval and 41 values for k in the $[0.80, 1.20]$ interval which results in $N_p = 1681$ possible (α, k) pairs. The WERs for both speakers for all VTLN methods available are presented in Table 1. N_p indicates the number of points in the grid search. For both speakers, the BPT offers the best perfor-

VTLN Method	sw04537A		sw20316A		N_p
	Errors	WER %	Errors	WER %	
no VTLN	162	30.62	232	50.00	1
Linear	127	24.01	221	47.63	41
Eide	128	24.20	221	47.63	41
1st BLT	131	24.76	219	47.20	41
BPT	115	21.74	216	46.55	1681
Reference	529 words		464 words		

Table 1: WERs for all available VTLN methods.

mances. For sw04537A, with a 30.62% baseline, an improvement of a 8.88% absolute gain can be achieved, 2.27% absolute better than the Linear VTLN, second best. Speaker sw20316A, with a 50.0% baseline, sees an improvement of 3.45% absolute accomplished by BPT, 0.65% absolute better than second best 1st-BLT. Table 2 presents the results for using a Maximum Likelihood (ML) approach to find β_s for the BPT alone. The

ML Method	sw04537A		sw20316A		N_e
	Errors	WER %	Errors	WER %	
Oracle Grid	121	22.87	223	48.06	1681
NM n-best	123	23.25	228	49.14	50
NM 1-best	123	23.25	219	47.20	50

Table 2: Results for ML selection of the BPT parameters.

N_e column indicates the number of likelihood function evaluations. The results on the ‘‘Oracle Grid’’ row are for a ML selection for β_s using the same grid as for the Oracle experiment. These results are our reference for ML selection of β_s . The following rows present results when the NM approach is utilized. Interestingly, the NM methods (n-best and 1-best rescoring) offer both encouraging results for sw04537A with only a slight increase in WER. For sw20316A, the case is a bit different as the NM n-best offers an improvement compared to the Oracle Grid results. In order to better understand these results, the ML values selected for α_s and k_s are shown in Table 3. From Ta-

Selection Method	sw04537A		sw20316A		N_e
	α_s	k_s	α_s	k_s	
Oracle	0.17	1.17	-0.05	1.0	1681
Oracle Grid	0.14	1.20	-0.05	1.01	1681
NM n-best	0.1340	1.2235	-0.0597	0.9904	50
NM 1-best	0.1341	1.2235	-0.0506	1.0058	50

Table 3: Parameters selection for both speakers.

ble 3, it is clear that NM algorithm selected values are quite close to the Oracle Grid values for α_s and k_s . The *Jacobian* of the transformation from \mathbf{X}_s to \mathbf{X}_s^β is not taken into account in our results. This transformation being non-linear, its Jacobian is not constant and not simple to evaluate. It is ignored for now.

Our results on two speakers proved that the NM algorithm provides valid solutions for our parameter set β_s for the BPT. The next natural step is to establish results on the Dev01 and Eval01 test sets. For both test sets, we estimated the β_s with

NM n-best and NM 1-best rescoring. The results are summarized in Table 4. For Dev01, a solid gain of 1.32% absolute can

VTLN Method	Hub-5 Dev01	Hub-5 Eval01
Baseline	44.47%	40.05%
NM n-best resc.	43.15%	38.92%
NM 1-best resc.	43.14%	39.04%

Table 4: WERs for Dev01 and Eval01 Testing sets.

be achieved. Both n-best and 1-best rescoring offers virtually the same performance. For Eval01, improvements of 1.13% absolute for n-best and 1.01% absolute for 1-best are achieved. The results on Eval01 are quite encouraging because it is a large test set of 120 speakers. The slight loss for the 1-best rescoring is acceptable in regards to the decrease in computation time.

4. Conclusions

Starting from the mathematical framework of BLT and departing from earlier works, we defined a new frequency transformation whose warping properties are quite meaningful for Speaker Normalization. Two degrees of freedom are available for our transformation allowing complex frequency axis warpings. However, it is still ‘‘constrained’’ by the physical meaning associated with it: a mapping from a prototype BP filter into a general BP filter. By means of Oracle experiments, we found some evidence that the BPT could offer better gains than classic one-parameter methods on a subset of speakers. We proposed a procedure based on the Nelder-Mead algorithm to estimate the two parameters for the BPT in a Maximum Likelihood framework. We established that such procedure does indeed provide valid parameters’ estimates. The NM procedure results on the Dev01 and Eval01 test sets indicate that a gain of 1.32% (Dev01) and 1.13% (Eval01) can be achieved by using BPT. Future works will consist in comparative results for several VTLN methods and BPT on a full test set like Eval01 as well as improving our NM procedure by compensating for the contribution of the transformation’s Jacobian.

5. References

- [1] E. Eide and H. Gish, ‘‘A parametric approach to vocal tract length normalization,’’ in *Proc. of ICASSP*, Atlanta, 1996.
- [2] Alejandro Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Ph.D. thesis, 1990.
- [3] Alan V. Oppenheim and D. H. Johnson, ‘‘Discrete representation of signals,’’ in *Proc. of IEEE*, 1972, vol. 60.
- [4] John W. McDonough, *Speaker Compensation with All-Pass Transforms*, Ph.D. thesis, Johns Hopkins University, 2000.
- [5] A.G. Constantinides, ‘‘Spectral transformations for digital filters,’’ in *Proc. of the IEE*, 1970, vol. 117, pp. 1585–1590.
- [6] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Processing*, Prentice Hall, 1993.
- [7] S. Matsoukas et al., ‘‘The 2001 byblos english lvsr system,’’ in *Proc. of ICASSP*, Orlando, 2002, vol. 1.
- [8] L. Lee and Rose R., ‘‘A frequency warping approach to speaker normalization,’’ *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, January 1998.
- [9] J. Lagarias et al., ‘‘Convergence properties of the nelder-mead simplex method in low dimensions,’’ *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1998.