Vocal Tract Normalization as Linear Transformation of MFCC

Michael Pitz and Hermann Ney

Chair of Computer Science VI (Lehrstuhl für Informatik VI)
RWTH Aachen – University of Technology
52056 Aachen, Germany
{pitz, ney}@informatik.rwth-aachen.de

Abstract

We have shown previously that vocal tract normalization (VTN) results in a linear transformation in the cepstral domain. In this paper we show that Mel-frequency warping can equally well be integrated into the framework of VTN as linear transformation on the cepstrum. We show examples of transformation matrices to obtain VTN warped Mel-frequency cepstral coefficients (VTN-MFCC) as linear transformation of the original MFCC and discuss the effect of Mel-frequency warping on the Jacobian determinant of the transformation matrix. Finally we show that there is a strong interdependence of VTN and Maximum Likelihood Linear Regression (MLLR) for the case of Gaussian emission probabilities.

1. Introduction

Vocal tract normalization (VTN) tries to compensate for the effect of speaker specific vocal tract lengths by warping the frequency axis of the power spectrum of the speech signal [1, 2, 3, 4]. The frequency axis is scaled by a warping function

$$\begin{array}{rcl} g_\alpha : [0,\pi] & \rightarrow & [0,\pi] \\ \omega & \rightarrow & \tilde{\omega} = g_\alpha(\omega) \end{array} \quad (1)$$

and the warped spectrum is defined as

$$|\{X(\omega)\}| = \left|\left\{\tilde{X}(g_\alpha(\omega))\right\}\right|$$

where the warping function g_α is assumed to be invertible, i.e. strictly monotonic and continuous. The frequency $\omega = \pi$ corresponds to the Nyquist frequency and the domain and co-domain are chosen to conserve bandwidth and information contained in the original spectrum.

We have shown in [5, 6] that in the framework of cepstral signal analysis VTN amounts to a linear transformation in the cepstral space for any arbitrary invertible warping function with domain and co-domain as given in Eq. (1). In that work we exemplary derived analytical solutions for the transformation matrices of piece-wise linear, quadratic, and bilinear warping functions.

The warped cepstral coefficients $\tilde{c}_n(\alpha), n = 1 \ldots N$ can be obtained by a linear transformation of the original cepstral coefficients $c_k, k = 1 \ldots K$ with a transformation matrix $\mathbf{A}(\alpha)$ of dimension $N \times K$:

This work was partially funded by the European Commission under the Human Language Technologies project CORETEX (IST-1999-11876), and by the DFG (Deutsche Forschungsgemeinschaft) under contract NE 572/4-1.

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n)\cos(g_\alpha^{(-1)}(\tilde{\omega})k) \quad (2)$$

with

$$s_k = \begin{cases} \frac{1}{2} & : \quad k = 0 \\ 1 & : \quad \text{else} \end{cases}$$

In the case of continuous spectra, there may be no upper limit for N and K. We have assumed that the original spectrum can be represented by a finite number of cepstral coefficients, for instance if it has been cepstrally smoothed already. In practice, however, we work with discrete spectra. Hence, N and K will be finite and equal to the number of spectral lines of the discrete Fourier spectrum. This number can be further reduced for cepstral smoothing.

In the following we will show that VTN warped Mel-frequency cepstral coefficients (VTN-MFCC) can also be obtained by a linear transformation of either the original plain cepstral coefficients or the original MFCC for arbitrary invertible warping functions. We will exemplary discuss transformation matrices obtained for a piece-wise linear warping function. Finally we will discuss a consequence of VTN being a linear transformation of the MFCC, namely a strong interdependence of VTN and Maximum Likelihood Linear Regression (MLLR). This interdependence can explain previous experimental results that improvements obtained by VTN and subsequent MLLR were not additive [7].

2. Integration of Mel Frequency Scale

Mel frequency warping is applied during signal analysis to adjust the spectral resolution to the human ear [8]:

$$f_{mel} = 2595 \cdot \lg\left(1 + \frac{f}{700\text{Hz}}\right) .$$

There are two possible ways to include Mel frequency warping into the framework of VTN as linear transformation:

A.) to express the VTN-MFCC as a linear function of the original, unwarped plain cepstral coefficients (CC)

or

B.) to express the VTN-MFCC as a linear function of the MFCC.

In the following we will calculate the MFCC directly on the power spectrum as described in [9] rather than using a filterbank.

2.1. From Plain CC to VTN-MFCC

We have shown in [5, 6] that a frequency warping of the spectrum with an arbitrary invertible function results in a linear transformation of the cepstral coefficients. Mel frequency warping can be considered as one special case of such a frequency warping and thus results in a linear transformation as well. Therefore the combination of VTN and subsequent Mel warping still amount to a linear transformation in the cepstral domain. VTN is typically applied before Mel scale warping; hence the combination of both warping steps becomes

$$g_{\text{mel}}(g_\alpha(\omega)) \;:\; \omega \rightarrow \tilde{\omega}_{\text{mel}} = B \cdot \lg\left(1 + \frac{g_\alpha(\omega) \cdot f_s}{2\pi \cdot 700\text{Hz}}\right) \quad (3)$$

where $g_\alpha(\omega)$ denotes the VTN warping function as before, f_s denotes the sampling frequency, and B is defined as

$$B = \frac{\pi}{\lg\left(1 + \frac{f_s}{2 \cdot 700\text{Hz}}\right)}$$

to meet the requirement $g_{\text{mel}}(\pi) = \pi$. Inserting Eq. (3) into Eq. (2) leads to

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \cos(\tilde{\omega}_{\text{mel}} n) \cos\left(g_\alpha^{(-1)}\left(g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}})\right)\, k\right) \quad (4)$$

Thus we can express the cepstral coefficients of the VTN-Mel-warped spectrum as linear transformation of the original, unwarped cepstral coefficients.

2.2. From MFCC to VTN-MFCC

We will see in Section 4 that VTN is equivalent to a parameterized constrained MLLR transformation. MLLR is a linear transformation of model parameters (means and variances) which were typically been estimated from MFCC feature vectors. Thus more interesting and of practical relevance is to express the VTN-Mel-warped cepstral coefficients as a function of the MFCC (i.e. without VTN) instead of the plain cepstral coefficients. The difficulty in the present case is that VTN is typically applied *before* Mel warping. We start with the definition of the VTN-Mel-warped cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$

$$\tilde{c}_n^{\text{mel}}(\alpha) = \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \ln\left|\hat{X}(\tilde{\omega}_{\text{mel}})\right| \cos(\tilde{\omega}_{\text{mel}} n) \;. \quad (5)$$

VTN is usually applied to original, i.e. non-Mel-scaled, spectrum ($\tilde{\omega}_{\text{mel}}$ denotes the VTN-Mel-warped frequency)

$$\tilde{\omega}_{\text{mel}} = g_{\text{mel}} \circ g_\alpha(\omega)$$

and the warped spectrum is given as

$$\left|\left\{\hat{X}(\tilde{\omega}_{\text{mel}})\right\}\right| = \left|\left\{X\left(g_\alpha^{(-1)}\left(g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}})\right)\right)\right\}\right| = \left|\left\{X(\omega)\right\}\right| \;.$$

We now expand the spectrum as function of the Mel-warped frequency ω_{mel} in terms of unnormalized (i.e. not VTN-warped) cepstral coefficients c_k^{mel}

$$\ln|X(\omega)|^2 = \ln\left|\hat{X}(\omega_{\text{mel}})\right|^2 = 2\sum_{k=0}^{K} c_k^{\text{mel}} \cos(\omega_{\text{mel}} k) \;. \quad (6)$$

As before, inserting Eq. (6) into Eq. (4) results in

$$\tilde{c}_n^{\text{mel}}(\alpha) = \sum_{k=0}^{K} c_k^{\text{mel}} \; \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \; \cos(\omega_{\text{mel}} k) \cdot \cos(\tilde{\omega}_{\text{mel}} n)$$

We now need to express the unnormalized Mel-scale frequency ω_{mel} as function of the VTN-warped Mel-scale frequency $\tilde{\omega}_{\text{mel}}$:

$$\omega_{\text{mel}} = g_{\text{mel}}(\omega) = g_{\text{mel}} \circ g_\alpha^{(-1)} \circ g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}}) \;.$$

Finally, we obtain

$$\tilde{c}_n^{\text{mel}}(\alpha) = \sum_{k=0}^{K} A_{nk}^{\text{mel}}(\alpha) \; c_k^{\text{mel}}$$

with

$$A_{nk}^{\text{mel}}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\text{mel}} \; \cos(\tilde{\omega}_{\text{mel}} n) \cos(g_{\text{mel}} \circ g_\alpha^{(-1)} \circ g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}}) \; k) \;. \quad (7)$$

Hence, the cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$ of the VTN-warped Mel-scale spectrum can be computed by a linear transformation of the unnormalized cepstral coefficients c_k^{mel} (without VTN warping). Because of the non-linear transformation the integral in Eq. (7) may hardly be solved analytically. Nevertheless, the transformation matrix can be calculated numerically.

We have calculated the transformation matrix numerically for a piece-wise linear warping function (dashed line in Fig. 1)

$$\omega \rightarrow \tilde{\omega} = g_\alpha(\omega) = \begin{cases} \alpha\omega & : \;\; \omega \leq \omega_0 \\ \alpha\omega_0 + \dfrac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & : \;\; \omega > \omega_0 \end{cases} \quad (8)$$

We choose the inflexion point ω_0, where the slope of the warping function changes, as follows:

$$\omega_0 \;\; = \;\; \begin{cases} \frac{7}{8}\pi & \alpha \leq 1 \\ \frac{7}{8 \cdot \alpha}\pi & \alpha > 1 \end{cases}$$

The resulting warping function $g_{\text{eff}} := g_{\text{mel}} \circ g_\alpha^{(-1)} \circ g_{\text{mel}}^{(-1)}$ reads

$$g_{\text{eff}}(\tilde{\omega}_{\text{mel}}) := g_{\text{mel}} \circ g_\alpha^{(-1)} \circ g_{\text{mel}}^{(-1)}(\tilde{\omega}_{\text{mel}}) = \begin{cases} B \cdot \log\left[1 + \frac{1}{\alpha}\left(10^{\tilde{\omega}_{\text{mel}}/B} - 1\right)\right] & : \;\; \omega_{\text{mel}} \leq g_{\text{mel}}(\omega_0) \\ B \cdot \log\left[1 + \frac{f_s \tilde{\omega}_0}{2 \cdot 700\text{Hz}}\left(\frac{1}{\alpha} - \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}\right) + \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}\left(10^{\tilde{\omega}_{\text{mel}}/B} - 1\right)\right] & : \;\; \omega_{\text{mel}} > g_{\text{mel}}(\omega_0) \end{cases} \quad (9)$$

and is shown in Fig. 1 (straight line). If we expand the effective warping function $g_{\text{eff}}(\tilde{\omega}_{\text{mel}})$ for $\omega_{\text{mel}} \leq g_{\text{mel}}(\omega_0)$ in a Taylor series about $\alpha = 1$ (for $\omega_{\text{mel}} \geq g_{\text{mel}}(\omega_0)$ the expansion is similar)

$$g_{\text{eff}}(\tilde{\omega}_{\text{mel}}) = \tilde{\omega}_{\text{mel}} - B\frac{1 - 10^{-\tilde{\omega}_{\text{mel}}/B}}{\ln(10)}(\alpha - 1) + \mathcal{O}\left((\alpha - 1)^2\right)$$

we see that the linear term dominates the expansion because the term $\frac{1 - 10^{-\tilde{\omega}_{\text{mel}}/B}}{\ln(10)}$ is small for $0 \leq \tilde{\omega}_{\text{mel}} \leq \pi$. Thus, $g_{\text{eff}}(\tilde{\omega}_{\text{mel}})$

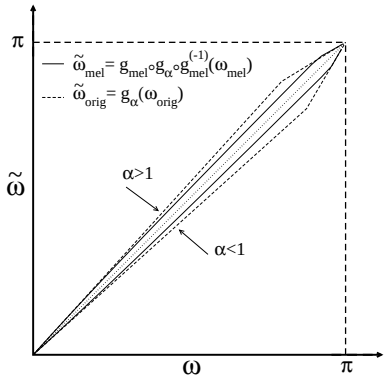

Figure 1: Effective warping function $g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ as function of the Mel frequency ω_{mel} (straight) in comparison to the warping function g_α for plain CC as function of the original frequency ω_{orig} (dashed) for $\alpha = 0.9$ and $\alpha = 1.1$

can be approximated by a linear function with an appropriate choice of an effective warping factor α_{eff}.

The cepstral coefficients $\tilde{c}_n^{\text{mel}}(\alpha)$ obtained by a linear transformation of MFCC with the matrix defined by Eq. (7) are identical to those calculated by explicitly warping the spectrum during signal analysis as presented in [9].

Transformation matrices for MFCC using piece-wise linear warping with warping factors ($\alpha = 0.9$ and $\alpha = 1.1$) are shown in Fig. 2 and 3. These matrices were calculated by solving Eq. (7) numerically without approximations.

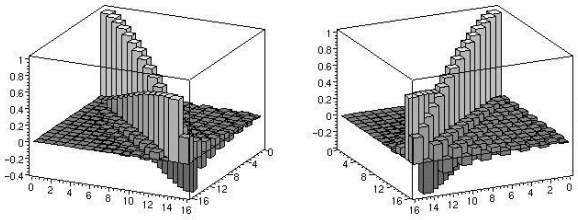

Figure 2: Matrix for piecewise linear warping function, $\alpha = 0.9$, Mel scale

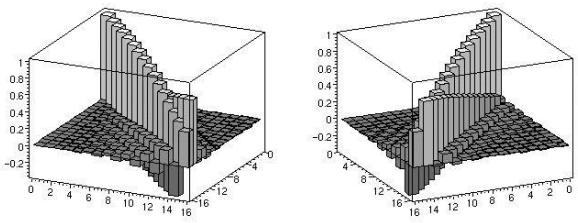

Figure 3: Matrix for piecewise linear warping function, $\alpha = 1.1$, Mel scale

In order to study the effect of the Mel-frequency warping on the transformation matrix, we will compare these matrices with those calculated analytically for computing VTN-CC as linear transformation of the plain CC (Fig. 4 and 5). We observe that the matrices for the Mel scale are more diagonally dominant than those for the original scale. Comparing the resulting warping function $\tilde{\omega}_{\text{mel}} = g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}(\omega_{\text{mel}})$ (straight line in Fig. 1) as function of the Mel frequency with the warping function g_α for plain CC (dashed line in Fig. 1) as function of the original frequency, we see that $g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ is much closer to identity than g_α. Therefore the transformation matrix for MFCC is more diagonally dominant than those obtained for plain cepstral coefficients. The general structure of the warping matrices as well as possible approximations are discussed in more detail in [6], also for bilinear and quadratic warping functions.

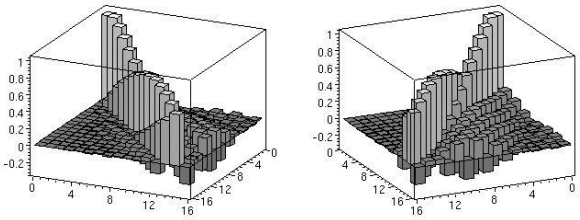

Figure 4: Matrix for piecewise linear warping function, $\alpha = 0.9$, no Mel scale

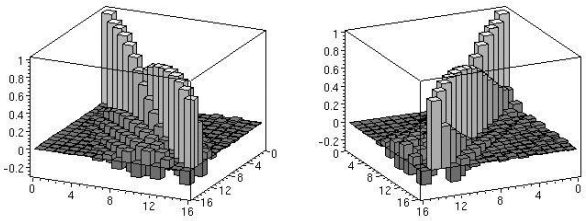

Figure 5: Matrix for piecewise linear warping function, $\alpha = 1.1$, no Mel scale

3. Discussion of Jacobian Determinant

To estimate the unknown warping factor α, we proceed as follows: For each speaker r we are given labeled training data (X_r, Wr) where X_r denote the sequence of acoustic data and W_r the sequence of spoken words. In recognition, a preliminary hypothesis of the unknown word sequence W_r can be obtained by a first recognition pass. Typically, we apply a maximum likelihood estimation of α

$$\hat{\alpha}_r = \underset{\alpha}{\operatorname{argmax}}\, p(X_r|W_r;\theta,\alpha) \tag{10}$$

$$= \underset{\alpha}{\operatorname{argmax}} \left\{ p_0(f_\alpha(X_r)|W_r;\theta_0) \cdot \left| \frac{df_\alpha(X_r)}{dX_r} \right| \right\} \tag{11}$$

In VTN the speaker normalization is usually not performed as a transformation of the acoustic vectors but by warping the power spectrum during signal analysis instead. Hence, the Jacobian determinant can hardly be calculated. In virtually all experimental studies the second term in Eq. (11), the Jacobian determinant, is neglected. Whether this is a good approximation or not will depend very much on how much the Jacobian determinant depends on α. Therefore it is good to study the second term as function of α. Expressing VTN as a matrix transformation of the acoustic vector ($x \rightarrow \mathbf{A}\, x$) now enables us study the Jacobian determinant $|\det \mathbf{A}|$ of the transformation.

A plot showing the dependency of the Jacobian determinant on the warping factor α has been computed numerically for piece-wise linear warping (Fig. 6). The dependence of the

Jacobian determinant on α is weaker for MFCC (dashed line in Fig. 6) because the effective warping function $g_{\text{mel}} \circ g_\alpha \circ g_{\text{mel}}^{(-1)}$ is much closer to identity than g_α for CC (cf. Fig. 1)

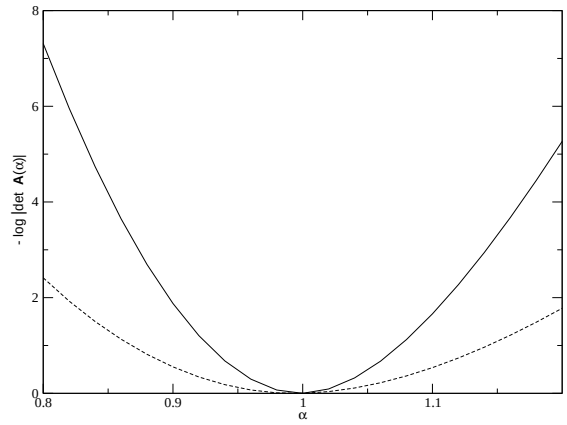

Figure 6: Plot of $-\log|\det \mathbf{A}|$ for piece-wise linear warping of 16 cepstral coefficients as function of α. Straight line: original frequency scale, dashed line: Mel frequency scale

4. Interdependence of VTN and MLLR

Most of today's automatic speech recognition systems make use of Hidden Markov Models (HMM) with Gaussian emission probability distributions

$$\mathcal{N}(x|\mu,\mathbf{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}} \exp\left\{-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\} .$$

with state dependent parameters μ and Σ. If the acoustic feature vector (essentially the cepstral coefficients) x is normalized with the VTN matrix $\mathbf{A}$, the Gaussian distribution changes to

$$\begin{aligned} x \rightarrow & y = \mathbf{A}\,x : \\ \mathcal{N}(x|\mu,\mathbf{\Sigma}) \rightarrow & \mathcal{N}(y|\mu,\mathbf{\Sigma}) \\ & = \mathcal{N}(x|\mathbf{A}^{-1}\,\mu, \mathbf{A}^{-1}\,\mathbf{\Sigma}\,\mathbf{A}^{-1^T}) \\ & = \mathcal{N}(x|\hat{\mu},\hat{\mathbf{\Sigma}}) \end{aligned}$$

with

$$\hat{\mu} = \mathbf{A}^{-1}\,\mu \qquad \text{and} \qquad \hat{\mathbf{\Sigma}} = \mathbf{A}^{-1}\,\mathbf{\Sigma}\,\mathbf{A}^{-\mathbf{1}\mathbf{T}} . \qquad (12)$$

Thus, a linear transformation of the observation vector x is equivalent to a linear transformation of the mean vector μ and an appropriate transformation of the covariance matrix $\mathbf{\Sigma}$.

The transformations in Eq. (12) describe a special case of MLLR which is called *constrained* MLLR [10, 11] (constrained refers to the use of the same matrix $\mathbf{A}$ for the transformation of the mean and variance).

In [7], Uebel and Woodland have found experimentally that improvements obtained by constrained MLLR and VTN were not additive. As we have shown, VTN may be viewed as a special case of constrained MLLR adaptation with an restriction to only one adjustable parameter (the warping parameter) which determines the matrix elements. The experiments were based on a MF-PLP signal analysis. The difference between MFCC and MF-PLP is mainly caused by different types of smoothing. This difference is not expected to effect the equivalence of VTN and linear transformations. Hence, the experiments support the analytic result that VTN is a special case of constrained MLLR.

5. Conclusion

We have shown in this work that VTN warped Mel-frequency cepstral coefficients (VTN-MFCC) can also be obtained by a linear transformation of either the original plain cepstral coefficients or the original MFCC for arbitrary invertible warping functions. The numerical values of VTN-MFCC computed with the presented approach were identical to those obtained by explicitly warping the spectrum during signal analysis. Expressing VTN as matrix transformation of MFCC allows us to compute the Jacobian determinant of the transformation, which has typically been neglected so far. Finally, we have shown that for the case of Gaussian emission probabilities VTN and MLLR are strongly interdependent, which can explain previous experimental results that improvements obtained by VTN and subsequent constrained MLLR were not additive.

6. References

[1] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 346–349.

[2] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 353–356.

[3] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, No. 2, Apr. 1977, pp. 183–192.

[4] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 339–341.

[5] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," in *Proc. ISCA Europ. Conf. on Speech Communication and Technology*, vol. 4, Aalborg, Denmark, Sept. 2001, pp. 2653 – 2656.

[6] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *submitted to IEEE Trans. on Speech and Audio Processing*, 2003.

[7] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalisation," in *Proc. ISCA Europ. Conf. on Speech Communication and Technology*, vol. 6, Budapest, Hungary, Sept. 1999, pp. 2527–2530.

[8] S. J. Young, *HTK: Hidden Markov Model Toolkit V1.4. User Manual*, Cambridge, England, Feb. 1993.

[9] S. Molau, M. Pitz, R. Schlüter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, June 2001, pp. 73–76.

[10] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, Sept. 1995.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.