Vocal Tract Normalization Equals Linear Transformation in Cepstral Space

Michael Pitz and Hermann Ney, *Member, IEEE*

***Abstract*—Vocal tract normalization (VTN) is a widely used speaker normalization technique which reduces the effect of different lengths of the human vocal tract and results in an improved recognition accuracy of automatic speech recognition systems.**

We show that VTN results in a linear transformation in the cepstral domain, which so far have been considered as independent approaches of speaker normalization. We are now able to compute the Jacobian determinant of the transformation matrix, which allows the normalization of the probability distributions used in speaker-normalization for automatic speech recognition. We show that VTN can be viewed as a special case of Maximum Likelihood Linear Regression (MLLR). Consequently, we can explain previous experimental results that improvements obtained by VTN and subsequent MLLR are not additive in some cases. For three typical warping functions the transformation matrix is calculated analytically and we show that the matrices are diagonal dominant and thus can be approximated by quindiagonal matrices.

***Index Terms*—Linear transformation, speaker adaptive modeling and training, speaker adaptive recognition, speech recognition, vocal tract (length) normalization.**

I. Introduction

A SPEAKER independent speech recognition system has to cope with a lot of variability in the acoustic signal. For example, varying transmission channels, noise, speakers, and speaking styles are sources of such irrelevant variabilities. From a more general perspective, this can be viewed as mismatch between training and testing condition of the ASR system. A lot of normalization (i.e., transformation of acoustic features) and adaptation (i.e., transformation of acoustic model parameters) schemes have been developed in the past few years to compensate for this mismatch in order to improve the accuracy of the ASR system [1].

A major part of the variability in the speech signal is caused by the speaker dependent vocal tract length. 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 [2]–[5]. In a simple physiological model, the human vocal tract is treated as a uniform tube of length L. According to this model a change in L by a certain factor α^{-1} results in a scaling of the frequency axis by α. Thus, for this model, the frequency axis should be scaled linearly to compensate for the variability caused by different vocal tracts of individual speakers. However, the simple tube model may not be the best for the human vocal tract so other frequency warping functions have been investigated [2]. In general, the frequency axis is scaled by a warping function g_α with a transformation parameter α

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

where ω denotes the original frequency and $\tilde{\omega}$ the warped frequency. The warping function g_α is assumed to be invertible, i.e., strictly monotonic and continuous (see Fig. 1). 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.

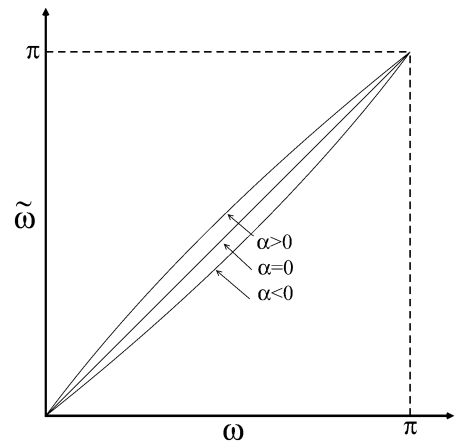

Fig. 1. Example of a VTN warping function $\bar{\omega} = g_\alpha(\omega)$ for different values of α.

All typical VTN approaches have in common that the warping function depends only on a few free parameters which control the amount of the frequency distortion. Even with only one free parameter (the warping factor α) to model the mismatch, VTN performs very well in a variety of recognition tasks. On the other hand, a few parameters allow to be estimated reliably on very little data which makes VTN a good choice for on-line speech recognition systems.

The relationship between VTN and linear transformations in the cepstral domain has been studied before. However, these investigations were restricted to a bilinear warping function [6, p. 119], [7, p. 113], conformal mappings [8], or were based on

Manuscript received May 5, 2003; revised March 10, 2004. This work was supported in part by the European Commission under the Human Language Technologies Project CORETEX (IST-1999-11876) and by the DFG (Deutsche Forschungsgemeinschaft) under Contracts NE 572/4-1 and NE 572/4-3. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng.

M. Pitz was with the Lehrstuhl für Informatik VI, RWTH Aachen University, 52056 Aachen, Germany. He is now with BMW Group Research and Technology, Hanauer Straße 46, 80992 München, Germany (e-mail: michael.pitz@bmw.de).

H. Ney is with the Lehrstuhl für Informatik VI, RWTH Aachen University, 52056 Aachen, Germany (e-mail: ney@informatik.rwth-aachen.de).

Digital Object Identifier 10.1109/TSA.2005.848881

1063-6676/$20.00 © 2005 IEEE

plausibility arguments and rough approximations [9]. In [10] it was shown that improvements in recognition accuracy gained by VTN and subsequent linear transformation (constrained MLLR) were not additive which indicates that both techniques may not be independent of each other.

We will show in this work that there is a strong interdependence of VTN and a linear transformation of the cepstral vector for arbitrary invertible warping functions without approximations. A related result has also been reported in [11] in the context of spectral distortion measures. For the case of Gaussian emission probabilities we will show the equivalence of two widely used normalization approaches (VTN and MLLR) which have so far been considered to be independent.

The remainder of the paper is organized as follows: We will show in Section II that VTN can always be expressed as linear transformation of the cepstral vector for *arbitrary* invertible warping functions. Than, in Section III, we will derive analytical results for the transformation matrix for three relevant warping functions. The general structure of the transformation matrix will be discussed in Section IV based on some weak assumptions on the warping function. In Section V we will show that VTN can also be expressed as linear transformation when using the Mel frequency scale. Some examples of spectra obtained by the method presented in this work will be shown in Section VI as well as investigations of the influence of cepstral smoothing at different stages during the signal analysis. We will discuss the correlation between VTN and MLLR as a consequence of VTN being a linear transformation in Section VII. In Section VIII we will analyze the Jacobian determinant of the transformation, which can now be calculated for the first time for any given warping function. In the conclusion in Section IX we will finally show some possible applications of the presented results and first experimental evaluatuions are given in Section X.

II. VTN Equals Linear Transformation in Cepstral Space

Most of today's ASR systems use some sort of either Mel-frequency cepstral coefficients (MFCCs) [12] or perceptual linear predictive (PLP) [13] coding signal analysis front end. In the following, we will show 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 (1).

Fig. 2 shows a traditional MFCC signal analysis front end (left side). After some preprocessing like preemphasis and windowing, the Fourier power spectrum of the speech waveform is computed for each time frame. The frequency axis of this power spectrum is warped by a warping function, usually piece-wise linear or bilinear, with a speaker dependent warping factor. Afterwards a filterbank is applied and the logarithm is taken. In a last step a discrete cosine transformation is applied to the log filter bank coefficients to remove the correlation between the different outputs. The dimensionality of the cepstral vector is reduced by omitting the highest cepstral coefficients for smoothing.

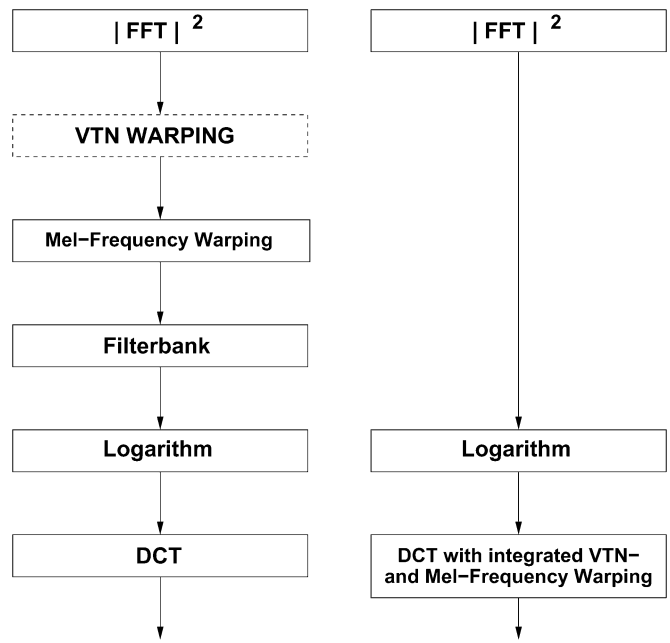

Fig. 2. Schemes of traditional MFCC computation (left) and integrated approach (right).

In [14] we proposed an improved signal analysis in which Mel-frequency warping and VTN are integrated into the discrete cosine transform (Fig. 2, right side). This new approach is mathematically simpler as the traditional one but gives very similar results in terms of recognition performance as well as the numerical values of the cepstral coefficients. In the remaining we will make use of this integrated MFCC computation scheme.

The cepstral coefficients c_k, $k = 0, \ldots, K$ of a spectrum $X(\omega)$ are defined by:

$$\begin{aligned} c_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega \mathrm{e}^{i\omega k} \ln |X(\omega)|^2, \quad k = 0, \ldots, K \\ &= \frac{s_k}{\pi}\int_{0}^{\pi} d\omega \cos(\omega k) \ln |X(\omega)|^2 \end{aligned} \tag{2}$$

where ω may either denote the true physical or the Mel frequency scale. We have introduced the symmetry factor s_k for convenience

$$s_k = \begin{cases} \frac{1}{2} & : k = 0 \\ 1 & : \text{else.} \end{cases} \tag{3}$$

We will start with plain cepstral coefficients (CC) and consider Mel-scale warping later because with plain CC the occurring integrals and thus the transformation matrix may be computed analytically. We will show in Section V that the results of the current Section are still valid for MFCC.

The transformation of spectral warping is defined as

$$\begin{aligned} \omega \rightarrow \tilde{\omega} &= g_\alpha(\omega) \\ |\{X(\omega)\}| &= \left|\left\{\tilde{X}(g_\alpha(\omega))\right\}\right| \\ &= \left|\left\{\tilde{X}(\tilde{\omega})\right\}\right|. \end{aligned}$$

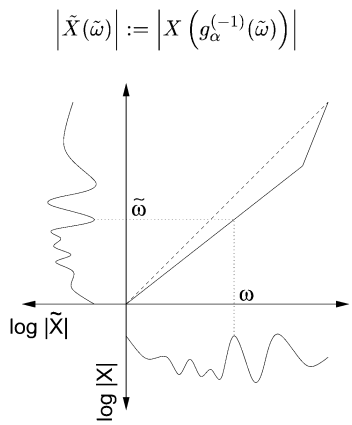

Fig. 3. Principle of vocal tract normalization (VTN).

That means, for the new frequency $\tilde{\omega}$ we compute the new spectrum $\tilde{X}(\tilde{\omega})$ according to (see also Fig. 3):

$$\left|\tilde{X}(\tilde{\omega})\right| := \left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|.$$

Hence, the n-th cepstral coefficient $\tilde{c}_n(\alpha)$, $n = 1, \ldots, N$ of the *warped* spectrum is given by

$$\tilde{c}_n(\alpha) = \frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\ \cos(\tilde{\omega}n)\ln\left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2. \quad (4)$$

We expand the spectrum $\ln\left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2$ in a Fourier series

$$\begin{aligned}\ln\left|X\left(g_\alpha^{(-1)}(\tilde{\omega})\right)\right|^2 &= \ln|X(\omega)|^2\\ &= 2\sum_{k=0}^{K} c_k\cos(\omega k)\\ &= 2\sum_{k=0}^{K} c_k\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right) \quad (5)\end{aligned}$$

where c_k denotes the k-th cepstral coefficient of the *unwarped* spectrum. 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. This assumption is not mandatory as long as the Fourier series is uniformly convergent, which will be the case for the spectra we deal with in practical applications. 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

We now insert (5) into (4), interchange integration and summation, and obtain

$$\begin{aligned}\tilde{c}_n(\alpha) &= \frac{2s_k}{\pi}\int_0^\pi d\tilde{\omega}\cos(\tilde{\omega}n)\sum_{k=0}^{K} c_k\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right)\\ &= \sum_{k=0}^{K} c_k\frac{2s_k}{\pi}\int_0^\pi d\tilde{\omega}\cos(\tilde{\omega}n)\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right)\\ &= \sum_{k=0}^{K} A_{nk}(\alpha)c_k \quad (6)\end{aligned}$$

with

$$A_{nk}(\alpha) = \frac{2s_k}{\pi}\int_0^\pi d\tilde{\omega}\cos(\tilde{\omega}n)\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right) \quad (7)$$

with the symmetry factor from (3)

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

Thus, the vector of warped cepstral coefficients $\tilde{c}_n(\alpha)$ is a linear transformation of the original cepstral coefficients c_k with a transformation matrix $\mathbf{A}(\alpha)$ of dimension $N \times K$.

In other words, we have expressed VTN warping with a warping function g_α as a linear transformation of the cepstral coefficients of the unwarped spectrum with transformation matrix $\mathbf{A}(\alpha)$ as given in (7). Note that the only assumption we have made regarding the warping function is invertibility.

For cepstral smoothing the number of cepstral coefficients may be reduced by omitting higher coefficients. In order to prevent a loss in spectral resolution the dimensionality reduction has to be applied to the *warped* cepstral coefficients. This will be discussed in more detail in Section VI.

III. Analytic Calculation of the Transformation Matrix

As we have seen in the last Section, VTN can always be expressed as a linear transformation in the cepstral domain independent of the functional form of the (invertible) warping function (cf. (6) and (7)). In the following we will analytically calculate the transformation matrix defined in (7) for 3 typical warping functions: piece-wise linear, quadratic and bilinear warping. The analytic calculation of the transformation matrix for other warping functions than those presented below, however, may not be as straightforward. Although the solutions for $\mathbf{A}(\alpha)$ will look quite different, we will see in Section IV that the resulting matrices have a common shape. But first we will present the detailed calculations.

A. Piece-Wise Linear Warping Function

In order to apply a piece-wise linear warping, we first compute the solution for a strictly linear warping function

$$\begin{aligned}g_\alpha &: \omega \to \tilde{\omega} = \alpha\cdot\omega\\ g_\alpha^{(-1)} &: \tilde{\omega} \to \omega = \alpha^{-1}\cdot\tilde{\omega}.\end{aligned}$$

This warping function does not meet the requirements of (1) as the co-domain is not equal to $[0, \pi]$ for $\alpha \neq 1$. We will neglect this requirement at first because the calculation for the piece-wise linear warping function (which meets the requirements of (1)) turns out to be very similar.

The entries $A_{nk}(\alpha)$ of the transformation matrix can be computed by elementary integration. For $\alpha \neq 1$ we obtain

$$\begin{aligned}A_{nk}(\alpha) &= \frac{2s_k}{\pi}\int_0^\pi d\tilde{\omega}\cos(\tilde{\omega}n)\cos(\alpha^{-1}\tilde{\omega}k)\\ &= \frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\big(\cos(\tilde{\omega}n + \alpha^{-1}\tilde{\omega}k)\\ &\qquad + \cos(\tilde{\omega}n - \alpha^{-1}\tilde{\omega}k)\big)\\ &= s_k\frac{\sin[(n+\alpha^{-1}k)\pi]}{(n+\alpha^{-1}k)\pi} + s_k\frac{\sin[(n-\alpha^{-1}k)\pi]}{(n-\alpha^{-1}k)\pi}.\end{aligned}$$

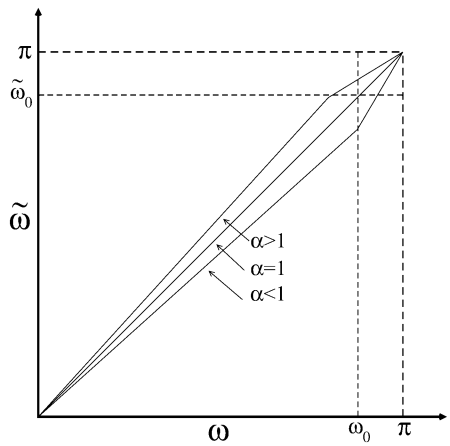

Fig. 4. Piece-wise linear warping functions for $\alpha = 0.9, 1.0, 1.1$.

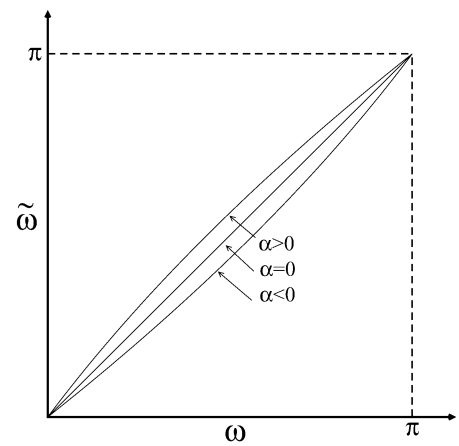

Fig. 5. Example of quadratic VTN warping functions $\bar{\omega} = g_\alpha(\omega)$ for $\alpha = -0.8, 0, +0.8$.

For $\alpha = 1$ this simplifies to

$$A_{nk}(1) = \delta_{nk}$$

because of the orthonormality of the cosine function.

To meet the requirement of invertibility, we now consider a piece-wise linear warping function [5], [15] as shown in Fig. 4:

$$\omega \to \tilde{\omega} = g_\alpha(\omega) = \begin{cases} \alpha\omega & : \omega \le \omega_0 \\ \alpha\omega_0 + \frac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & : \omega > \omega_0. \end{cases} \tag{8}$$

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

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & \alpha \le 1 \\ \frac{7}{8 \cdot \alpha}\pi & \alpha > 1\,. \end{cases}$$

The transformation matrix $A_{nk}(\alpha)$ is computed similarly to the linear case but the integration is split into two parts

$$A_{nk}(\alpha, \tilde{\omega}_0) = \frac{2s_k}{\pi}\left(\int_0^{\tilde{\omega}_0} + \int_{\tilde{\omega}_0}^{\pi}\right) d\tilde{\omega} \times \cos(\tilde{\omega}n)\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right)$$

with $\tilde{\omega}_0 = \alpha \cdot \omega_0$.

Noting that the solution for $\alpha = 1$ remains the same as in the linear case, we obtain for $\alpha \neq 1$

$$\begin{aligned} A_{nk}(\alpha) = {} & s_k \frac{\sin[(n - \alpha^{-1}k)\tilde{\omega}_0]}{(n - \alpha^{-1}k)\pi} \\ & + s_k \frac{\sin[(n + \alpha^{-1}k)\tilde{\omega}_0]}{(n + \alpha^{-1}k)\pi} \\ & - s_k \frac{\sin[(n - \alpha^{-1}k)\tilde{\omega}_0]}{\left(n - \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}k\right)\pi} \\ & - s_k \frac{\sin[(n + \alpha^{-1}k)\tilde{\omega}_0]}{\left(n + \frac{\pi - \alpha^{-1}\tilde{\omega}_0}{\pi - \tilde{\omega}_0}k\right)\pi}. \end{aligned} \tag{9}$$

This matrix can now be used for VTN alternatively to the conventional approaches like explicitly warping the power spectrum during signal analysis or the integrated approach described in [14]. A detailed discussion of this warping matrix will be given in Section IV.

B. Quadratic Warping Function

In order to study the effect of the warping function on the transformation matrix, we will calculate the matrix for a quadratic warping function (Fig. 5). Although it has not been studied in literature in such detail like piecewise or bilinear warping functions, a quadratic warping function is the next step beyond linear warping in a sense of a power series expansion.

We define the quadratic warping function as follows:

$$g_\alpha^{(-1)} : \tilde{\omega} \to \omega = \tilde{\omega} + \alpha\left(\frac{\tilde{\omega}}{\pi} - \left(\frac{\tilde{\omega}}{\pi}\right)^2\right). \tag{10}$$

We start again with (7)

$$\begin{aligned} A_{nk}(\alpha) &= \frac{2s_k}{\pi}\int_0^\pi d\tilde{\omega}\cos(\tilde{\omega}n)\cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right) \\ &= \frac{s_k}{\pi}\int_0^\pi d\tilde{\omega}\left[\cos\left(\tilde{\omega}n + g_\alpha^{(-1)}(\tilde{\omega})k\right) \right. \\ &\qquad \left. + \cos\left(\tilde{\omega}n - g_\alpha^{(-1)}(\tilde{\omega})k\right)\right] \\ &= \frac{1}{\pi}\left(I_+ + I_-\right) \end{aligned} \tag{11}$$

with

$$\begin{aligned} I_\pm &= \int_0^\pi d\tilde{\omega}\cos\left\{\tilde{\omega}n \pm \left[\tilde{\omega} + \alpha\left(\frac{\tilde{\omega}}{\pi} - \left(\frac{\tilde{\omega}}{\pi}\right)^2\right)\right]k\right\} \\ &= \int_0^\pi d\tilde{\omega}\cos\left\{\left(n \pm k\left(1 + \frac{\alpha}{\pi}\right)\right)\tilde{\omega} \mp k\frac{\alpha}{\pi^2}\tilde{\omega}^2\right\}. \end{aligned} \tag{12}$$

The integrand of (12) has the form $\cos(ax^2 + bx)$ with $a = \mp k\alpha/\pi^2$ and $b = (n \pm k(1 + \alpha/\pi))$. To calculate $I_\pm$ we make use of the identity

$$\begin{aligned} \cos(ax^2 + bx) = {} & \cos\left(a\left(x + \frac{b}{2a}\right)^2\right)\cos\left(\frac{b^2}{4a}\right) \\ & + \sin\left(a\left(x + \frac{b}{2a}\right)^2\right)\sin\left(\frac{b^2}{4a}\right). \end{aligned} \tag{13}$$

The resulting integrals from (13) can now be solved elementary with the solutions

$$\int dx \sin\left(a\left(x+\frac{b}{2a}\right)^2\right) = \sqrt{\frac{\pi}{2a}}\mathrm{S}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right)$$
$$\int dx \cos\left(a\left(x+\frac{b}{2a}\right)^2\right) = \sqrt{\frac{\pi}{2a}}\mathrm{C}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right)$$

where $\mathrm{S}(x)$ and $\mathrm{C}(x)$ denote the Fresnel sine and cosine which are defined as

$$\mathrm{S}(x) = \int_0^x dt \sin\left(\frac{\pi}{2}t^2\right)$$
$$\mathrm{C}(x) = \int_0^x dt \cos\left(\frac{\pi}{2}t^2\right)$$

respectively. Thus, integration of (13) yields

$$\int dx \cos(ax^2+bx) = \sqrt{\frac{\pi}{2a}}\left[\cos\left(\frac{b^2}{4a}\right)\mathrm{C}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right) + \sin\left(\frac{b^2}{4a}\right)\mathrm{S}\left(\frac{2ax+b}{\sqrt{2\pi a}}\right)\right].$$

Finally, the solution of (11) for the quadratic warping function as defined in (10) becomes

$$\begin{aligned} A_{nk}(\alpha) = {} & \frac{\pi s_k}{\sqrt{2\pi k\alpha}}\cos\left(\frac{(n\pi+k(\pi+\alpha))^2}{4k\alpha}\right) \\ & \times\left[\mathrm{C}\left(\frac{n\pi+k(\pi+\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{C}\left(\frac{n\pi+k(\pi-\alpha)}{\sqrt{2\pi ka}}\right)\right] \\ & + \frac{\pi s_k}{\sqrt{2\pi k\alpha}}\cos\left(\frac{(n\pi-k(\pi+\alpha))^2}{4k\alpha}\right) \\ & \times\left[\mathrm{C}\left(\frac{n\pi-k(\pi-\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{C}\left(\frac{n\pi-k(\pi+\alpha)}{\sqrt{2\pi ka}}\right)\right] \\ & + \frac{\pi s_k}{\sqrt{2\pi k\alpha}}\sin\left(\frac{(n\pi+k(\pi+\alpha))^2}{4k\alpha}\right) \\ & \times\left[\mathrm{S}\left(\frac{n\pi+k(\pi+\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{S}\left(\frac{n\pi+k(\pi-\alpha)}{\sqrt{2\pi ka}}\right)\right] \\ & + \frac{\pi s_k}{\sqrt{2\pi k\alpha}}\sin\left(\frac{(n\pi-k(\pi+\alpha))^2}{4k\alpha}\right) \\ & \times\left[\mathrm{S}\left(\frac{n\pi-k(\pi-\alpha)}{\sqrt{2\pi ka}}\right) - \mathrm{S}\left(\frac{n\pi-k(\pi+\alpha)}{\sqrt{2\pi ka}}\right)\right]. \end{aligned} \tag{14}$$

This matrix will also be discussed in Section IV.

C. Bilinear Warping Function

Another frequently used warping function is the bilinear transformation (BLT). In order to calculate the transformation matrix for a BLT we will switch to the complex z-domain instead of the real frequency ω because the occurring integrals will then be easier to compute. The BLT for a complex number z is defined by

$$\tilde{z} = \frac{z+\alpha}{1+\alpha z} \text{ with } z = \mathrm{e}^{i\omega}, \quad \tilde{z} = \mathrm{e}^{i\tilde{\omega}}$$

with the (real) warping parameter α, $|\alpha| < 1$.

As the BLT has a simple form for complex numbers we use the (complex) z-transform, which is equivalent to the Fourier transform for $z = \mathrm{e}^{i\omega}$

$$\log[X(z)] = \sum_{k=0}^{\infty} c_k z^{-k} \leftrightarrow c_k = \frac{1}{2\pi i}\oint dz \log[X(z)] z^{k-1}.$$

The z-transform of warped spectrum is given by

$$\log[\hat{X}(\hat{z})] = \sum_{n=0}^{\infty} \hat{c}_n \hat{z}^{-n} \leftrightarrow \hat{c}_n = \frac{1}{2\pi i}\oint d\hat{z} \log[X(\hat{z})] \hat{z}^{n-1}.$$

Frequency warping is defined as

$$\hat{X}(\mathrm{e}^{ig_\alpha(\omega)}) = X(\mathrm{e}^{i\omega}) \text{ or } \hat{X}(\hat{z}) = X(z)$$

and thus

$$\hat{c}_n = \frac{1}{2\pi i}\oint dz \sum_{k=0}^{\infty} c_k \hat{z}^{-k} z^{n-1}.$$

Interchanging integration and summation yields

$$\hat{c}_n = \sum_{k=0}^{\infty} \frac{1}{2\pi i}\oint dz \hat{z}^{-k} z^{n-1} \cdot c_k.$$

Again, we see that warping of the frequency axis of a given spectrum amounts to a linear transformation in the cepstral domain

$$\hat{c}_n(\alpha) = \sum_{k=0}^{\infty} A_{nk}(\alpha)\, c_k$$

with the transformation matrix

$$A_{nk}(\alpha) = \frac{1}{2\pi i}\oint dz \hat{z}^{-k} z^{n-1}.$$

The warping matrix $A_{nk}(\alpha)$ for a bilinear warping function has previously been calculated in [7]

$$A_{nk} = \frac{1}{2\pi i}\oint dz \left(\frac{z-\alpha}{1-\alpha z}\right)^{-k} z^{n-1}.$$

The integration is carried out using the Cauchy integral formula

$$\frac{1}{n!}\left.\frac{d^{(n)}F(z)}{dz^{(n)}}\right|_{z=z_0} = \frac{1}{2\pi i}\oint d\xi \frac{F(\xi)}{(\xi-z_0)^{(n+1)}}$$

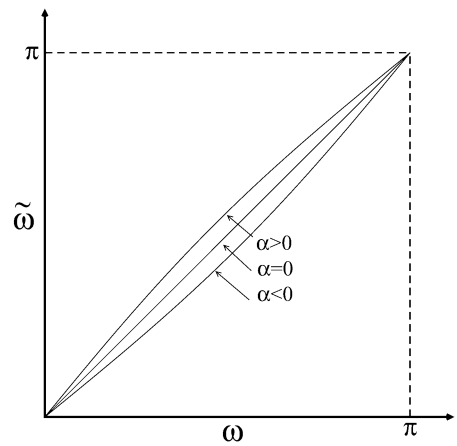

Fig. 6. Example of bilinear VTN warping functions $\tilde{\omega} = g_\alpha(\omega)$ for $\alpha = -0.1, 0, 0.1$.

which yields

$$A_{nk}(\alpha) = \frac{1}{(k-1)!} \sum_{m=\max(0,k-n)}^{k} \binom{k}{m} \times \frac{(m+n-1)!}{(m+n-k)!} (-1)^m \alpha^{(2m+n-k)}. \quad (15)$$

IV. Discussion of the Structure of the Transformation Matrix

After having obtained analytical solutions for the transformation matrix $A_{nk}(\alpha)$ in the last Section, we now want to discuss the structure of the resulting matrices. We will see that the matrices have a common structure even though the functional forms in (9), (14), and (15) look quite different. Motivated by this common structure we will give a possible approximation to reduce the number of matrix elements to be calculated.

In the following we show pictures of the transformation matrices for values of the warping parameter α which occur typically in speech recognition experiments. As can be seen from Figs. 7–12, all matrices are dominated by the diagonal elements and the shapes of the matrices are similar.

We will now show that the dominance of the diagonal elements is a consequence of a general characteristic of typical warping functions rather than the actual functional form. This will lead to an approximation of the transformation matrix by a quindiagonal matrix.

Warping factors estimated for real speakers lead to warping functions which deviate only slightly from unity. This can also be seen from Figs. 4–6 where we used typical values of the warping factor α. Hence we write the warping function in the following form:

$$g_\alpha^{(-1)}(\tilde{\omega}) = \tilde{\omega} - \delta_\alpha(\tilde{\omega}) \text{ with } \delta_\alpha(\tilde{\omega}) \ll 1. \quad (16)$$

Further we assume the deviation from unity $\delta_\alpha(\tilde{\omega}) := \tilde{\omega} - g_\alpha^{(-1)}(\tilde{\omega})$ to be smooth and to be either concave or convex.

We start again with (7)

$$\begin{aligned} A_{nk}(\alpha) &= \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega} \cos(\tilde{\omega} n) \cos\left(g_\alpha^{(-1)}(\tilde{\omega})k\right) \\ &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \left[\cos\left(\tilde{\omega} n + g_\alpha^{(-1)}(\tilde{\omega})k\right) \right. \\ &\qquad \left. + \cos\left(\tilde{\omega} n - g_\alpha^{(-1)}(\tilde{\omega})k\right)\right]. \end{aligned}$$

Substituting (16) into (7) and using

$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y)$$

yields

$$\begin{aligned} A_{nk}(\alpha) &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \cos[(n+k)\tilde{\omega}] \cos[k\delta_\alpha(\tilde{\omega})] \\ &\quad - \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \sin[(n+k)\tilde{\omega}] \sin[k\delta_\alpha(\tilde{\omega})] \\ &\quad + \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \cos[(n-k)\tilde{\omega}] \cos[k\delta_\alpha(\tilde{\omega})] \\ &\quad + \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \sin[(n-k)\tilde{\omega}] \sin[k\delta_\alpha(\tilde{\omega})]. \end{aligned}$$

We now expand this equation for small $\delta_\alpha(\tilde{\omega})$ up to linear terms in $\delta_\alpha(\tilde{\omega})$ for $n \neq k$

$$\begin{aligned} A_{nk}(\alpha) &= \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \cos[(n+k)\tilde{\omega}] \\ &\quad - k \sin[(n+k)\tilde{\omega}]\delta_\alpha(\tilde{\omega}) \\ &\quad + \frac{s_k}{\pi} \int_0^\pi d\tilde{\omega} \cos[(n-k)\tilde{\omega}] \\ &\quad + k \sin[(n-k)\tilde{\omega}]\delta_\alpha(\tilde{\omega}) + \mathcal{O}(\delta_\alpha^2). \end{aligned}$$

We have assumed that $\delta_\alpha(\tilde{\omega})$ is smooth and has a uniform curvature. Thus and because of $\delta_\alpha(\tilde{\omega}) \ll 1$, $\delta_\alpha(\tilde{\omega})$ can be considered as constant w.r.t. $\tilde{\omega}$ within each period

$$\left[(m-1)\frac{2\pi}{n \pm k}, m\frac{2\pi}{n \pm k}\right]$$

$(m = 1, \ldots, \lfloor (n \pm k)/2 \rfloor)$ if the sine term oscillates quickly, i.e., for larger values of $(n+k)$ and $(n-k)$, respectively. Hence, $A_{nk}(\alpha)$ can be approximated by

$$A_{nk}(\alpha) \cong \delta_{nk} + \frac{k}{\pi} \int_0^\pi d\tilde{\omega} \sin[(n-k)\tilde{\omega}]\delta_\alpha(\tilde{\omega}) \quad (17)$$

where the second term contributes significantly only for small values of $(n-k)$, $n \neq k$, i.e., close to the diagonal elements and becomes small for large values of $|n-k|$. Thus the diagonal and only a few off-diagonal elements with small values of $|n-k|$ will dominate the matrix.

For $n = k$ the expansion has to be extended up to quadratic terms in $\delta_\alpha(\tilde{\omega})$ because the term linear in $\delta_\alpha(\tilde{\omega})$ vanishes as can be seen from (17)

$$A_{nn} = 1 - s_n \frac{n}{2} \delta_\alpha^2(\tilde{\omega}). \quad (18)$$

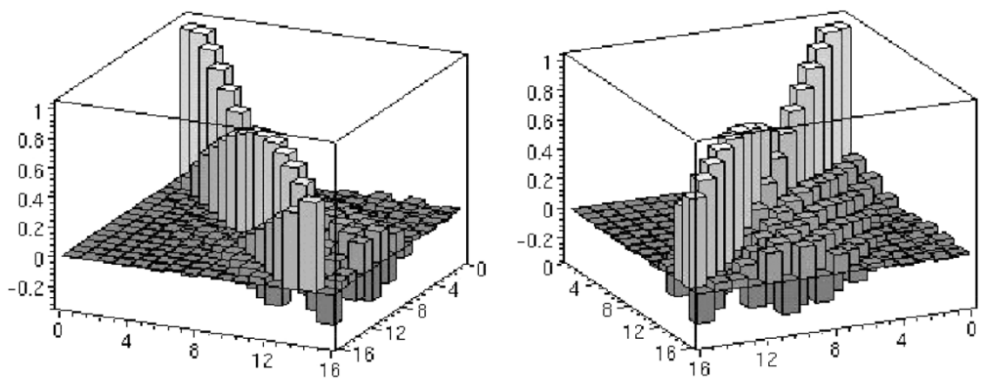

Fig. 7. Matrix for piecewise linear warping function, $\alpha = 0.9$.

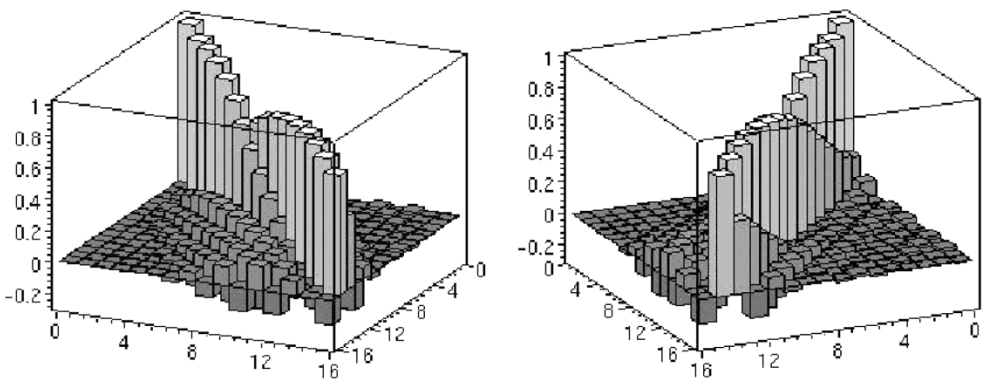

Fig. 8. Matrix for piecewise linear warping function, $\alpha = 1.1$.

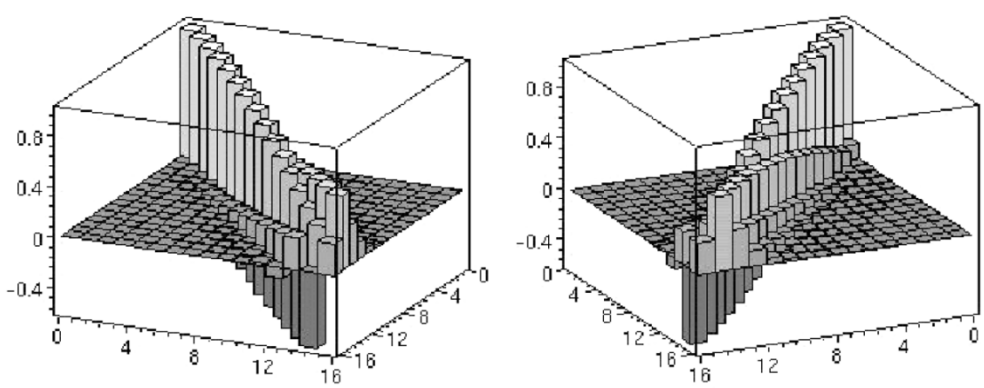

Fig. 9. Matrix for quadratic warping function, $\alpha = -0.5$.

Combining (17) and (18) we yield

$$A_{nk}(\alpha) \cong \left(1 - \frac{s_n n}{2}\delta_\alpha^2(\tilde{\omega})\right)\delta_{nk} + \frac{k}{\pi}\int_0^\pi d\tilde{\omega}\sin[(n-k)\tilde{\omega}]\delta_\alpha(\tilde{\omega}). \quad (19)$$

In combination with Figs. 7–12 we see that the matrix $A_{nk}(\alpha)$ can be approximated by a quindiagonal matrix for values of the warping factor α typical for automatic speech recognition, i.e.,

$$A_{nk}(\alpha) \cong 0 \text{ for } |n-k| > 2.$$

We will see in Section V that the matrix can be even approximated by a tridiagonal matrix when using the Mel frequency scale.

V. Integration of Mel Frequency Scale

We will show in this Section that Mel frequency scaling can equally well be integrated into the framework of VTN as linear transformation on the cepstrum. Mel frequency warping is applied during signal analysis to adjust the spectral resolution to that of the human ear [16]

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

So far, we have not considered Mel frequency warping. There are two possible ways to include Mel frequency warping into the framework of VTN as linear transformation of the cepstral coefficients (CC)

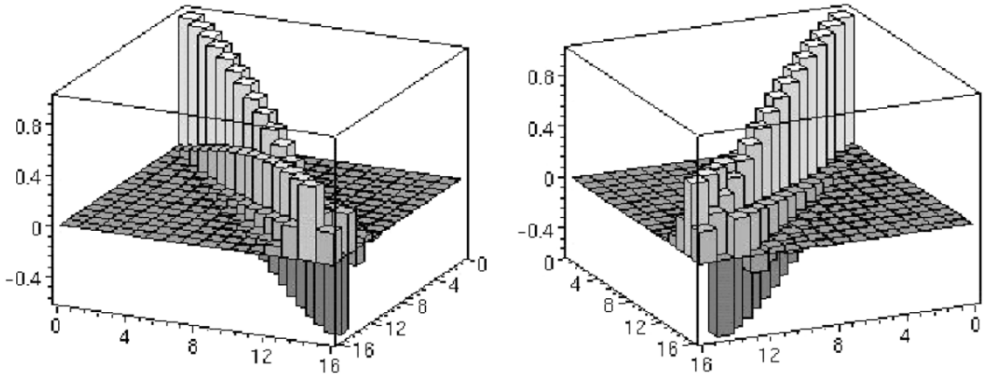

Fig. 10. Matrix for quadratic warping function, $\alpha = +0.5$.

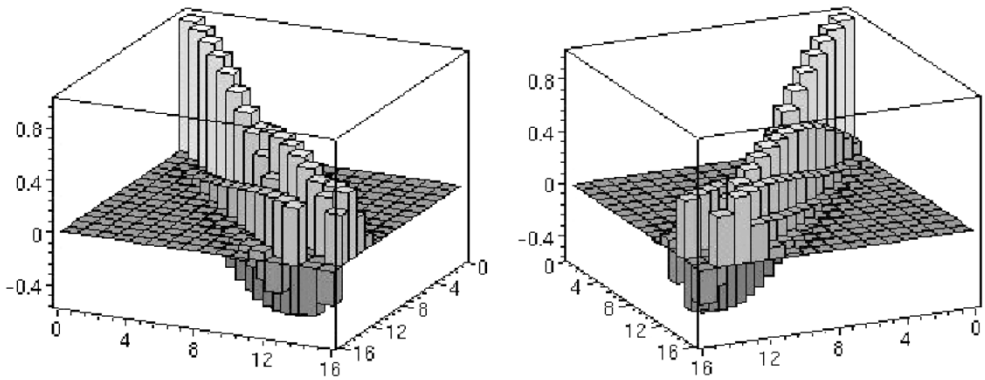

Fig. 11. Matrix for bilinear warping function, $\alpha = +0.1$.

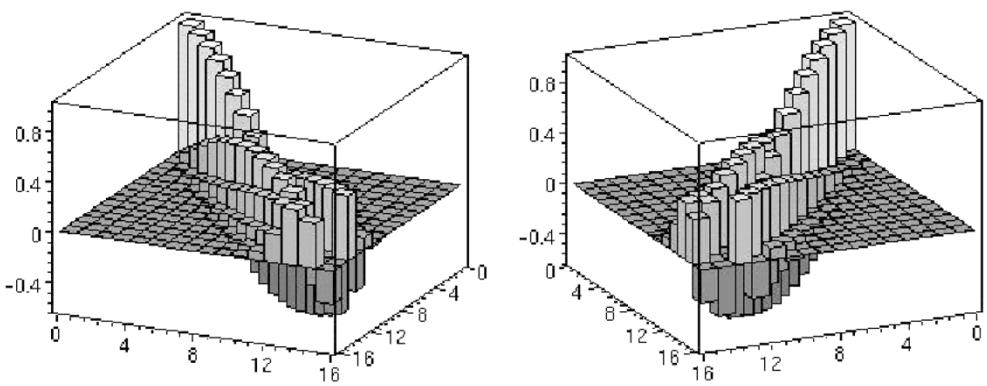

Fig. 12. Matrix for bilinear warping function, $\alpha = -0.1$.

A) to express the VTN-Mel warped CC as a linear function of the original, unwarped CC, or;
B) to express the VTN-Mel warped CC as a linear function of Mel-warped CC (MFCC) (without VTN).

A. From Plain CC to VTN-Mel Warped CC

We have shown in Section II 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. Thus, the combination of VTN and subsequent Mel warping still amounts to a linear transformation in the cepstral domain. VTN is typically applied before Mel scale warping; hence the combination of both warping steps becomes

$$g_{\mathrm{mel}}(g_\alpha(\omega)) : \omega \rightarrow \tilde{\omega}_{\mathrm{mel}} = B \cdot \lg\left(1 + \frac{g_\alpha(\omega) \cdot f_s}{2\pi \cdot 700\,\mathrm{Hz}}\right) \quad (20)$$

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

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

to meet the requirement $g_{\mathrm{mel}}(\pi) = \pi$. Inserting (20) into (7) leads to

$$A_{nk}(\alpha) = \frac{2s_k}{\pi} \int_0^\pi d\tilde{\omega}_{\mathrm{mel}} \cos(\tilde{\omega}_{\mathrm{mel}} n) \times \cos\left(g_\alpha^{(-1)}\left(g_{\mathrm{mel}}^{(-1)}(\tilde{\omega}_{\mathrm{mel}})\right) k\right). \quad (21)$$

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

B. From MFCC to VTN Warped MFCC

We will show in Section VII that VTN is equivalent to a parameterized constrained MLLR transformation. MLLR transforms parameters of the ASR (means and variances) which have typically been estimated from Mel warped 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 unwarped 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 $\tilde{c}_n^{\mathrm{mel}}(\alpha)$

$$\tilde{c}_n^{\mathrm{mel}}(\alpha) = \frac{s_k}{\pi} \int_0^{\pi} d\tilde{\omega}_{\mathrm{mel}} \ln \left| \hat{X}(\tilde{\omega}_{\mathrm{mel}}) \right| \cos(\tilde{\omega}_{\mathrm{mel}} n). \quad (22)$$

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

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

and the warped spectrum $\left\{\hat{X}(\tilde{\omega}_{\mathrm{mel}})\right\}$ is obtained according to

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

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

$$\ln |X(\omega)|^2 = \ln \left|\hat{X}(\omega_{\mathrm{mel}})\right|^2 = 2 \sum_{k=0}^{K} c_k^{\mathrm{mel}} \cos(\omega_{\mathrm{mel}} k). \quad (23)$$

As before, inserting (23) into (21) results in

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

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

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

Finally, we obtain

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

with

$$A_{nk}^{\mathrm{mel}}(\alpha) = \frac{2 s_k}{\pi} \int_0^{\pi} d\tilde{\omega} \cos(\tilde{\omega} n) \times \cos\left(g_{\mathrm{mel}} \circ g_\alpha^{(-1)} \circ g_{\mathrm{mel}}^{(-1)}(\tilde{\omega}) k\right). \quad (24)$$

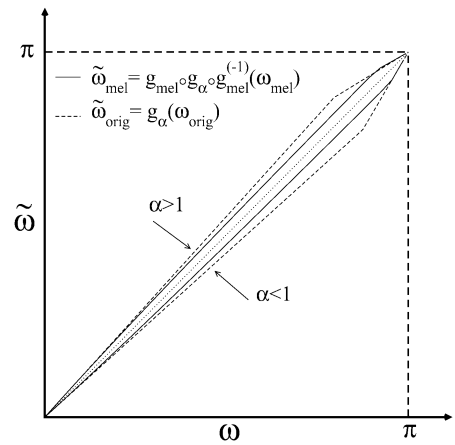

Fig. 13. Effective warping function $g_{\mathrm{mel}} \circ g_\alpha \circ g_{\mathrm{mel}}^{(-1)}$ for MFCC as function of the Mel frequency ω_{mel} (straight) in comparison to the warping function g_α for plain CC as function of the original frequency ω_{orig} (dashed).

Hence, the cepstral coefficients $\tilde{c}_n^{\mathrm{mel}}(\alpha)$ 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 highly nonlinear transformation $g_{\mathrm{mel}} \circ g_\alpha^{(-1)} \circ g_{\mathrm{mel}}^{(-1)}$, the integral in (24) may not be solved analytically. Nevertheless, the transformation matrix can be calculated numerically.

The effective warping function $\tilde{\omega}_{\mathrm{mel}} = g_{\mathrm{mel}} \circ g_\alpha \circ g_{\mathrm{mel}}^{(-1)}(\omega_{\mathrm{mel}})$ is shown in Fig. 13 and examples for transformation matrices are shown in Figs. 14 and 15.

The cepstral coefficients $\tilde{c}_n^{\mathrm{mel}}(\alpha)$ obtained by the method presented here are identical to those calculated by explicitly warping the spectrum during signal analysis as presented in [14].

Comparing the resulting warping function $\tilde{\omega}_{\mathrm{mel}} = g_{\mathrm{mel}} \circ g_\alpha \circ g_{\mathrm{mel}}^{(-1)}(\omega_{\mathrm{mel}})$ (straight line in Fig. 13) for MFCC as function of the Mel frequency with the warping function g_α for plain CC (dashed line) as function of the original frequency, we see that $g_{\mathrm{mel}} \circ g_\alpha \circ g_{\mathrm{mel}}^{(-1)}$ is much closer to identity than g_α. Thus, the transformation matrix is even more diagonally dominant than those obtained for plain cepstral coefficients (compare Figs. 14 and 15 with Figs. 7 and 8). Therefore, the transformation matrix for MFCC can be approximated by a *tri*diagonal matrix rather than a *quin*diagonal matrix (cf. Section IV).

VI. Examples of Spectra Warped by Linear Transformation

In this section, we will show some examples of spectra obtained by applying the linear transformation to the cepstral vectors and discuss the influence of cepstral smoothing at different stages during signal analysis.

In the traditional or the integrated signal analysis approach (see Section II), the smoothing has to be applied by omitting higher *warped* cepstral coefficients because VTN is applied before the cepstrum transformation or integrated into the cepstrum transformation. When using the transformation matrix calculated in this paper the smoothing can be applied at two different

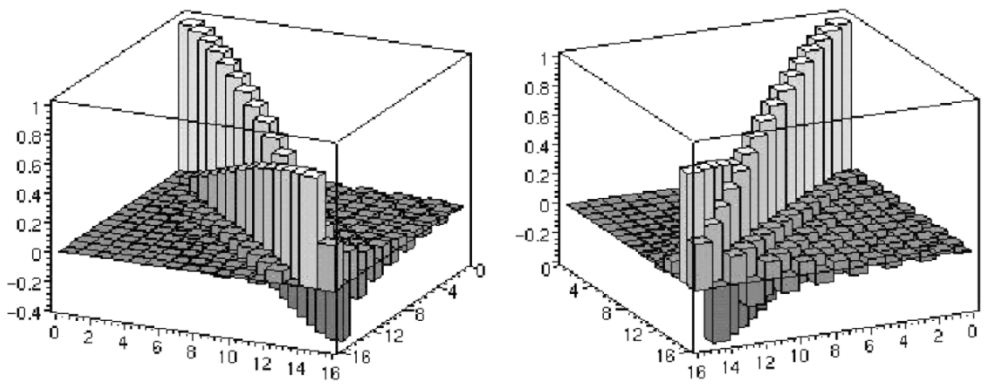

Fig. 14. Matrix for piecewise linear warping function, $\alpha = 0.9$, Mel scale.

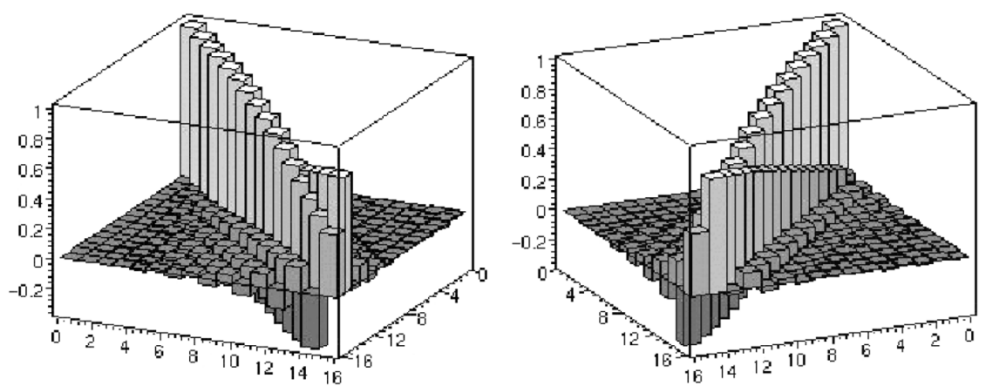

Fig. 15. Matrix for piecewise linear warping function, $\alpha = 1.1$, Mel scale.

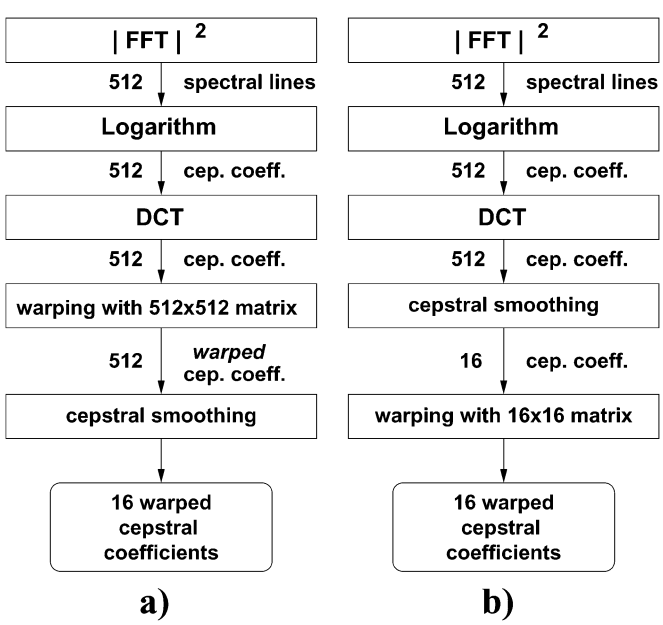

Fig. 16. Schemes for different stages of cepstral smoothing: (a) smoothing after warping and (b) smoothing before warping.

stages: to the *warped* cepstral coefficients (scheme a in Fig. 16) or to the *unwarped* cepstral coefficients (scheme b in Fig. 16). Smoothing by omitting higher unwarped cepstral coefficients (scheme b) results in a loss of frequency resolution but has the advantage of faster computation because less unwarped cepstral coefficients have to be calculated. Additionally, the multiplication of the matrix $A_{nk}(\alpha)$ with the cepstral vector is faster because $A_{nk}(\alpha)$ is of lower dimension. We will discuss the differences between both smoothing schemes in the following. A sample spectrum (Fig. 17, $\alpha = 1.0$) with $N = 512$ spectral lines was transformed into $K = 512$ cepstral coefficients by a discrete cosine transform (DCT)

$$c_k = \frac{4}{N} \sum_{n=0}^{N/2-1} \ln \left| X\left(\mathrm{e}^{i(2\pi n/N)}\right) \right|^2 \cos\left(\frac{2\pi n}{N} k\right).$$

Then the cepstral vector has been transformed using (9) into a piece-wise linear warped cepstral vector of 512 coefficients for warping factors $\alpha = 0.8$ and $\alpha = 1.2$, respectively. Afterwards, the inverse DCT has been applied to the warped cepstral vector in order to obtain a warped spectrum. This last transformation has been carried out for demonstration only; in practice the warped cepstral vector is used for further processing. The resulting spectra are shown in Fig. 17. A comparison of the warped cepstral coefficients obtained by the method presented here with those computed from the spectrum as described in [14] reveals no differences.

As an additional example we show the effect of cepstral smoothing in Fig. 18. Again, the spectrum shown in Fig. 17 has been transformed into 512 cepstral coefficients and has now been smoothed by transforming back with only the first 16 cepstral coefficients ($\alpha = 1.0$ in Fig. 18). The warped

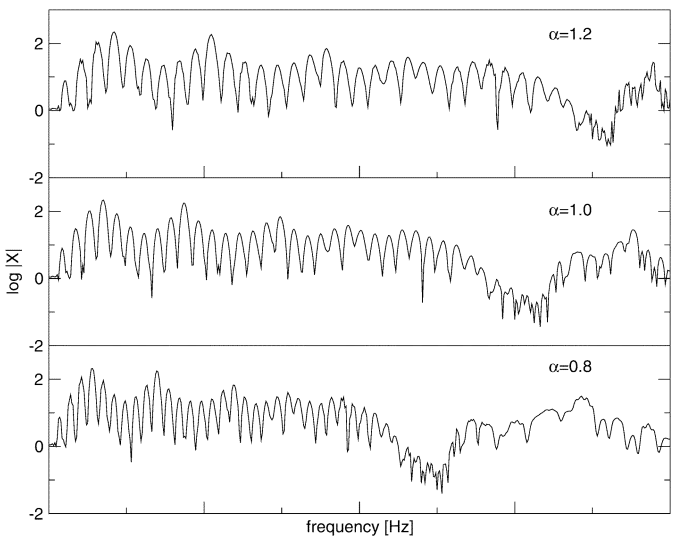

Fig. 17. Example of warped spectra with warping factors $\alpha = 0.8$ and $\alpha = 1.2$.

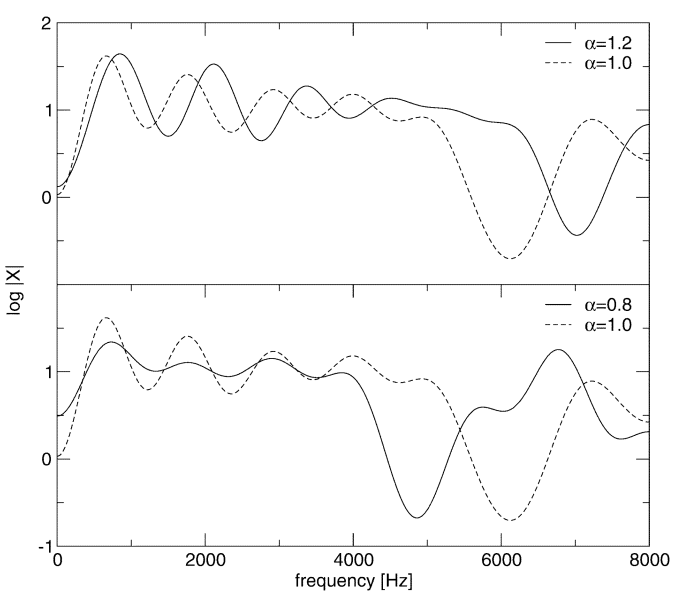

Fig. 18. Example of a smoothed spectrum; the cepstrum was warped with a 512×512 matrix and subsequently reduced to 16 coefficients (scheme a).

spectra have been obtained by calculating 512 cepstral coefficients, transforming them with (9) into 512 warped cepstral coefficients, and subsequent smoothing by transforming back with only the first 16 warped cepstral coefficients (scheme a in Fig. 16). As said before, we can exactly reproduce the warping obtained from [14] only if we apply scheme a. If we first smooth by calculating only the first 16 cepstral coefficients and warp hereafter using a 16×16 matrix (scheme b), we obtain slightly different results. The difference between both methods is shown in Fig. 19. These differences arise mainly in the part of the spectrum where the slope of the warping function is larger than 1 (cf. Fig. 4), i.e., for $\omega > \omega_0$ for $\alpha = 0.8$ and $\omega < \omega_0$ for $\alpha = 1.2$ ($\omega_0 = (7/8)\pi \hat{=} 7000$ Hz).

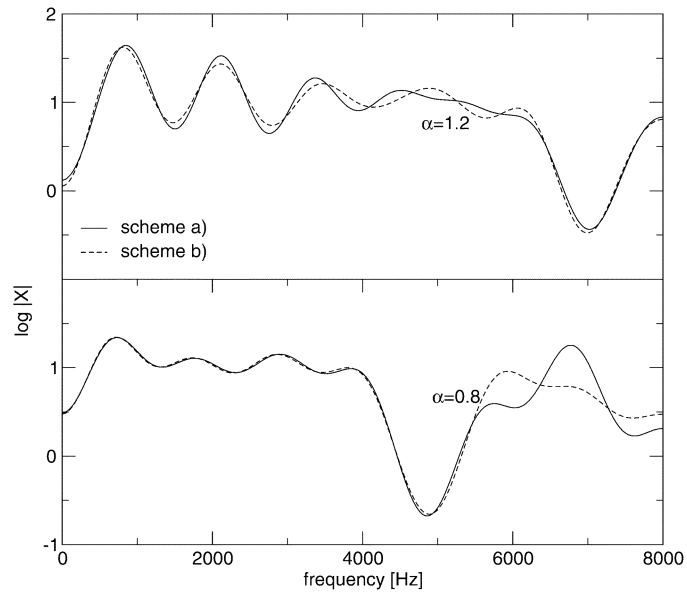

Fig. 19. Effect of different order of warping and smoothing.

VII. Interdependence of VTN and MLLR

Adaptation and normalization are commonly viewed as different techniques to reduce the mismatch between training and testing conditions [1]. We have shown in [17] that adaptation and normalization can be put together in the same mathematical framework. Moreover they are identical in terms of Bayes' decision rule. As a consequence of VTN being a linear transformation of the cepstral vector we can show a strong interdependence of two widely used normalization (VTN) and adaptation (MLLR) techniques if Gaussian emission probabilities are applied.

Most of today's ASR 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

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

with

$$\hat{\mu} = \mathbf{A}^{-1}\mu \text{ and } \hat{\mathbf{\Sigma}} = \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-1^T}. \tag{25}$$

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 (25) describe a special case of MLLR which is called *constrained* MLLR [18], [19]. The attribute constrained refers to the use of the same matrix $\mathbf{A}$ for the transformation of the mean and variance.

In [10], Uebel and Woodland have found empirically 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.

VIII. Jacobian Determinant for VTN

To estimate the unknown warping factor α, we proceed as follows: For each speaker r we are given labeled training data (X_r, W_r) where X_r denote the sequence of acoustic data and W_r the sequence of spoken words [17]. 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 to obtain the unknown warping factor α of a feature transformation function g_α [20]

$$\hat{\alpha}_r = \underset{\alpha}{\operatorname{argmax}} \left\{ p(g_\alpha(X_r)|W_r) \cdot \left| \frac{dg_\alpha(X_r)}{dX_r} \right| \right\}. \tag{26}$$

The last term denotes the Jacobian determinant of the transformation. It can be omitted if no direct comparison of probability values is carried out with differently "normalized" distributions. In a typical VTN approach the warping factor α is chosen by comparing the scores obtained by a forced alignment of the same utterance warped with a set of discrete warping factors. Here the Jacobian determinant needs to be taken into account [20]. Often the Jacobian determinant is assumed to be flat as function of α and is approximated to be independent of α and thus neglected.

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 (26), 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. 20). As one can see, the function $-\log|\det \mathbf{A}|$ penalizes warping factors with larger deviations from 1 and thus keeps the transformation closer to identity. The dependence of the Jacobian determinant on α is weaker for MFCC (dashed line in Fig. 20) 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. 13).

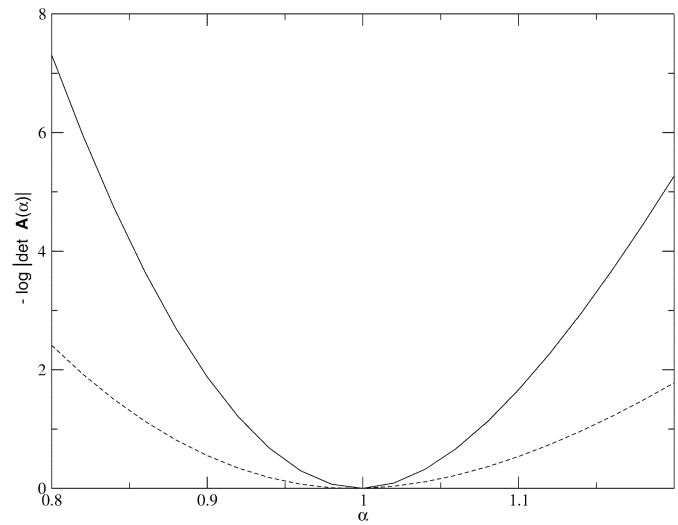

Fig. 20. Plot of $-\log|\det \mathbf{A}|$ for piece-wise linear warping of 16 cepstral coefficients as function of α. Straight line: original frequency scale, dashed line: Mel frequency scale.

The warping factor α is usually determined by warping the speech signal with a discrete set of warping factors and choosing the most likely one according to (26), neglecting the Jacobian determinant. With the limitation of possible warping factors to a set of discrete values close to the identity transformation the span of possible transformations is reduced. Thus the errors caused by improper normalization of the probability distribution by neglecting the Jacobian determinant are small and do not cause VTN to fail.

When not restricting the span of possible transformations as with Maximum Likelihood Linear Transforms (MLLT) [19], the normalization of the acoustic feature vectors will fail without proper normalization of the probability distributions. We have observed in experiments that unrestricted linear transformation with neglecting the Jacobian determinant cause the automatic speech recognition system to spuriously recognize silence only. A similar result has been reported in [9]. In the following we want to give an empirical explanation of these results. Let us assume one tries to transform the acoustic feature vectors with a transformation matrix A without any restrictions to the values of the matrix elements and neglect the Jacobian determinant. This results in a density function

$$d_A(x|\mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \times \exp\left\{ -\frac{1}{2}(\mathbf{A}x - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{A}x - \mu) \right\}$$

which is not normalized. If the transformation matrix $\mathbf{A}$ is estimated using a maximum (log)likelihood criterion, the following will most probably happen: Among the possible means μ of the ASR system one or more means are assigned to the *silence* model, which contains usually low spectral energy and thus small cepstral coefficients. The transformation will now map the acoustic feature vectors onto the silence model by simply shrinking the values of the cepstral coefficients[1]. As a simplified example, assume that the mean vector of the silence model

[1]If the silence model is treated separately, the estimated transformation will map the feature vectors to the mean of a consonant model with low spectral energy

consists of zeroes only. If the transformation matrix is identical to $\mathbf{0}$ we have a perfect match of all feature vectors with the silence model independent of their actual values. But if the density is properly normalized by taking the Jacobian determinant into account, this results in an additional term $\log|\det \mathbf{A}|$ in the log-likelihood function

$$\log \mathcal{L}\left(x_1^T; d_A(x|\mu,\Sigma)\right) = \sum_{t=1}^{T} \log d_A(x_t|\mu,\Sigma) + T \cdot \log|\det \mathbf{A}|.$$

This second term $\log|\det \mathbf{A}|$ plays the role of a penalty for the transformation onto the silence model because it penalizes transformation to small cepstral values. In our example, this would result in an infinite penalty which balances the perfect match of the first term.

The reason why VTN in contrast to MLLT does not fail without proper normalization lies in the limitation of the warping factors and the specification of the warping function. The resulting transformation has not enough degrees of freedom and thus the matrix is close to the identity matrix.

IX. Conclusions

In this section, we want to discuss some consequences and applications of our presented results.

One obvious application is the correct treatment of the Jacobian determinant. With the results presented in this paper it is for the first time possible to take the Jacobian determinant into account for the most widely used warping functions. Although VTN works quite well without taking the Jacobian determinant into account a systematic study of the effects has not been done yet. The correct treatment should result in an improved warping factor estimation and thus improved recognition accuracy.

Currently, the warping factor is estimated similarly in nearly all VTN approaches. The spectrum is warped with a discrete set of warping factors and a forced alignment of the utterance is performed using a given transcription (either the correct one or obtained by a preliminary recognition pass). The warping factor for which the alignment yields the best score is then chosen. This procedure is quite time consuming as it requires a signal analysis and a time alignment for each warping factor in the set. Although there are some approaches to speed up the warping factor estimation [17], calculating the warping factor analytically from the cepstral vectors would be preferable. As the transformation matrix can now be calculated analytically, the warping factor could be directly calculated rather than being estimated by a grid search. Unfortunately, the resulting objective function for an HMM with Gaussian emission probability with variance Σ_s and mean μ_s and posterior probability $\gamma_s(t)$ for being in state s at time t

$$\alpha^{opt} = \underset{\alpha}{\operatorname{argmax}} \sum_{t=1}^{T}\sum_{s=1}^{S}\left\{\gamma_s(t)\left(\log|\det \mathbf{A}(\alpha)| - \frac{1}{2}(\mathbf{A}(\alpha)x_t-\mu_s)^T\Sigma_s^{-1}(\mathbf{A}(\alpha)x_t-\mu_s)\right)\right\} \quad (27)$$

can hardly be solved explicitly, especially for MFCC (cf. (24)).

Using the standard grid search approach for estimating the warping factor, the approximation of the transformation matrix derived in Sections IV and V may be of less importance. In that case, the computing the approximated matrix has similar costs as computing the complete matrix A_{nk}. But the approximation may be useful in another way: as we have shown in Section VII, VTN can be considered as a special case of constrained MLLR. Thus, the approximation motivated in Sections IV and V can be used for the case of a feature space transform as described in [19]. In that case, a tridiagonal matrix could result in a similar performance with much less parameters. First experimental results are presented in the following section, a detailed analysis goes beyond the scope of this paper and requires additional studies.

X. Experiments

We have shown in Sections IV and V that the VTN warping matrix may be approximated by a band-diagonal structured matrix. Similar approximations have been proposed empirically in [10], [21]. Although we have shown the band-structure for the VTN transformation matrix A, a similar structure can also been anticipated for the corresponding MLLR matrix $W = A^{-1}$. It has been shown in [22] that the entries of inverses of band matrices decay fast enough to justify a similar approximation, though the number of bands may be higher for the inverse matrix.

The experimental setup is as follows. We have restricted the ordinary MLLR matrix $\mathbf{W}$ to a band structure: $W_{ij} = 0$ for $j < i-\delta$ and $j > i+\delta$, i.e., we used only δ bands. Similar results have been reported in [21], where the authors used an iterative approach to solve the occurring equations. Unfortunately, the convergence of the iteration is not guaranteed and in fact we observed convergence problems using this iterative approach. On the other hand, the estimation of the MLLR matrix is a quadratic optimization problem and thus can be solved in closed form. As the solution for the MLLR matrix can be obtained row-wise, it is easy to restrict the resulting matrix to a band structure. We use the notation introduced in [23]

$$\hat{\mu}_{s,r} = \boldsymbol{W}'_{s,r}\mu_s + b_{s,r} \quad (28)$$

where s denotes the HMM state and r the speaker. For notational convenience, (28) is usually rewritten in the form

$$\hat{\mu}_{s,r} = \boldsymbol{W}_{s,r}\xi_s \quad (29)$$

where ξ_s denotes the extended mean vector $\xi_s = \left[1\mu_s^\top\right]^\top$ and $\boldsymbol{W}_{s,r}$ is the $n\times(n+1)$-matrix $[b_{s,r}\boldsymbol{W}'_{s,r}]$. The adaptation matrix $\boldsymbol{W}_{s,r}$ is estimated by maximum likelihood, given adaptation data x_1^T and a word sequence w_1^N

$$\boldsymbol{W}_{s,r}^{\mathrm{ML}} = \underset{W_{s,r}}{\operatorname{argmax}}\, p\left(x_1^T|w_1^N,\mu_s,\boldsymbol{\Sigma}_s;\mathbf{W}_{\mathbf{s,r}}\right). \quad (30)$$

Assuming diagonal covariance matrices

$$\boldsymbol{\Sigma}_s = \operatorname{diag}\left(\sigma_1^2,\ldots,\sigma_d^2,\ldots,\sigma_D^2\right)$$

TABLE I
RECOGNITION RESULTS (WER IN %) ON THE WSJ SPOKE 3 CORPUS FOR DIFFERENT AMOUNTS OF ADAPTATION DATA AND VARYING NUMBER OF BANDS δ FOR THE MLLR MATRIX

		Number of adaptation utterances				
		1	5	10	20	40
baseline		34.7				
Number of bands	0	27.7	23.1	20.2	17.3	17.4
	1	27.0	22.2	18.6	16.3	15.5
	2	**27.0**	22.0	17.9	15.6	14.5
	3	27.3	21.3	17.7	14.9	14.0
	4	27.5	**21.2**	16.7	14.1	13.3
	5	28.2	21.9	**16.7**	14.2	13.2
	7	28.8	22.4	16.8	**14.0**	12.9
	full	43.8	33.2	23.2	15.6	**12.5**

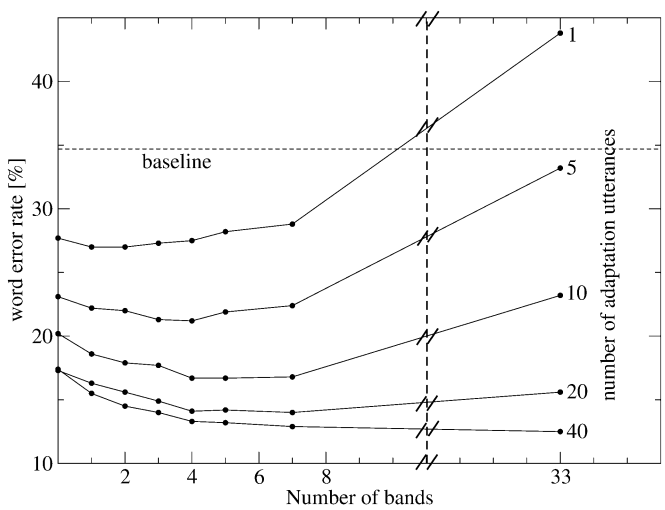

Fig. 21. Plot of recognition results from Table I.

the solution for the i-th row $w^{(i)}$ of $\mathbf{W}$ is given by the following equation:

$$\sum_{k=0,i-\delta}^{i+\delta} G_{jk}^{(i)} w_k^{(i)} = Z_j^{(i)} \quad j = 0, i-\delta, \ldots, i+\delta \tag{31}$$

with

$$\boldsymbol{G}^{(i)} = \sum_{s=1}^{S} \frac{1}{\sigma_{si}^2} \xi_s \xi_s^\top \sum_{t=1}^{T} \gamma_s(t) \tag{32}$$

$$Z^{(i)} = \sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_s(t) \frac{1}{\sigma_{si}^2} x_{ti} \xi_s^\top. \tag{33}$$

The $k = 0$ and $j = 0$ terms in (31) result from the bias term b_s, r. Thus, for each row we have to invert only a part of $\boldsymbol{G}^{(i)}$ of size[2] $2\delta + 1$ for the band matrix and 1 for the bias term $(2\delta + 1 + 1) \times (2\delta + 1 + 1)$ to obtain the solution of (31).

We used the Spoke3 test set of the Wall Street Journal task for the experimental evaluation. This consists of 40 utterances for test and adaptation, respectively, from ten nonnative speakers. The decision-tree clustered across-word models were trained on the SI-84 training set. The acoustic model consists of 4000 HMM states with 185k Gaussian densities and one pooled diagonal covariance. An LDA is applied to seven consecutive vectors resulting in a 32 dimensional feature vector. Details of the RWTH speech recognition system can be found in [24]. We used a trigram language model trained on the lm training data provided by NIST and a 5k vocabulary. The baseline speaker independent (SI) word error rate (WER) is 34.7%, the SI WER on the Nov94 WSJ0-5k evaluation corpus is 3.75%.

As can be seen from Table I and Fig. 21, the largest improvement is obtained from the first few bands of the MLLR matrix. This supports our theoretical findings from Sections IV and V that the transformation matrix can be approximated by a tri- or quindiagonal matrix. Additionally, the full matrix is superior for the complete adaptation set only. Except for very few adaptation data (1 utterance), the best choice for δ is 4, which is very close to the respective optimal value.

[2] $2\delta + 1$ for the band matrix and 1 for the bias term.

XI. SUMMARY

We have shown that vocal tract normalization can be expressed as a linear transformation of the cepstral vector for *arbitrary* invertible warping functions. For the case of piece-wise linear, quadratic and bilinear warping we have exemplarily derived an analytic solution for the transformation matrix. Using special properties of typical warping functions we have shown that the transformation matrix can be approximated by a quindiagonal matrix for plain cepstral coefficients or a tridiagonal matrix for MFCC, respectively. Experimental studies show that a band diagonal restriction of the matrix in MLLR adaptation clearly outperforms the full transform for small amounts of adaptation data and is only slightly inferior to the full transform for large amounts of adaptation data. Computing the transformation matrix for VTN allows us to properly normalize the probability distribution with the Jacobian determinant of the transformation. We have shown that VTN amounts to a special case of MLLR, which can explain previous experimental results that improvements obtained by VTN and subsequent constrained MLLR were not additive.

REFERENCES

[1] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, Sofia Antinopolis, France, Aug. 2001, pp. 11–19.
[2] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 346–349.
[3] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 353–356.
[4] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. ASSP-25, Apr. 1977, pp. 183–192.
[5] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Atlanta, GA, May 1996, pp. 339–341.
[6] A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1990.
[7] J. W. McDonough, "Speaker compensation with all-pass transforms," Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, 2000.

[8] J. W. McDonough, G. Zavaliagkos, and H. Gish, "An approach to speaker adaptation based on analytic functions," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, Atlanta, GA, May 1996, pp. 721–724.

[9] S. Cox, "Speaker normalization in the MFCC domain," in *Int. Conf. on Spoken Language Processing*, vol. 2, Bejing, China, Oct. 2000, pp. 853–856.

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

[11] F. K. Nocerino, L. R. Rabiner, and D. H. Klatt, "Comparative study of several distortion measures for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, Apr. 1985, pp. 25–28.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Speech Audio Process.*, vol. 28, pp. 357–366, Aug. 1980.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Jun. 1990.

[14] 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, Jun. 2001, pp. 73–76.

[15] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, Phoenix, AZ, Apr. 1999, pp. 761–764.

[16] S. J. Young, *HTK: Hidden Markov Model Toolkit V1.4. User Manual* Cambridge, U.K., 1993.

[17] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 415–426, Sep. 2002.

[18] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

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

[20] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 190–202, May 1996.

[21] M. Afify and O. Siohan, "Constrained maximum likelihood linear regression for speaker adaptation," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, Beijing, China, Oct. 2000, pp. 861–864.

[22] S. Demko, W. F. Moss, and P. W. Smith, "Decay rates for inverse band matrices," *Math. Comput.*, vol. 43, no. 168, pp. 491–499, Oct. 1984.

[23] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[24] A. Sixtus and H. Ney, "From within-word model search to across-word model search in large vocabulary continuous speech recognition," *Comput. Speech Lang.*, vol. 16, no. 2, pp. 245–271, May 2002.

Michael Pitz received the Diploma degree in physics from RWTH Aachen University, Aachen, Germany, in 1996 and the Dr.rer.nat. degree in computer science from RWTH Aachen University in 2005.

From November 1996 to March 1998, he was a Research Assistant with the Institute of Theoretical Physics B, RWTH, and the II. Physikalisches Institut, RWTH, where he worked on statistical physics and condensed matter physics. In April 1998, he joined the Lehrstuhl für Informatik VI, RWTH. His research interests cover automatic speech recognition, pattern recognition, acoustic and stochastic modeling, signal analysis and speaker adaptation and normalization.

Hermann Ney (M'86) received the Diploma degree in physics from Goettingen University, Goettingen, Germany, in 1977, and the Dr.-Ing. degree in electrical engineering from Braunschweig University of Technology, Braunschweig, Germany, in 1982.

He has been working in the field of speech recognition, natural language processing, and stochastic modeling for more than 20 years. He has authored and co-authored more than 200 papers. In 1977, he joined Philips Research in Germany. In 1985, he was appointed Department Head. All of his career at Philips was in research and advanced development of basic technology for pattern recognition, speech recognition, and spoken language systems. From October 1988 to October 1989, he was a Visiting Scientist at Bell Laboratories, Murray Hill, NJ. In July 1993, he joined the Computer Science Department of RWTH Aachen, University of Technology, as a Full Professor. His responsibilities include planning, directing and carrying out research for national, European, and industrial sponsors and supervising Ph.D. students. He has been and is a peer reviewer for a number of major scientific journals. He is on the editorial board of several major scientific journals.

Dr. Ney was on the Executive Board of the German Section of the IEEE from 1992 to 1998. For the term 1997 to 2000, he was a member of the Speech Technical Committee of the IEEE.