Vocal Tract Normalization Equals Linear Transformation in Cepstral Space

Michael Pitz, Sirko Molau, Ralf Schlüter, Hermann Ney

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

Abstract

We show that vocal tract normalization (VTN) frequency warping results in a linear transformation in the cepstral domain. For the special case of a piece-wise linear warping function, the transformation matrix is analytically calculated. This approach enables us 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.

1. Introduction

Vocal tract normalization (VTN) tries to compensate for the effect of speaker dependent vocal tract lengths by warping the frequency axis of the power spectrum [2, 5, 3, 9, 10]:

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

The warping function g_α is assumed to be invertible, i.e. strictly monotonic and continuous (see Figure 1).

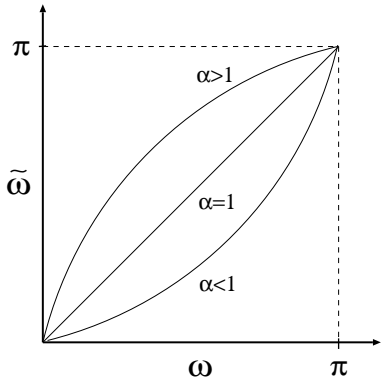

Figure 1: Example of VTN warping functions $\tilde{\omega}_\alpha$ for different values of α.

The relationship between VTN frequency warping and linear transformations in the cepstral domain has been studied before [1, p.199],[4]. However, these investigations were based on special assumptions:

- The VTN frequency warping is restricted to a bilinear transformation [1, p.119],[4].
- The cepstral representation is based on an all-pass or LPC model [4].

In contrast, we will show that there is a general equivalence of VTN frequency warping and a linear transformation of the cepstral vector, independent of these assumptions. A related result has been reported in [6] in the context of spectral distortion measures.

The remainder of the paper is organized as follows: In the second paragraph we show that VTN amounts to a linear transformation of the acoustic vector. The transformation matrices for the cases of linear and piece-wise linear warping are analytically derived in the third paragraph, followed by some examples obtained by warping a given spectrum with our approach. Then we discuss the implications for the normalization of probability distributions when transforming the random variables. The paper is summarized in section 6.

2. Cepstral Representation of VTN Frequency Warping

We consider cepstral coefficients c_k defined by:

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

where ω may either denote the true physical or the Mel frequency scale. Note that the conventional definition of c_0 differs by a factor of 2.
The $n-$th cepstral coefficient of the warped spectrum is

$$\tilde{c}_n(\alpha) = \frac{1}{\pi}\int_0^\pi d\tilde{\omega} \lg |X(e^{ig_\alpha^{(-1)}(\tilde{\omega})})|^2 \cdot \cos(\tilde{\omega} n) \,.$$

In order to obtain the value of the warped power spectrum for a given frequency, we access the unwarped spectrum at the frequency determined by the inverse warping function. This is necessary as in practice only the discrete unwarped spectrum is given. Explicit spectral interpolation for warping is avoided this way.

Now we expand the spectrum $\lg \left|X(e^{ig_\alpha^{(-1)}(\tilde{\omega})})\right|^2$ in a Fourier series:

$$\lg \left|X(\mathrm{e}^{i\omega})\right|^2 = 2\sum_{k=0}^{K} c_k \cos(\omega k) \,.$$

where c_k denotes the k-th cepstral coefficient of the unwarped

spectrum. Interchanging integration and summation yields:

$$\begin{aligned}\tilde{c}_n(\alpha) &= \frac{2}{\pi}\int_0^{\pi}\cos(\tilde{\omega}n)\sum_{k=0}^{K}c_k\cos(g_\alpha^{(-1)}(\tilde{\omega})k)\;d\tilde{\omega}\\ &= \sum_{k=0}^{K}c_k\;\frac{2}{\pi}\int_0^{\pi}\cos(\tilde{\omega}n)\cos(g_\alpha^{(-1)}(\tilde{\omega})k)\;d\tilde{\omega}\\ &= \sum_{k=0}^{K}A_{nk}(\alpha)\;c_k \qquad (3)\end{aligned}$$

with

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

Thus, the vector of warped cepstral coefficients is a linear transformation of the original cepstral coefficients with a transformation matrix $\mathbf{A}(\alpha)$ of dimension $N \times K$. In the case of continuous spectra there may be no upper limit for N and K. In practice, however, we work with discrete spectra. Hence, N and K will be finite, but not necessarily have the same value. Choosing a smaller value of N results in a smoothing of the power spectrum and eliminates the pitch.

3. Analytic Calculation of the Transformation Matrix

3.1. Linear Warping Function

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

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

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

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

For $\alpha = 1$ this simplifies to

$$A_{nk}(1) = \begin{cases}2 & : \quad n=k=0\\ \delta_{nk} & : \quad \text{else}\end{cases}$$

because of the orthonormality of the cosine function. Note that the value for $n = k = 0$ results from our special definition of the zeroth cepstral coefficient c_0.

3.2. Piece-wise Linear Warping Function

To meet the requirement of invertibility, we now consider a piece-wise linear warping function [10, 11] with two parameters (α, ω_0) as shown in Figure 2:

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

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

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

Hence, $g_{(\alpha,\omega_0)}$ depends solely on α.

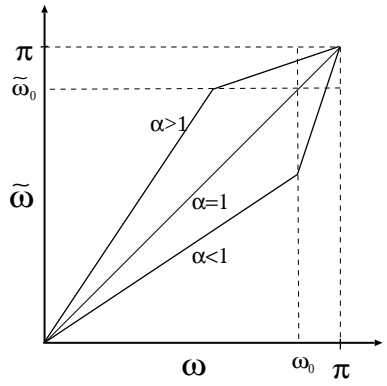

Figure 2: Piece-wise linear warping functions for different values of α

The transformation matrix $A_{nk}(\alpha)$ is computed similar to the linear case:

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

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 $\alpha \neq 1$:

$$\begin{aligned}A_{nk}(\alpha) =& \frac{\sin(n-\alpha^{-1}k)\tilde{\omega}_0}{(n-\alpha^{-1}k)\pi}+\frac{\sin(n+\alpha^{-1}k)\tilde{\omega}_0}{(n+\alpha^{-1}k)\pi}\\ &-\frac{\sin(n-\alpha^{-1}k)\tilde{\omega}_0}{\left(n-\frac{\pi-\alpha^{-1}\tilde{\omega}_0}{\pi-\tilde{\omega}_0}k\right)\pi}-\frac{\sin(n+\alpha^{-1}k)\tilde{\omega}_0}{\left(n+\frac{\pi-\alpha^{-1}\tilde{\omega}_0}{\pi-\tilde{\omega}_0}k\right)\pi}\;.\end{aligned} \qquad (4)$$

This matrix can now be used for VTN alternatively to explicit warping the discrete-frequency power spectrum or the integrated approach described in [5].

3.3. General Warping Functions

We would like to stress again that VTN can always be written as a linear transformation in the cepstral domain independent of the functional form of the invertible warping function (see eqn. (3)). The analytic calculation of the transformation matrix for a non–linear warping function, however, is not as straightforward as in the piece-wise linear case presented above.

4. Examples

In this section we will show some examples of spectra obtained by applying the linear transformation to the cepstral vectors. A sample spectrum (Figure 3, $\alpha = 1.0$) with $N = 512$ spectral lines was transformed into $K = 512$ cepstral coefficients by a discrete cosine transform (DCT):

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

Then the cepstral vector has been transformed into a piece-wise linearly warped (4) 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. A comparison of the warped cepstral coefficients obtained by the method presented here with those computed from the spectrum as described in [5] reveals no differences.

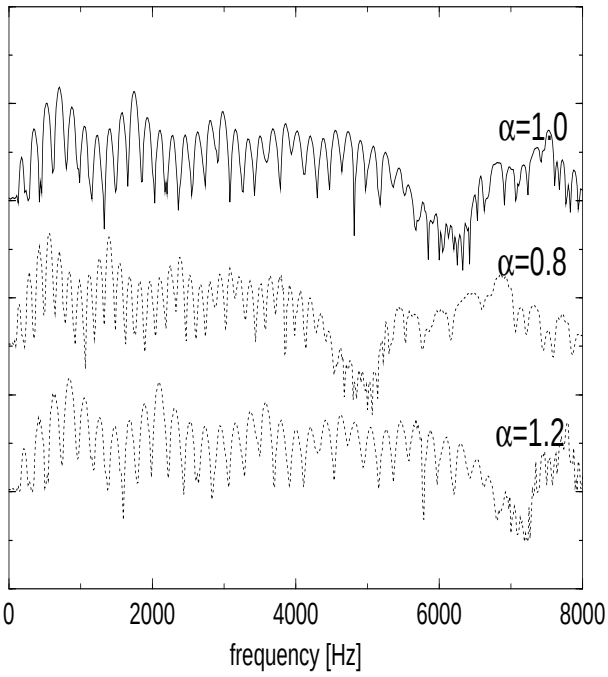

Figure 3: Example of warped spectra with warping factors $\alpha = 0.8$ and $\alpha = 1.2$.

As an additional example we show the effect of cepstral smoothing in Figures 4 and 5. Again, the spectrum shown in Figure 3 has been transformed into 512 cepstral coefficents and has now been smoothed by transforming back with only the first 16 cepstral coefficients ($\alpha = 1$ in Figs. 4, 5). The warped spectra have been obtained by calculating 512 cepstral coefficients, transforming them with (4) into 512 warped cepstral coefficients, and subsequent smoothing by transforming back with only the first 16 warped cepstral coefficients. It should be noted that this time we can exactly reproduce the warping obtained from [5] only if we first compute all 512 cepstral coefficients, warp them using (4), and smooth at this point using only the first 16 of the obtained cepstral coefficients. If we first smooth by calculating only the first 16 cepstral coefficients and warp hereafter using a 16×16 matrix, we obtain slightly different results. The difference between both methods is shown in Figure 6.

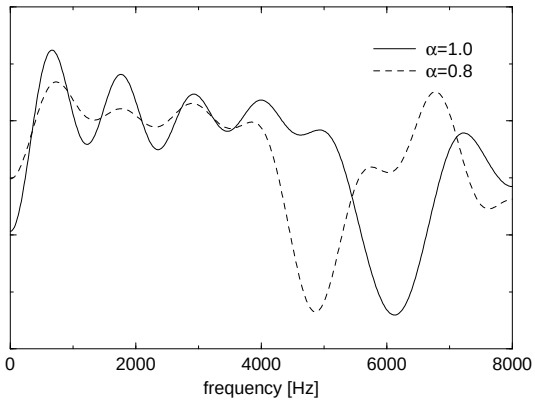

Figure 4: Example of a smoothed spectrum; the cepstrum was warped with a 512×512 matrix ($\alpha = 0.8$) and subsequently reduced to 16 coefficients.

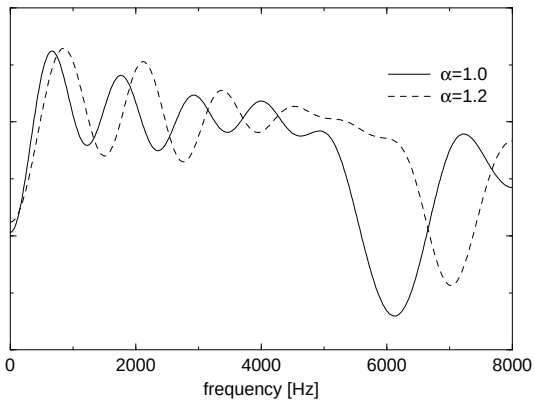

Figure 5: Example of a smoothed spectrum; the cepstrum was warped with a 512×512 matrix ($\alpha = 1.2$) and subsequently reduced to 16 coefficients.

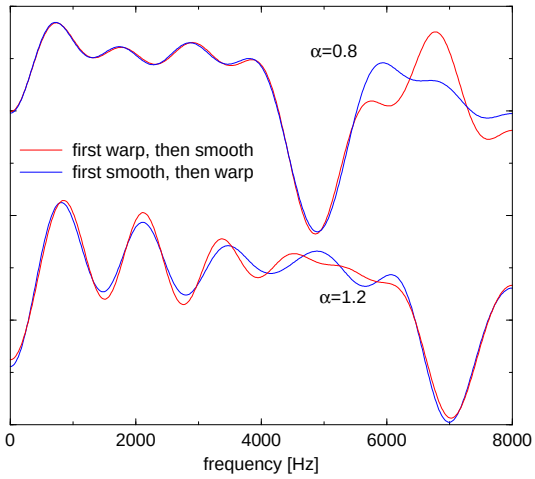

Figure 6: Effect of different order of warping and smoothing

5. Speaker Normalization

In speaker *normalization* the acoustic observation vector is modified, whereas speaker *adaptation* modifies the acoustic model parameters. This will cause the probability distribution to be not properly normalized anymore. To re-normalize the transformed distributions, the Jacobian of the transformation must be taken into account [4, 7].

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 can hardly be calculated. The warping factor α is usually determined by a maximum likelihood criterion. If the correct normalization is neglected, systematic errors in estimating α may occur.

Expressing VTN as a matrix transformation of the acoustic vector ($x \rightarrow \mathbf{A}\,x$) enables us to take the Jacobian into account:

$$\begin{aligned}\mathcal{N}(x|\mu,\Sigma) &\rightarrow \mathcal{N}(\mathbf{A}\,x|\mu,\Sigma)\\ &= \mathcal{N}(x|\mathbf{A^{-1}}\,\mu,\mathbf{A^{-1^T}}\,\Sigma\mathbf{A^{-1}})\\ &= \frac{1}{\sqrt{\det 2\pi\mathbf{A^{-1^T}}\,\Sigma\mathbf{A^{-1}}}}\exp\{\dots\}\\ &= \frac{|\det\mathbf{A}|}{\sqrt{\det 2\pi\Sigma}}\exp\{\dots\}\end{aligned}$$

where in the last step $\mathbf{A}$ is assumed to be square. The practical influence of the Jacobian is subject of current research. A qualitative plot showing the dependency of the Jacobian determinant on the warping factor alpha has been computed numerically for piece-wise linear warping (Figure 3).

The dependency of $|\det\mathbf{A}(\alpha)|$ on α can be used for a refined estimation of α in speaker normalization.

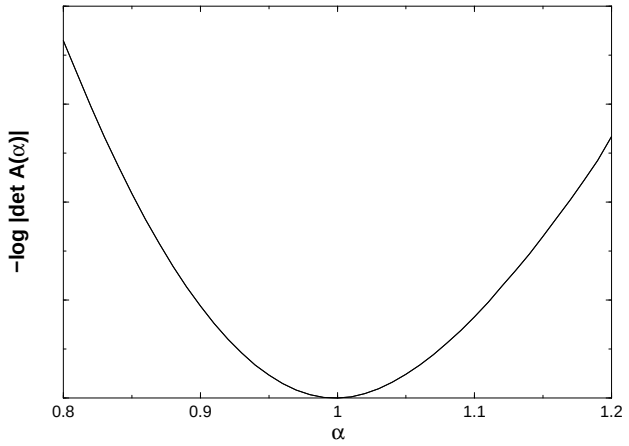

Figure 7: Plot of $-\log\det|\mathbf{A}(\alpha)|$ for piece-wise linear warping as function of α. The scaling of the ordinate is intentionally left out as it depends on the number of cepstral coefficients.

6. Conclusion

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 warping we derived an analytic solution for the transformation matrix. This allows us to re-normalize the probability distribution with the Jacobian of the transformation.

7. References

[1] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition", *Ph. D. Thesis*, Carnegie Mellon University, Pittsburgh, PA, USA, September 1990.

[2] E. Eide, H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, Vol. 1, pp. 346-349, Atlanta, GA, May 1996.

[3] L. Lee, R. Rose "Speaker Normalization Using Efficient Frequency Warping Procedures" Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, Vol. 1, pp. 353-356, Atlanta, GA, May 1996.

[4] J. McDonough, "Speaker Normalization With All-Pass Transforms", Technical Report No. 28, Center for Language Speech Prcessing, The Johns Hopkins University, Baltimore, MD, USA, Sep. 1998 (http://www.clsp.jhu.edu/people/jmcd/postscript/all-pass.ps).

[5] S. Molau, M. Pitz, R. Schlüter, H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum" Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, Salt Lake City, UT, June 2001, to appear.

[6] F. K. Nocerino, L. R. Rabiner and D. H. Klatt, "Comparative Study of Several Distortion Measures for Speech Recognition", Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, pp. 25-28, Atlanta, GA, Apr. 1985.

[7] A. Sankar, C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol 4, No. 3, May 1996.

[8] L.F. Uebel, P.C. Woodland, "An Investigation into Vocal Tract Length Normalisation", Proc. *6th Europ. Conf. on Speech Communication and Technology*, Vol. 6, pp. 2527-2530, Budapest, Hungary, Sep. 1999.

[9] H. Wakita: "Normalization of Vowels by Vocal Tract Length and its Application to Vowel Identification." *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-25, No. 2, pp. 183-192, April 1977.

[10] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, "Speaker Normalization on Conversational Telephone Speech," Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, Vol. 1, pp. 339-341, Atlanta, GA, May 1996.

[11] L. Welling, S. Kanthak, H. Ney, "Improved Methods for Vocal Tract Normalization," Proc. *Int. Conf. on Acoustic, Speech and Signal Processing*, Vol. 2, pp. 761–764, Phoenix, AZ, April 1999.