Enhancing Vocal Tract Length Normalization with Elastic Registration for Automatic Speech Recognition

Florian Müller and Alfred Mertins

Institute for Signal Processing, University of Lübeck, Lübeck, Germany
{mueller,mertins}@isip.uni-luebeck.de

Abstract

Vocal tract length normalization (VTLN) is commonly applied utterance-wise with a warping function that makes the assumption of a linear dependence between the vocal tract length and the location of the formants. In this work we propose a data-driven method for enhancing the performance of systems that already use standard VTLN. The method is based on elastic registration to estimate optimal non-parametric transformations to further reduce inter-speaker variabilities. Results show that the proposed method can increase the performance of monophone systems such that it reaches that of a triphone system.

Index Terms: automatic speech recognition, vocal tract length normalization, elastic registration

1. Introduction

Speaker-normalization and -adaptation methods are commonly used in speaker-independent automatic speech recognition (ASR) systems to handle inter-speaker variability. While "speaker-adaptation" usually refers to an adaptation of the acoustic model parameters with a maximum-likelihood linear regression (MLLR) approach [1], the term "speaker-normalization" is mostly used in the context of vocal tract length normalization (VTLN) methods [2], which try to compensate for the effects of different vocal tract lengths (VTL) on the feature extraction stage. In it's standard way, this compensation is working on the whole utterance by either warping the frequency centers of the used filter bank or by warping the frequency axis of the output of the filter bank. Assuming a lossless, uniform tube model of length l, the resonance frequencies F_i occur at $F_i = (2i - 1) \cdot c/(4l)$, $i = 1, 2, 3, \ldots$, where c is the speed of sound. This linear scaling of the resonances for different tube lengths is the basis for the often used piecewise-linear warping function as described, for example, in [3]. Different types of other warping functions were analyzed [4], but did not show any significant advances with respect to accuracy compared to piecewise-linear warping.

In this work we propose a method that accounts for two additional factors that are not or only roughly accounted for in the commonly used VTLN approach: Usually, the whole utterance of a single speaker is warped with only a single warping factor. While this approach mitigates the average effect of different VTLs on a per-speaker basis, it does not consider the fact that the VTL of a single speaker changes when producing phonemes where, for example, the lips are lengthened or the larynx is lowered [5]. There are works that follow the idea of using more than one warping parameter for normalizing the time-frequency (TF) representation of an utterance of a single speaker: [6] proposed a region-based VTLN approach where a parameter for a piecewise-linear warping function is estimated for up to five phoneme groups during decoding. A method for a frame-wise warping parameter estimation was proposed by [7], where the Viterbi search space is augmented with a search for an optimal warping parameter and the corresponding decoder is referred to as "MATE decoder". Both methods use a piecewise-linear warping function.

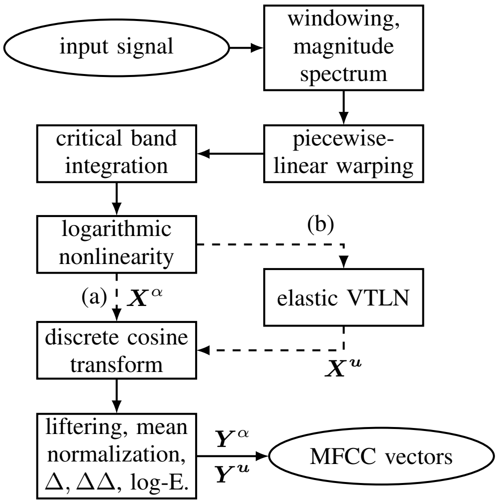

input signal
windowing, magnitude spectrum
piecewise-linear warping
critical band integration
logarithmic nonlinearity
(b)
elastic VTLN
(a)
$\boldsymbol{X}^{\alpha}$
discrete cosine transform
$\boldsymbol{X}^{\boldsymbol{u}}$
liftering, mean normalization, Δ, $\Delta\Delta$, log-E.
$\boldsymbol{Y}^{\alpha}$
$\boldsymbol{Y}^{\boldsymbol{u}}$
MFCC vectors

Figure 1: *Computation of VTL normalized MFCC vectors (a) in its common form denoted as $\boldsymbol{Y}^{\alpha}$, (b) with the proposed enhancement denoted as $\boldsymbol{Y}^{\boldsymbol{u}}$.*

In this work we present a data-driven method for refining the TF representation as output of the commonly used one-parameter VTLN approach. Our proposed method makes use of elastic registration with specific constraints for the task of VTLN for ASR. The resulting warping functions are non-parametric and allow for a high degree of freedom. The next section describes the idea of the proposed method and gives some details about its implementation. Section 3 explains the experiments and analyzes the method with respect to the resulting ASR performance. The paper is concluded in Section 4.

2. VTLN and Elastic Registration

In this work we use mel frequency cepstral coefficients (MFCC). The procedure for the computation of MFCC vectors with an integrated (optional) warping of the frequency axis is illustrated in Figure 1. However, the proposed method can be used with any feature type with an intermediate spectral representation. An often used implementation of VTLN follows the procedures for speaker-adaptive training (SAT) and a two-pass decoding strategy as described in [8]. The method proposed in this work can be regarded as an additional feature enhancement step to standard VTLN.

2.1. Standard Vocal Tract Length Normalization

In this work, we consider a set of global warping factors $\alpha = \{-0.88, -0.9, \ldots, 1.12\}$, where we refer to $\alpha_N = 1$ as the "neutral warping factor" in the following. The SAT procedure with VTLN according to [8] can be summarized as follows: First, let $r = 1, \ldots, R$ be utterance indices. Using the non-normalized observations $\boldsymbol{Y}_r$ an acoustic model $\boldsymbol{\lambda}$ with single Gaussians per state is estimated,

$$\boldsymbol{\lambda} = \arg\max_{\widehat{\boldsymbol{\lambda}}} \prod_{r=1}^{R} p\left(\boldsymbol{Y}_r \,|\, W_r;\, \widehat{\boldsymbol{\lambda}}\right). \tag{1}$$

Second, for each utterance the warping factor $\alpha^{(r)}$ is determined with the model $\boldsymbol{\lambda}$ and the ground-truth transcriptions $\boldsymbol{W}^{(r)}$ in a maximum likelihood sense,

$$\alpha_r = \arg\max_{\alpha} p(\boldsymbol{Y}_r^{\alpha} \,|\, W_r;\, \boldsymbol{\lambda}), \quad r = 1, \ldots, R. \tag{2}$$

As third step, a VTL normalized acoustic model $\boldsymbol{\lambda}'$ is estimated using the normalized observations $\boldsymbol{Y}_r^{\alpha_r}$ for each utterance r,

$$\boldsymbol{\lambda}' = \arg\max_{\widehat{\boldsymbol{\lambda}}} \prod_{r=1}^{R} p\left(\boldsymbol{Y}_r^{\alpha_r} \,|\, W_r;\, \widehat{\boldsymbol{\lambda}}\right). \tag{3}$$

For the recognition of a given observation sequence $\boldsymbol{Y}$ with the SAT acoustic model $\boldsymbol{\lambda}'$, a suboptimal two-pass strategy [8] can be applied as follows: A first decoding pass with non-normalized observations $\boldsymbol{Y}$ and acoustic model $\boldsymbol{\lambda}$ yields a hypothesized transcription $\widetilde{W}$,

$$\widetilde{W} = \arg\max_{W} \{P(W) \cdot p\,(\boldsymbol{Y} \,|\, W;\, \boldsymbol{\lambda})\}. \tag{4}$$

Given the normalized model $\boldsymbol{\lambda}'$ and the hypothesis $\widetilde{W}$, a warping factor $\widetilde{\alpha}$ is selected that yields the highest likelihood,

$$\widetilde{\alpha} = \arg\max_{\alpha} p\left(\boldsymbol{Y}^{\alpha} \,|\, \widetilde{W};\, \boldsymbol{\lambda}'\right). \tag{5}$$

A second decoding pass with normalized observations $\boldsymbol{Y}^{\widetilde{\alpha}}$ and normalized model $\boldsymbol{\lambda}'$ yields the final transcription,

$$\arg\max_{W} \left\{P(W) \cdot p\left(\boldsymbol{Y}^{\widetilde{\alpha}} \,|\, W;\, \boldsymbol{\lambda}'\right)\right\}. \tag{6}$$

2.2. Elastic Vocal Tract Length Normalization

The idea of the VTLN approach that normalizes the frequency axis of the spectrograms as described above can be seen as trying to deform the magnitude spectrum such that the deformed spectrum is more similar to a corresponding spectrum that would have been generated by a speaker associated with a neutral warping factor. Ideally, the deformation is context-dependent and has a high degree of freedom, which allows for the modeling of a wide range of spectral effects due to different VTLs.

Let us assume we have filter bank outputs $\boldsymbol{X}^{\alpha}$ that have been normalized with the VTLN approach as summarized in Section 2.1 and let $\boldsymbol{g} = (g_1, g_2, \ldots, g_G)$ refer to the indices of utterances associated with the neutral warping parameter α_N. Furthermore, let $\boldsymbol{\Lambda}$ be a Gaussian mixture model (GMM) based acoustic model whose parameters have been trained on the normalized outputs $\boldsymbol{X}^{\alpha}$ that are associated with the neutral warping parameter α_N,

$$\boldsymbol{\Lambda} = \arg\max_{\widehat{\boldsymbol{\Lambda}}} \prod_{k=1}^{G} p\left(\boldsymbol{X}_{g_k}^{\alpha_N} \,|\, W_{g_k};\, \widehat{\boldsymbol{\Lambda}}\right). \tag{7}$$

Due to the GMM (here with M Gaussians) the probability density function (PDF) modeled by a single state j of an acoustic model is given by

$$b_j(\boldsymbol{x}_t) = \sum_{m=1}^{M} c^{(jm)} \mathcal{N}\left(\boldsymbol{x}_t;\, \boldsymbol{\mu}^{(jm)},\, \boldsymbol{\Sigma}^{(jm)}\right), \tag{8}$$

where $\boldsymbol{x}_t$ is a single observation vector, $c^{(jm)}$ is a weighting coefficient, and $\mathcal{N}(\,\cdot\,;\, \boldsymbol{\mu},\, \boldsymbol{\Sigma})$ is a multivariate Gaussian PDF with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$\mathcal{N}(\boldsymbol{x};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n\,|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}. \tag{9}$$

Obviously, the likelihood $b_j(\boldsymbol{x}_t)$ in Eq. (8) can be maximized with

$$\boldsymbol{x}^{(j)} = \arg\max_{\widehat{\boldsymbol{x}}_t} b_j(\widehat{\boldsymbol{x}}_t) = \sum_{m=1}^{M} c^{(jm)} \boldsymbol{\mu}^{(jm)}, \tag{10}$$

and it can be seen that the maximum can be determined if the state j is known. Now, let $S_r(\boldsymbol{X}, \boldsymbol{\lambda}', W) = (s_1, s_2, \ldots, s_T)$ denote the state sequence of utterance r that is estimated with forced-alignment based on an observation sequence $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_T\end{bmatrix}$, an acoustic model $\boldsymbol{\lambda}'$, and a given transcription W. The acoustic likelihood for $\boldsymbol{X}$ and S_r given $\boldsymbol{\Lambda}$ is

$$p(\boldsymbol{X}, S_r \,|\, \boldsymbol{\Lambda}) = \prod_{t=1}^{T} b_{s_t}(\boldsymbol{x}_t). \tag{11}$$

Eq. (11) would be maximized with

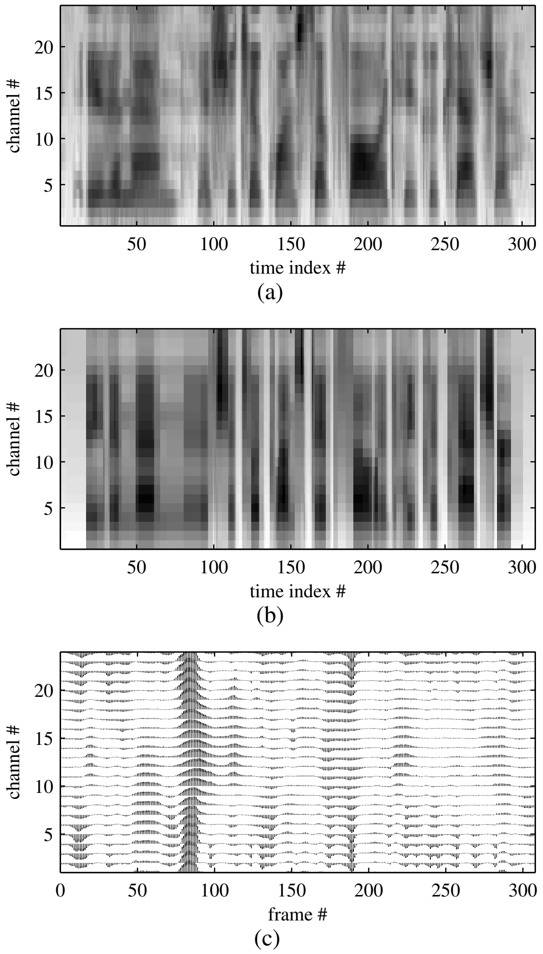
$$\boldsymbol{X}^* = \begin{bmatrix}\boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \ldots & \boldsymbol{x}_T^*\end{bmatrix} \quad \text{where} \quad \boldsymbol{x}_t^* = \boldsymbol{x}^{(s_t)}. \tag{12}$$

Figure 2(a) shows an exemplary filter bank output $\boldsymbol{X}$ of a single utterance. Using a three-state left-to-right monophone model $\boldsymbol{\lambda}'$, a forced-alignment W was estimated, which yields a state-sequence $S(\boldsymbol{X}, \boldsymbol{\lambda}', W)$. The optimal observation sequence $\boldsymbol{X}^*$ according to Eq. (12) is shown in Figure 2(b).

We want to describe the spectral effects due to VTL changes for each frame of a whole utterance. The key idea of the proposed method in this work is to find a transformation such that a transformed observation sequence is similar to its optimal observation sequence. This procedure is called "registration" and is actively researched within the field of image processing. As is described in more detail in the following, the objective function to be optimized contains a term that is based on the linearized elastic potential. Therefore, we refer to the proposed method as "elastic VTLN".

2.2.1. Elastic Registration

Details to the following introduction about the applied registration approach can be found in [9]. In general, the goal of registration can be stated as follows: Given a reference $\boldsymbol{R}$ and template $\boldsymbol{T}$ and a mapping $\boldsymbol{R}, \boldsymbol{T} : \mathbb{R}^2 \to \mathbb{R}$, we want to find a displacement $\boldsymbol{u} : \mathbb{R}^2 \to \mathbb{R}^2$, such that the transformed template $\boldsymbol{T}^{\boldsymbol{u}} := \boldsymbol{T}(x - \boldsymbol{u}(x))$ is similar to $\boldsymbol{R}$. For the computation of $\boldsymbol{T}^{\boldsymbol{u}}$ a linear interpolation scheme is used in this work and the boundaries of $\boldsymbol{T}$ were extended with linear regression. The similarity is quantified with a distance measure $\mathcal{D}[\boldsymbol{R}, \boldsymbol{T}^{\boldsymbol{u}}] : \mathbb{R}^2 \to \mathbb{R}$. By introducing a regularization term $\mathcal{S}[\boldsymbol{u}] : \mathbb{R}^2 \to \mathbb{R}$ prior knowledge can be introduced and



(a)

(b)

(c)

Figure 2: (a) original observation sequence $\boldsymbol{X}$, (b) optimal observation sequence $\boldsymbol{X}^*$, (c) exemplary displacement field $\boldsymbol{u}$.

the numerical solution becomes more stable. The constrained optimization problem then reads

$$\min_{\boldsymbol{u}} \mathcal{D}\left[\boldsymbol{R},\boldsymbol{T}^{\boldsymbol{u}}\right]+\nu\mathcal{S}\left[\boldsymbol{u}\right] \quad \text{subject to} \quad \boldsymbol{u}\in\mathcal{M}, \tag{13}$$

where $\nu \in \mathbb{R}^+$ is a regularization parameter, and $\mathcal{M}$ is a set of admissible transformations. As distance measure $\mathcal{D}$ the correlation-based distance measure [9] is used,

$$\mathcal{D}^{\text{corr}}[\boldsymbol{R},\boldsymbol{T}^{\boldsymbol{u}}]=\left\langle \frac{\boldsymbol{R}-\boldsymbol{\mu}(\boldsymbol{R})}{\boldsymbol{\sigma}(\boldsymbol{R})}, \frac{\boldsymbol{T}^{\boldsymbol{u}}-\boldsymbol{\mu}(\boldsymbol{T}^{\boldsymbol{u}})}{\boldsymbol{\sigma}(\boldsymbol{T}^{\boldsymbol{u}})}\right\rangle_{L_2}, \tag{14}$$

where $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ denote the mean and standard deviation, respectively. The choice for the regularizer in this work can be motivated, e.g., by considering the spectral effects of spatially restricted VTL changes. By means of an articulatory speech synthesis model it is shown in [5] that an elongation at the lips, the larynx, or a mid segment yield a warping of resonance frequencies that is not linear with frequency. In the two-dimensional case the elastic regularizer $\mathcal{S}^{\text{elast}}$ [9] can be seen as a rubber foil that induces tension if deformed. For two dimensions it is defined as

$$\mathcal{S}^{\text{elast}}[\boldsymbol{u}]=\frac{1}{2}\int_{\Omega}\sum_{d=1}^{2}\rho\left\|\nabla\boldsymbol{u}_d\right\|^2+(\rho+\kappa)(\operatorname{div}\boldsymbol{u})^2\,\mathrm{d}x, \tag{15}$$

where $\rho, \kappa \in \mathbb{R}^+$ are the so-called Navier-Lamé constants, which control the elastic behavior of the deformation, ∇ denotes a gradient, and div the divergence operator.

For the optimization of Eq. (13) we use the first-optimize-then-discretize approach. That means, a minimizer of the objective function is determined first that leads to a nonlinear system of partial differential equations (PDE). Then, the PDE is discretized and solved with a fixed-point iteration scheme in this work. There exist efficient algorithms for solving the occurring linear system of equations in each iteration [9]. To constrain the possible solutions with displacements along the subband axis, the displacements that occur along the time axis are set to zero in each iteration of the numerical solution while keeping the displacements along the subband axis. In this work the Navier-Lamé constants were set to $\rho = 1$ and $\kappa = 0$, which is a common choice [9]. As an example, a displacement field for the reference and template signals shown in Figure 2 (b) and (a), respectively, can be seen in Figure 2 (c). The displacements along the subband axis for each component are clearly visible. The chosen regularization parameter yields spatially restricted and smooth displacements.

2.2.2. Using Elastic Registration for VTLN: Elastic VTLN

The standard VTLN approach can be used for SAT, as well as for VTL normalization during recognition. By making use of elastic VTLN, we propose procedures for both cases to enhance the overall performance of the ASR system in the following.

Starting with a SAT acoustic model $\boldsymbol{\lambda}'$, the following method aims to further decrease the effects of inter-speaker-variabilities that result in translations along the subband axis. In a first step, an acoustic model $\boldsymbol{\Lambda}$ is trained only on utterances that are associated with the neutral warping parameter (see Section 2.2). With the ground-truth labels of the training data, a maximum-likelihood (ML) state alignment is computed. For each training observation sequence $\boldsymbol{X}_r$, an optimal observation sequence $\boldsymbol{X}_r^*$ is generated and a displacement field $\boldsymbol{u}_r$ is estimated with $\boldsymbol{X}_r^*$ being the reference and $\boldsymbol{X}_r$ being the template,

$$\boldsymbol{u}_r=\arg\min_{\widehat{\boldsymbol{u}}}\mathcal{D}^{\text{corr}}\left[\boldsymbol{X}_r^*,\boldsymbol{X}_r^{\widehat{\boldsymbol{u}}}\right]+\nu\mathcal{S}^{\text{elast}}\left[\widehat{\boldsymbol{u}}\right]. \tag{16}$$

The application of the displacements for each utterance yields a warped spectral representation $\boldsymbol{X}_r^{\boldsymbol{u}}$. A subsequent computation of cepstral-coefficient based features on the basis of the warped representations (cf. Figure 1) yields the final observations $\boldsymbol{Y}_r^{\boldsymbol{u}}$. These are used for a re-estimation of the acoustic model parameters, which leads to the final acoustic model $\boldsymbol{\lambda}''$,

$$\boldsymbol{\lambda}''=\arg\max_{\widehat{\boldsymbol{\lambda}}}\prod_{r=1}^{R}p\left(\boldsymbol{Y}_r^{\boldsymbol{u}_r}\mid W_r,\widehat{\boldsymbol{\lambda}}\right). \tag{17}$$

Similar to the standard VTLN approach, the decoding of features with elastic VTLN uses the hypothesis $\widetilde{W}$ from a first decoding pass for an ML state-alignment. The output of the state-alignment is used to generate a hypothetically optimal observation sequence $\widetilde{\boldsymbol{X}}^*$ that, in turn, is used as reference for a subsequent elastic registration. The resulting displacement $\boldsymbol{u}$ is then used to compute a deformed spectral representation $\boldsymbol{X}^{\boldsymbol{u}}$. The deformation spectral values are used for the extraction of cepstral-coefficient based features $\boldsymbol{Y}^{\boldsymbol{u}}$. A second decoding pass yields the final transcription.

3. Experiments

The TIMIT corpus with its standard training and test sets (without SA sentences) was used here. The training set consists

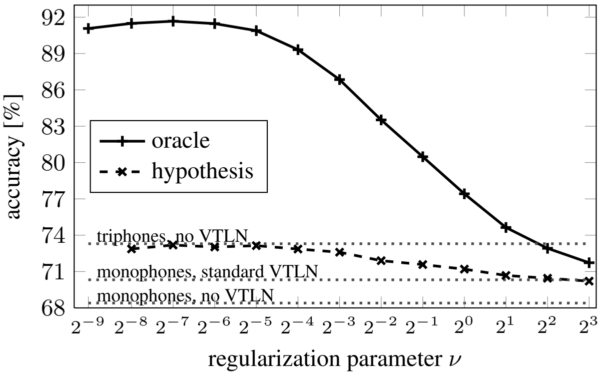


Figure 3: Resulting accuracies using elastic VTLN for oracle (solid) and hypothetical (dashed) transcriptions, as well as baseline accuracies.

of 3696 utterances from 462 different speakers. The test set consists of 1344 utterances from another 168 different speakers. Following the standard procedure for TIMIT, the initial phoneme set was folded to 48 phonemes. Three-state left-to-right monophone models with up to 16 Gaussians and diagonal covariance matrices together with bigram statistics were used. For the computation of the recognition accuracy, the transcriptions were further folded to 39 phonemes. We decided to use monophone models in this work to decrease the computational load, thus, making the analysis of the proposed elastic VTLN method more feasible. The feature extraction follows the procedure as depicted in Figure 1 and yields 39 dimensional MFCC vectors. The baseline accuracy of the system without VTLN is 68.4%, and 70.3% with standard VTLN. As additional baseline, triphone modeling with rule-based state-clustering yields an accuracy of 73.3% without VTLN, and 74.5% with standard VTLN.

In a first step, an upper bound for the accuracy obtained with elastic VTLN was determined. This was done by estimating state alignments based on oracle transcriptions for both the training as well as for the test utterances. Features were computed with the resulting deformations as described in Section 2.2.2 and recognitions experiments were conducted with the monophone system. The accuracies for different choices of the regularization weights ν are shown in Figure 3 as solid line. The impact of a large regularization coefficient is clearly visible: The larger the weight, the smaller are the resulting displacements towards optimal spectral representations. An optimal choice for ν w.r.t. accuracy is given by $\nu = 0.008$. As is described next, this holds for the use of both the oracle as well as the hypothesized transcription.

The potential of elastic VTLN is clearly shown with the accuracy reaching 91.7% with the oracle-transcription based monophone system. However, in practice, a hypothesized transcription from the first decoding pass has to be used for the normalization. To see how elastic VTLN performs under practical conditions, hypothesized transcriptions as output of the standard VTLN approach were used for the computation of the displacement fields in a second experiment. The results are shown in Figure 3 as dashed line. It can be seen that a large regularization weight yields no performance improvements in comparison to standard VTLN. However, when choosing $\nu = 0.008$, the accuracy of the monophone system can be increased by more than four percentage points, reaching the accuracy of the triphone system. At this point it is noteworthy, that the enhanced hypothesis could be used for another elastic VTLN pass, which should further increase the accuracy.

4. Conclusions and Outlook

We presented a method that we refer to as "elastic VTLN" for enhancing the standard VTLN approach. The method is data-driven and makes use of elastic registration with nonparametric deformations as output. Using elastic VTLN, the results show that it is possible to enhance the performance of a monophone system such that it reaches that of a triphone system.

The choice of both the distance measure as well as the regularization method can have a considerable effect on the solution. It is shown in the experimental part that the choices for this work yield promising results. However, additional experiments will have to show if other measures or regularizers are even more beneficial. Another common approach for registration methods is the introduction of an additional penalty term. The objective function used within this work does not account for energy preservation (w.r.t. the spectral values) during the computation of the transformation. An appropriate penalty term could take care for this. We assume that due to the normalization during the subsequent feature extraction in this work, the effect of not considering energy preservation is mitigated. Nevertheless, a subtle analysis might provide further performance improvements. Due to the small size of training and test data provided by the TIMIT corpus, these results can only be seen as preliminary ones and have to be verified on a larger corpus with a more competitive acoustic modeling. A comparison of elastic VTLN with regional VTLN and VTLN with the MATE decoder is also part of future work.

A Matlab implementation of the registration method that was used for the experiments of this work will be available at http://www.isip.uni-luebeck.de/download.

5. Acknowledgements

This work has been supported by the German Research Foundation under Grant No. ME1170/4-1.

6. References

[1] 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.

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

[3] T. Hain, P. C. Woodland, T. R. Niesler, and E. W. D. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," in *Proc. Int. Conf. Audio, Speech, and Signal Processing*, Phoenix, USA, May 1999, pp. 57–60.

[4] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalisation," in *Proc. 6th European Conf. Speech Communication and Technology (EUROSPEECH'99)*, Budapest, Hungary, Sept. 1999, pp. 2527–2530.

[5] S. Mathur, B. Story, and J. Rodriguez, "Vocal-tract modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays," *IEEE Tran. Audio Speech and Language Processing*, vol. 14, no. 5, pp. 1754 – 1762, Sept. 2006.

[6] M. G. Maragakis and A. Potamianos, "Region-based vocal tract length normalization for ASR," in *Proc. Interspeech-2008*, Brisbane, Australia, Sept. 2008, pp. 1365–1368.

[7] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *Proc. Interspeech-2005*, Lisbon, Portugal, Sept. 2005, pp. 3009–2012.

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

[9] J. Modersitzki, *Numerical Methods for Image Registration*. New York: Oxford University Press, 2004.