On the Use of a Wave-Reflection Model for the Estimation of Spectral Effects due to Vocal Tract Length Changes with Application to Automatic Speech Recognition

Florian Müller, Alfred Mertins

Institut for Signal Processing, University of Lübeck, 23562 Lübeck, Email: {mueller, mertins}@isip.uni-luebeck.de

Abstract

Vocal tract length normalization (VTLN) is commonly used in state-of-the-art automatic speech recognition (ASR) systems to reduce the mismatch between speaker-dependent formant frequency scalings. Usually, the normalization is done by a piece-wise linear scaling of the filter bank center frequencies. The linear scaling is motivated by a uniform acoustic tube model that does not take any loss effects into account. Furthermore, it is known that a change in vocal tract length (VTL) yields different spectral effects for different phonemes. However, these phoneme-dependent differences are usually not explicitly considered in the common VTLN processing. In this work, we consider a vocal tract model that has been developed within the field of articulatory speech synthesis. The model mimics the vocal tract geometry for different phonemes and simulates naturally occurring loss effects like yielding wall vibrations and viscous losses. An elastic registration method is used to determine the relating transforms between the spectral envelopes of vocal tracts with different lengths. The resulting warping functions are analyzed w.r.t. their application for VTLN in ASR systems.

Introduction

Vocal tract length normalization (VTLN) [7] in current automatic speech recognition (ASR) systems for the compensation of the effects of different vocal tract lengths (VTL) is based on a lossless, uniform tube model [1]. With a length l, such a model has resonances at frequencies $F_i = c/(4l)$, $i = 1, 2, 3$, where c is the speed of sound. To normalize this linear scaling effect while maintaining the original frequency range a piecewise-linear warping function [7] is commonly used in practice.

This work presents a model-driven method to describe the spectral effects due to different VTLs. A wave-reflection model [6] is used that simulates loss effects due to wall vibrations, viscous fluid losses, heat conduction losses, and kinetic pressure losses. As is described in more detail later the spectral effects are described in terms of a displacement field, which is estimated with an elastic registration approach [4].

Synthesis with a Wave-Reflection Model

The wave-reflection model used in this work is based on the Kelly-Lochbaum model [2]. It models a one-dimensional sound wave propagation within a vocal tract that is modeled as a sequence of short tubes. The choice of diameters for the individual tubes defines the resonance frequencies of the tract model. For the synthesis of vowels the diameters can be set with area functions [6] that are formulated in terms of a neutral diameter function and proportional amounts of basis functions Φ_1, Φ_2. The model simulates loss effects like wall vibrations, viscous fluid losses, and heat conduction losses based on lumped element circuit approximations [5]. Besides more sophisticated methods for changing the simulated VTL of the model [3], the choice of different numbers of concatenated tube elements already allows for discrete VTL changes. It has been shown that this model is able to produce natural sounding utterances for speech synthesis.

Estimating the Effects of Vocal Tract Length Changes

Details to the following introduction about the applied registration approach can be found in [4]. In general, the goal of registration can be stated as follows: Given a reference $\boldsymbol{R}$ and template $\boldsymbol{T}$ and a mapping $\boldsymbol{R}, \boldsymbol{T} : \mathbb{R}^q \to \mathbb{R}$, where q denotes the dimensionality, we want to find a displacement $\boldsymbol{u} : \mathbb{R}^q \to \mathbb{R}^q$, 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 $\mathcal{D}[\boldsymbol{R}, \boldsymbol{T}^{\boldsymbol{u}}] : \mathbb{R}^q \to \mathbb{R}$. By introducing a regularization term $\mathcal{S}[\boldsymbol{u}] : \mathbb{R}^q \to \mathbb{R}$ prior knowledge can be introduced and the numerical solution becomes more stable. The constrained optimization problem then reads

$$\min_{\boldsymbol{u}} \mathcal{D}[\boldsymbol{R}, \boldsymbol{T}^{\boldsymbol{u}}] + \nu \mathcal{S}[\boldsymbol{u}] \quad \text{subject to} \quad \boldsymbol{u} \in \mathcal{M}, \tag{1}$$

where $\nu \in \mathbb{R}^+$ is a regularization parameter, and $\mathcal{M}$ is a set of admissible transformations. In this work, the so-called sum-of-squared differences (SSD) is used as distance measure,

$$\mathcal{D}^{\text{SSD}}[\boldsymbol{R}, \boldsymbol{T}^{\boldsymbol{u}}] := \int_{\Omega} \left(\boldsymbol{T}^{\boldsymbol{u}}(x) - \boldsymbol{R}(x)\right)^2 \, \mathrm{d}x. \tag{2}$$

For the regularization the elastic regularizer $\mathcal{S}^{\text{elast}}[\boldsymbol{u}]$ according to [4] is used,

$$\mathcal{S}^{\text{elast}}[\boldsymbol{u}] = \frac{1}{2} \int_{\Omega} \sum_{d=1}^{2} \rho \|\nabla \boldsymbol{u}_d\|^2 + (\rho + \kappa)(\operatorname{div} \boldsymbol{u})^2 \, \mathrm{d}x, \tag{3}$$

where the Navier-Lamé constants were set to $\kappa = 0$ and $\rho = 1$, which is a common choice. For the numerical solution the first-optimize-then-discretize approach

was chosen and the resulting partial differential equations were solved with a time-marching approach according to [4].

Estimating the Spectral Effects and Recognition Experiments

The estimation of the warping functions is split into three stages: First, vowel sounds are synthesized with the tract model. Nine VTLs linearly located within the range from 14.3 cm to 20.6 cm are simulated. For each of the ten vowel configurations from [6] 121 variations are generated with slight differences in the corresponding weighting coefficients for Φ_1, Φ_2. For each tract configuration a 500 ms speech signal with a sampling rate of 16 kHz is generated and passed to a gammatone filter bank with 110 channels. In the second step displacements $\boldsymbol{u}$ for each utterance were computed with the elastic registration approach as described above. The mean logarithmized filter bank outputs resulting from each tract configuration are used as template $\boldsymbol{T}$ and the corresponding model with "neutral" VTL of 17.5 cm is used as reference $\boldsymbol{R}$. Generally, the choice of the regularization parameter ν is task dependent and, thus, different values for this parameter are considered in the experiments, $\nu \in [0.004, 128]$. The results of the second step are deformations for each considered tract configuration. To allow for the generation of one-parameter warping functions a principal component analysis (PCA) is used and the first PCA component $\boldsymbol{v}$ is then used for generating a set of N warping functions $\mathbf{W}$,

$$\mathbf{W} = \{\, \alpha_k \boldsymbol{v} \,|\, k = 1, 2, \ldots, N \,\}, \tag{4}$$

where α_k is chosen such that an appropriate warping range is covered. The warping functions of $\mathbf{W}$ are used for speaker-adaptive training (SAT) and for the a maximum-likelihood grid-search VTLN [7] during the decoding stage.

Phoneme recognition experiments were conducted on the TIMIT corpus. Following the standard procedure, the SA sentences were not considered and the phoneme set was folded from 48 to 39 phonemes for accuracy evaluation. Three-state left-to-right monophone models with diagonal covariances and a bigram language model were used. Features were extracted with a 110-channel gammatone filter bank. Twelve cepstral coefficients were computed from the filter bank output together with log-energy and $\Delta, \Delta\Delta$-features.

The results of the phoneme recognition experiments are shown in Table 1. Baselines are shown in the upper two rows of the table and illustrate the benefits of VTLN. The remaining rows show accuracies for different choices of the regularizer weight ν. It can be seen that the highest accuracy for these experiments is obtained with a choice of $\nu = 8$. The corresponding set of warping functions is shown in Figure 1.

Conclusion and Outlook

In this work we have shown a model-driven approach for estimating warping functions for VTLN. For this, we made use of a lossy wave-refelction model originally developed for speech synthesis and of elastic registration. Phoneme recognition results show that the proposed methods can be used for the estimation of warping functions and a slight improvement in accuracy compared the the standard VTLN approach could be observed.

Future work will investigate the use of phoneme dependent warping functions. Furthermore, other choices of distance measures or regularizers could also yield increases in accuracy.

Table 1: Results of the phoneme recognition experiments.

warping function	accuracy [%]
-	66.4
(piecewise-)linear	68.0
elastic, $\nu = 0.004$	67.7
elastic, $\nu = 0.5$	68.0
elastic, $\nu = 2$	68.2
elastic, $\nu = 8$	**68.3**
elastic, $\nu = 32$	68.1
elastic, $\nu = 128$	68.0

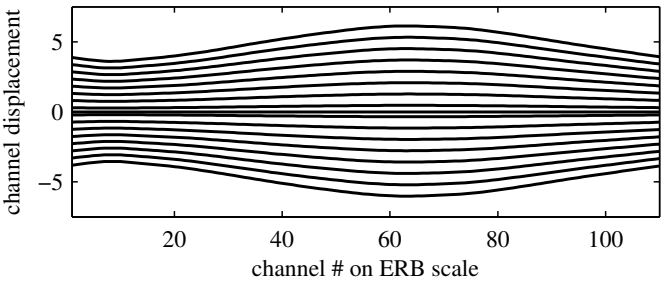

Figure 1: Set of resulting warping functions with $\nu = 8$.

Acknowledgement

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

References

[1] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, New York, 1993.

[2] J. Kelly and C. Lochbaum. Speech synthesis. In *Proc. Speech Communications Seminar*, Stockholm, Sweden, 1963. Royal Institute of Technology.

[3] 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*, 14(5):1754 – 1762, Sept. 2006.

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

[5] B. H. Story. *Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract*. PhD thesis, University of Iowa, 1995.

[6] B. H. Story. A parametric model of the vocal tract area function for vowel and consonant simulation. *J. Acoust. Soc. Am.*, 5(5):3231–3254, May 2005.

[7] L. Welling, H. Ney, and S. Kanthak. Speaker adaptive modeling by vocal tract normalization. *IEEE Trans. Speech and Audio Processing*, 10(6):415–426, Sept. 2002.