Breaking a Paradox: Applying VTLN to Residuals

David Sündermann
Universitat Politècnica de Catalunya
08034 Barcelona, Spain
david@suendermann.com

Harald Höge
Siemens Corporate Technology
81739 Munich, Germany
harald.hoege@siemens.com

Tim Fingscheidt
TU Braunschweig
38106 Braunschweig, Germany
t.fingscheidt@tu-braunschweig.de

Abstract

Vocal tract length normalization (VTLN) is a speaker normalization technique that tries to compensate for the effect of speaker-dependent vocal tract lengths. According to the source-filter model of speech production, the vocal tract can be represented by features as linear predictive or cepstral coefficients, whereas the residual is an estimate of the underlying excitation, i.e. the voice source. In speech recognition, VTLN is performed before deriving the features. Interestingly, recent studies on VTLN-based voice conversion suggest that the application of VTLN to the residuals, that should hardly contain vocal tract information, may result in a voice that is very similar to a voice generated in the conventional way, i.e., by applying VTLN to the original speech frames. In this paper, we discuss an objective experiment that is to confirm this observation.

1 Introduction

In speech recognition, vocal tract length normalization is a standard technique for speaker normalization [1]. It compensates for speaker-dependent vocal tract lengths by warping the frequency axis of the magnitude spectrum of the speech frames before deriving the features. Here, the latter are mostly mel frequency cepstral coefficients, whereas in other disciplines of speech processing as speech coding or synthesis, features types as linear predictive coefficients or line spectral frequencies are widely used. According to the source-filter model of speech production [2], all these features are supposed to represent the vocal tract, whereas the prediction error, referred to as *residual*, is an estimate of the voice source, i.e. the excitation. Consequently, VTLN is supposed to influence the vocal tract features rather than the residual which is usually ignored in speech recognition.

Apart from its use in speech recognition, VTLN can be applied to voice conversion where a source voice is manipulated according to certain criteria (e.g. to mimic a pre-defined target speaker or to change voice characteristics as gender or age) [3].

Our recent studies on residual prediction [4] and text-independent voice conversion [5] suggested the application of VTLN to the speech residuals before inverse filtering using the vocal tract features as filter coefficients.

Interestingly, we observed that often the order of performing VTLN and inverse filtering can be changed resulting in very similar voices. This seems to be a paradox since VTLN aims at warping the frequency axis of the speech frame to change formant frequencies and bandwidths that highly depend on the vocal tract length. The formant information, however, should be contained in the features, and the residuals should ideally be a white noise without characteristical formant structure.

This paper reports about an objective comparison of performing VTLN before and after inverse filtering and shows that the contribution of the residual to the spectral warping is considerable and dominates particularily for small vocal tract length changes.

2 Choosing a Warping Function

In speech recognition literature, several functions that describe the warping of the frequency axis when performing VTLN are proposed [6]. Our experience shows that they behave similarly when used for voice conversion [3]. Furthermore, we found that particularily the piece-wise linear warping function has the advantage that it may be directly applied in time domain without the detour through the frequency domain and back [7]. In the present investigation, we use a special case of the latter with one segment. The strength of the warping, i.e. the strength of the vocal tract length change, is influenced by the *warping factor*, in the following referred to as α.

3 Experiments

To objectively demonstrate the aformentioned paradoxical perceptive effect, we compare the magnitude spectra of the speech after applying VTLN to the speech frames with that after applying VTLN to the residuals and inverse filtering with the unchanged linear predictive coefficients. For this purpose, we use the sum of the Euclidean distances between the aforementioned spectra, in the following referred to as D_r.

To have a standard of comparsion, we furthermore measure the sum of the distances between the spectra of the original speech and those of the warped speech; this results in the distance D_w. Now, we introduce the quotient $e = \frac{D_r}{D_w}$, a measure that becomes smaller, the more similar the spectra of the two compared VTLN application variants are with respect to the similarity of warped and unwarped spectra. I.e., small values of e would confirm our thesis.

3.1 The Experimental Corpus

As test database, we took a corpus that was used in the first evaluation campaign of the speech-to-speech translation project TC-Star [8]: A Spanish speech corpus consisting of 50 phonetically balanced utterances of four speakers (two male and two female).

3.2 The Experimental Setup

Since the error measure e introduced above seemed to strongly depend on the warping

e	$\alpha = 0.95$	$\alpha = 1.05$
male1	0.41	0.29
male2	0,34	0,22
female1	0,47	0,28
female2	0,63	0,40

Table 1: e scores of different speakers

factor α, we carried out experiments with different warping factors and speakers.

We know from speech recognition literature [9] that $0.88 \leq \alpha \leq 1.12$ is a reasonable region of the warping factor for the application of VTLN to speaker normalization. There, every voice is to be transformed to an average voice.

When applying VTLN to voice conversion, we sometimes want to convert voices from one extreme to the other. Consequently, the warping factors to map both extremes can vary about the double of those which are expected for voice normalization, i.e., we investigate the warping factors $0.76 \leq \alpha \leq 1.24$. As an example, in Figure 1, we display the dependence of the e score on the warping factor for one of the male voices.

Furthermore, Table 1 shows results for different speakers and decreasing and increasing the vocal tract length. The warping factors are examples taken from the voice conversion experiments in the first TC-Star evaluation campaign [8].

3.3 Interpretation

Figure 1 shows that the smaller the vocal tract change is, the more effective works the residual VTLN compared with the VTLN applied to the speech signal. The examples of Table 1 result in e scores of about one third, i.e., the distance between both techniques is about one third of that

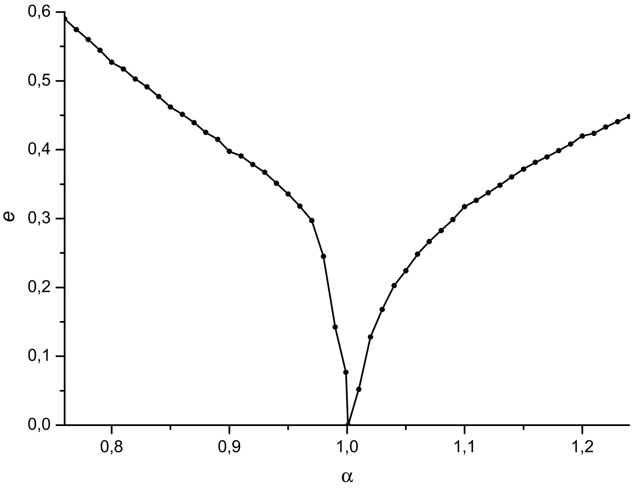

Figure 1: Dependence of the e score on the warping factor α

between warped and unwarped speech. Consequently, the resulting speech of both techniques seems to be rather similar. This outcome agrees with our perceptive experience and hence confirms this paper's thesis.

As another interesting result, we clearly see that the results for male are better than those of female voices and the results for increasing the vocal tract length are better than those for decreasing (femalizing). This observation confirms that voice conversion for female voices is harder than that for male [5].

4 References

[1] D. Pye and P. C. Woodland, "Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition," in *Proc. of the ICASSP'97*, Munich, Germany, 1997.

[2] A. Acero, "Source-Filter Models for Time-

Scale Pitch-Scale Modification of Speech," in *Proc. of the ICASSP'98*, Seattle, USA, 1998.

[3] D. Sündermann, H. Ney, and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in *Proc. of the ASRU'03*, Virgin Islands, USA, 2003.

[4] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion," in *Proc. of the ICASSP'05*, Philadelphia, USA, 2005.

[5] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-Independent Voice Conversion Based on Unit Selection," in *Proc. of the ICASSP'06*, Toulouse, France, 2006.

[6] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, 2005.

[7] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "Time Domain Vocal Tract Length Normalization," in *Proc. of the ISSPIT'04*, Rome, Italy, 2004.

[8] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. v. d. Heuvel, H.-U. Hain, and X. S. Wang, "TTS: Specifications of LR, Systems' Evaluation and Modules Protocols," in *Submitted to the LREC'06*, Genoa, Italy, 2006.

[9] L. F. Uebel and P. C. Woodland, "An Investigation into Vocal Tract Length Normalization," in *Proc. of the Eurospeech'99*, Budapest, Hungary, 1999.