TIME DOMAIN VOCAL TRACT LENGTH NORMALIZATION

David Sündermann, Antonio Bonafonte

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
C/ Jordi Girona, 1 i 3, 08034 Barcelona, Spain
{suendermann,antonio}@gps.tsc.upc.es

Hermann Ney

RWTH Aachen – University of Technology
Computer Science Department
Ahornstr. 55, 52056 Aachen, Germany
ney@cs.rwth-aachen.de

Harald Höge

Siemens AG
Corporate Technology
Otto-Hahn-Ring 6, 81739 Munich, Germany
harald.hoege@siemens.com

ABSTRACT

Recently, the speaker normalization technique VTLN (vocal tract length normalization), known from speech recognition, was applied to voice conversion. So far, VTLN has been performed in frequency domain. However, to accelerate the conversion process, it is helpful to apply VTLN directly to the time frames of a speech signal. In this paper, we propose a technique which directly manipulates the time signal. By means of subjective tests, it is shown that the performance of voice conversion techniques based on frequency domain and time domain VTLN are equivalent in terms of speech quality, while the latter requires about 20 times less processing time.

1. INTRODUCTION

Vocal tract length normalization (VTLN) [1] tries to compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis of the phase and magnitude spectrum. In speech recognition, VTLN aims at the normalization of a speaker's voice to remove individual speaker characteristics and, thus, improve the recognition performance [2].

The same technique can be used for voice conversion [3], which is the modification of a source speaker's voice in order to sound like another speaker [4]. For instance, voice conversion is applied to speech synthesis systems to change the identity of the system's standard speaker in a fast and comfortable way. Here, the process is not a normalization (mapping of several speakers to a certain individual) but the other direction (transforming a standard speaker to several well-distinguishable individuals). This consideration led to the term *reverse VTLN* when referring to the usage as voice conversion technique [5]. To simplify matters, in the following, we continue to utilize *VTLN* in connection with voice conversion.

In speech recognition, most parts of the signal processing are performed in frequency domain. Hence, VTLN is applied to the frequency spectrum, cf. Section 2. In the following, we will refer to this technique as *FD-VTLN* (frequency domain VTLN).

In contrast to speech recognition, concatenative speech synthesis predominantly operates in time domain. For instance, the concatenation of speech segments and the prosodical manipulation (intonation, speaking rate, etc.) are often based on TD-PSOLA (time domain pitch-synchronous overlap and add) [6]. The application of FD-VTLN to speech synthesis requires the transformation from time to frequency domain and the other way around using DFT (discrete Fourier transformation) and inverse DFT, respectively.

However, when a speech synthesis system is to be used inside an embedded environment, each negligible operation must be avoided due to very limited processing resources [7]. This is the motivation why VTLN should be directly applied to the time frames of a signal processed by a speech synthesizer before being concatenated and prosodically manipulated by means of TD-PSOLA. In the following, we refer to this technique as *TD-VTLN* (time domain VTLN). In Section 3, we describe how TD-VTLN can be derived from FD-VTLN and address a computing time comparison between both techniques.

The equivalence of FD-VTLN and TD-VTLN in terms of voice conversion performance (speech quality and success of the voice identity conversion) is investigated with the help of subjective tests in Section 4.

2. FREQUENCY DOMAIN VTLN

2.1. Preprocessing

Since the advantages of pitch-synchronous speech modification and analysis are well-studied, this approach has been also successfully applied to voice conversion [8].

To extract pitch-synchronous frames from a given speech signal, we use the algorithm described in [9]. In voiced regions, the frame lengths depend on the fundamental frequency, in unvoiced regions, the pitch extraction algorithm utilizes a mean approximation.

By applying DFT without zero padding to the frames, we obtain complex-valued spectra with distinct numbers of spectral lines. In the following, these spectra are referred to as X.

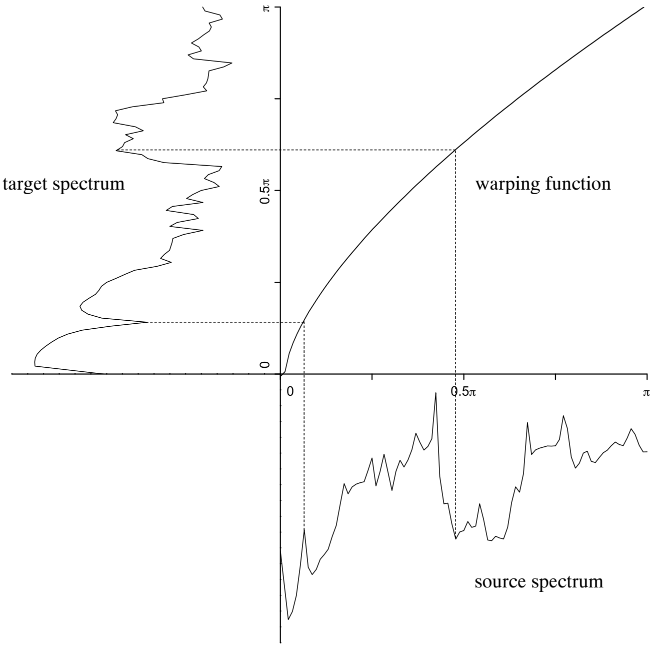

Fig. 1. Warping the magnitude spectrum: an example

2.2. Warping Functions

The realization that the warping of the frequency axis of the magnitude spectrum can lead to a considerable speech recognition performance gain yielded several more or less well-studied warping functions. They can be distinguished regarding the number of parameters describing the particular function and their linearity or nonlinearity, respectively. In Table 1, we show a categorization of the warping functions used in literature.

In general, a warping function is defined as $\tilde{\omega}(\omega|\xi_1,\xi_2,\ldots); 0 \le \omega,\tilde{\omega} \le \pi$, where $\xi_1,\xi_2,\ldots$ are the *warping parameters* and ω is the normalized frequency with π corresponding to half the sampling frequency according to the Nyquist criterion. In Figure 1, we show an example source spectrum, a warping function and the resulting target spectrum.

3. TIME DOMAIN VTLN

3.1. Choosing a Warping Function

When we apply VTLN to voice conversion, it does not play an important role which particular warping function is used since they result in very similar spectra [3]. Hence, the converted speech of different warping functions is hardly perceptually distinguishable. At least, this statement is true for remaining in the same row of Table 1. The effect of increasing the number of warping parameters on the quality and capability of VTLN-based voice conversion has not yet been adequately tested.

parameters	linear	nonlinear
one	• piece-wise linear with two segments – asymmetric [11] – symmetric [13]	• bilinear [10] • power [1] • quadratic [12]
several	• piece-wise linear with several segments [3]	• allpass transform [14]

Table 1. Categorization of VTLN warping functions

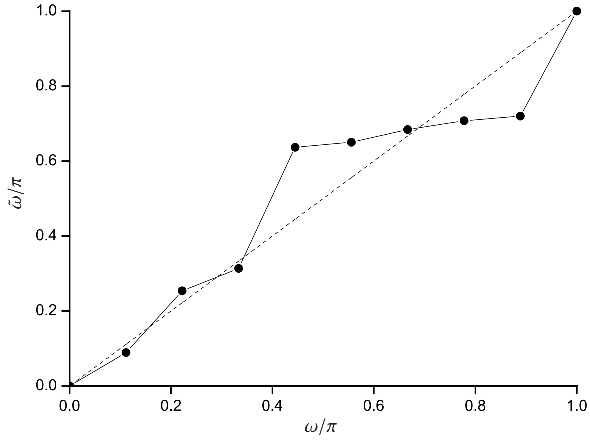

Fig. 2. Example of a piece-wise linear warping function

In the following, we limit our considerations to the piece-wise linear warping function with several segments that includes the two-segment function as a special case, cf. [3]:

$$\tilde{\omega}(\omega|\omega_1^I,\tilde{\omega}_1^I) = \alpha_i\omega+\beta_i \quad \text{for} \quad \omega_i \le \omega < \omega_{i+1}; i = 0\ldots I \qquad (1)$$

$$\text{with} \quad \alpha_i = \frac{\tilde{\omega}_{i+1}-\tilde{\omega}_i}{\omega_{i+1}-\omega_i}, \quad \beta_i = \tilde{\omega}_{i+1} - \alpha_i\omega_{i+1} \quad \text{and}$$

$$0 = \omega_0 < \omega_1 < \cdots < \omega_I < \omega_{I+1} = \pi, \text{ for } \tilde{\omega}_i \text{ equivalent.}$$

An example of this monotonous and bounded function is displayed in Figure 2.

Now, we have a look at the warped spectrum $\tilde{X}$ derived from X applying the warping function $\tilde{\omega}$. The following relation holds:

$$\tilde{X}(\tilde{\omega}(\omega)) = X(\omega) \quad \Longrightarrow \quad \tilde{X}(\omega) = X(\tilde{\omega}^{-1}(\omega))\,. \qquad (2)$$

Then, we determine the inverse function of $\tilde{\omega}$, cf. (1):

$$\tilde{\omega}^{-1}(\omega) = \frac{\omega-\beta_i}{\alpha_i}$$

$$\frac{\omega_i-\beta_i}{\alpha_i} \le \omega < \frac{\omega_{i+1}-\beta_i}{\alpha_i}; \quad i = 0\ldots I\,.$$

This equation can be rewritten as

$$\tilde{\omega}^{-1}(\omega) = \sum_{i=0}^{I} \frac{\omega-\beta_i}{\alpha_i} R(\alpha_i\omega+\beta_i|\omega_i,\omega_{i+1}) \qquad (3)$$

$$\text{with} \quad R(\omega|\omega',\omega'') = \operatorname{rect}\left[\frac{\omega-\frac{\omega''+\omega'}{2}}{\omega''-\omega'}\right]; \; \omega' < \omega''$$

$$= \begin{cases} 1 & : \quad \omega' < \omega < \omega'' \\ 0.5 & : \quad \omega = \omega' \vee \omega = \omega'' \\ 0 & : \quad \text{otherwise} \end{cases}$$

Inserting (3) into (2) yields

$$\tilde{X}(\omega) = \sum_{i=0}^{I} X\left(\frac{\omega-\beta_i}{\alpha_i}\right) R(\alpha_i\omega+\beta_i|\omega_i,\omega_{i+1})\,. \qquad (4)$$

3.2. What Happens in Time Domain?

To answer this question, we exploit several properties of the DFT, in particular, the scaling, shifting, convolution, and linearity rule.

To begin with, we describe the first term in (4) in time domain. Here, we use the time signal x from which the spectrum X was computed by applying DFT:

$$X(\omega) = \mathcal{F}\{x(t)\}(\omega) \quad \Longrightarrow$$

$$u(t) := \mathcal{F}^{-1}\left\{X\left(\frac{\omega-\beta}{\alpha}\right)\right\}(t) = \alpha e^{i\beta t}x(\alpha t)\ . \qquad (5)$$

The second term is transformed by means of the time correspondance of the frequency domain rect function, cf. (3),

$$\mathcal{F}^{-1}\{\mathrm{rect}\,(\omega)\}(t) = \frac{T}{\pi t}\sin\left(\frac{t}{2T}\right)\ , \qquad (6)$$

where T is the time frame length. By again utilizing the scaling and shifting rules, we obtain a rather complicated term for the time correspondance $r(t)$ derived from $R(\alpha\omega+\beta)$.

In time domain, a multiplication between two spectral functions corresponds to a convolution, thus we can connect the results from (5) and (6) according to (4) as follows

$$\mathcal{F}^{-1}\left\{X\left(\frac{\omega-\beta}{\alpha}\right)\cdot R(\alpha\omega+\beta)\right\}(t) = (u * r)(t)\ . \quad (7)$$

Finally, two points have to be taken into account:

• As (5) suggests, we still have a complex-valued time signal after the convolution. This is due to the DFT that delivers a symmetric magnitude and phase spectrum in the range $-\pi \leq \omega \leq \pi$. So far, we only have considered the positive part of the frequency axis, but the steps described in this section must also be applied to the negative part. Both contributions obtained from (7) are then summed and result in a real-valued time signal.

• Since we deal with discrete time signals, the scaling $x(\alpha t)$, cf. (5), is carried out by means of cubic spline interpolation with a certain number of interpolation points according to the current frame length. Depending on the warping parameter α, this interpolation would either cover only a part of the original time frame ($\alpha < 1$) or stretch across a span of time that is greater than the original frame ($\alpha > 1$). A straight-forward summation of the time signals obtained through (7), suggested by applying the DFT linearity rule to (4), would lead to time signals with unreasonable properties as signal jumps. Therefore, we propose extending the scaling to cover the complete original frame, that, consequently, yields scaled frames of different lengths. Finally, these frames are joined within the TD-PSOLA processing step, which has to be carried out $I+1$ times.

3.3. On the Computational Complexity of TD-VTLN

In Section 1, we have argued, that we want to use TD-VTLN for accelerating the conversion process. However, through the usage of the convolution operation which has a complexity order of $O(T^2)$, we will not essentially reduce the computing time[1]. Only the special case $I = 0$, i.e. the linear warping function with exactly one segment can essentially speed up the computation: In (4), the summation is omitted and we have $\beta = 0$. Furthermore, $R(\alpha\omega)$ becomes 1.0 for the complete considered spectrum, thus, the convolution (7) need not be performed and we obtain the warped time frame

$$\tilde{x}(t) = u(t) = \alpha x(\alpha t)\ .$$

	FD-VTLN	TD-VTLN
DFT	$4T^2 - 2T$	–
spline interpolation	$40T$	$40T$
IDFT	$4T^2 - 2T$	–
PSOLA	$4T$	$4T$
total	$8T^2 + 40T$	$44T$

Table 2. FD vs. TD-VTLN: breakdown of operations ($I = 0$)

Table 2 shows a comparison between FD and TD-VTLN with respect to the required operations. When we take the average frame lengths from the experimental corpus described in Section 4.1, $T_f = 101$ and $T_m = 140$ for the female and the male speaker, respectively, we obtain an acceleration by a factor of about 19 for the female and about 26 for the male speaker replacing FD-VTLN by TD-VTLN.

4. EXPERIMENTS

In the last section, we have shown that for the special case $I = 0$ the computing time can be essentially reduced. However, we have to control, if this simplification affects the conversion quality compared to the standard case $I = 1$, cf. [3] and [5]. This section addresses the subjective evaluation of the presented TD-VTLN technique.

4.1. The Corpus

The corpus utilized in this work contains several hundred Spanish sentences uttered by a female and a male speaker. The speech signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16 kHz.

4.2. Defining the Warping Parameters

As mentioned in Section 1, in speech synthesis, VTLN is used to create new voices that are sufficiently distinguishable from the original. To investigate this effect, we estimate the warping parameters in the way that the converted spectra that stem from speech of a source speaker maximally approach the corresponding spectra of a target speaker's speech. To obtain these corresponding spectra, we apply dynamic time warping to the speech signals based on equivalent utterances of both speakers (text-dependent approach). The cost function, which is to be minimized, is derived from the objective error criterion described in [15] and leads to the following equation:

$$\begin{aligned}\alpha &= \arg\min_{\alpha'}\sum_{n=1}^{N} w_n d(\tilde{X}_n(\alpha'), Y_n)\\ &\approx \sum_{n=1}^{N} w_n \arg\min_{\alpha'} d(\tilde{X}_n(\alpha'), Y_n)\end{aligned}$$

[1] The computational complexity order of the DFT algorithm is $O(T^2)$ as well.

	FD-VTLN	TD-VTLN	total
source speaker	20%	16%	18%
target speaker	29%	36%	32%
neither	50%	48%	49%

Table 3. Results of the extended ABX test

	FD-VTLN	TD-VTLN	total
female-male	3.3	3.4	3.3
male-female	2.6	2.6	2.6
total	3.0	3.0	

Table 4. Results of the MOS test

$$\text{with} \quad w_n = \frac{\sqrt{E(X_n)E(Y_n)}}{\sum\limits_{\nu=1}^{N} \sqrt{E(X_\nu)E(Y_\nu)}}$$

$$\text{and} \quad d(X,Y) = E\left[\frac{X}{\sqrt{E(X)}} - \frac{Y}{\sqrt{E(Y)}} \right] .$$

Here, N is the number of training frames and $E(X)$ is the signal energy of the spectrum X.

4.3. Subjective Evaluation

By means of the method described in the last section, we determined the warping parameter α for the two gender combinations utilizing 10 training sentences. Then, we applied both FD and TD-VTLN and both gender combinations to 8 sentences of the corpus, obtaining a total of 32 converted sentences. From these, 8 sentences were randomly selected in the way that each gender-VTLN combination was represented by exactly two sentences. This randomization was carried out again for each of the 14 participants, 12 of whom were specialists in speech processing.

At first, the participants were asked if the converted voice sounds similar to the source or to the target voice or to neither of them (extended ABX test). This was to control the capability of VTLN-based voice conversion to generate new voices. Furthermore, they were asked to assess the overall sound quality of the converted speech on a mean opinion score (MOS) scale between 1 (very bad) and 5 (very good). Table 3 reports the results of the extended ABX test and Table 4 those of the MOS rating depending on the VTLN technique and the gender combination.

4.4. Interpretation

The outcomes of the subjective tests discussed in the last section can be interpreted as follows:

• VTLN-based voice conversion features the capability to manipulate a given voice in such a way that the result is sufficiently different from the original to be perceived as another voice: Only 18% of the example sentences were recognized as spoken by the source speaker, cf. Table 3.

• On the other side, VTLN-based voice conversion is not appropriate to imitate a certain speaker's voice: Table 3 reports that only 32% of the examples were perceived to be uttered by the target speaker whose voice characteristics led to the warping parameter α, cf. Section 4.2.

• As Table 4 shows, the overall sound quality of the two compared techniques FD and TD-VTLN is equivalent, although the former is based on the conventional piece-wise linear warping function with two segments [13] (I=1), whereas the latter uses the special case introduced in Section 3.3 (I=0). The average MOS corresponds to that reported in the literature dealing with VTLN-based voice conversion, cf. [5].

• At least for the corpus our tests were based on, the conversion from a male to a female voice resulted in an essentially worse MOS than the other direction, cf. Table 4. This result confirms the objective error measures reported in [3].

5. CONCLUSION

This paper addresses the transformation of the spectral warping as a part of FD-VTLN to the time domain. We refer to this technique as time domain vocal tract length normalization (TD-VTLN). When we apply TD-VTLN to voice conversion, the computational costs can be reduced by a factor of about 20 while keeping the sound quality and the ability of voice identity conversion.

6. REFERENCES

[1] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *Proc. of the ICASSP'96*, Atlanta, USA, 1996.

[2] 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.

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

[4] E. Moulines and Y. Sagisaka, "Voice Conversion: State of the Art and Perspectives," *Speech Communication*, vol. 16, no. 2, 1995.

[5] M. Eichner, M. Wolff, and R. Hoffmann, "Voice Characteristics Conversion for TTS Using Reverse VTLN," in *Proc. of the ICASSP'04*, Montreal, Canada, 2004.

[6] F. J. Charpentier and M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," in *Proc. of the ICASSP'86*, Tokyo, Japan, 1986.

[7] A. W. Black and K. A. Lenzo, "Flite: A Small Fast Run-Time Synthesis Engine," in *Proc. of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, UK, 2001.

[8] A. Kain and M. W. Macon, "Spectral Voice Transformations for Text-to-Speech Synthesis," in *Proc. of the ICASSP'98*, Seattle, USA, 1998.

[9] V. Goncharoff and P. Gries, "An Algorithm for Accurately Marking Pitch Pulses in Speech Signals," in *Proc. of the SIP'98*, Las Vegas, USA, 1998.

[10] A. Acero and R. M. Stern, "Robust Speech Recognition By Normalization of the Acoustic Space," in *Proc. of the ICASSP'91*, Toronto, Canada, 1991.

[11] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalization on Conversational Telephone Speech," in *Proc. of the ICASSP'96*, Atlanta, USA, 1996.

[12] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," in *Proc. of the Eurospeech'01*, Aalborg, Denmark, 2001.

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

[14] J. McDonough, W. Byrne, and X. Luo, "Speaker Normalization with All-Pass Transforms," in *Proc. of the ICSLP'98*, Sydney, Australia, 1998.

[15] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A First Step Towards Text-Independent Voice Conversion," in *Proc. of the ICSLP'04*, Jeju, Korea, 2004.