Frequency Domain vs. Time Domain VTLN

David Sündermann, Antonio Bonafonte
Universitat Politècnica de Catalunya (UPC)
Department of Signal Theory and Communications
08034 Barcelona, Spain
{suendermann,antonio}@gps.tsc.upc.es

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

Hermann Ney
RWTH Aachen
Computer Science Department
52056 Aachen, Germany
ney@cs.rwth-aachen.de

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 compare the standard approach with 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.

Keywords: voice conversion, VTLN, embedded speech processing

1 Introduction

Vocal tract length normalization (VTLN) (Eide and Gish, 1996) 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 (Pye and Woodland, 1997).

The same technique can be used for voice conversion (Sündermann, Ney, and Höge, 2003), which is the modification of a source speaker's voice in order to sound like another speaker (Moulines and Sagisaka, 1995). 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 (Eichner, Wolff, and Hoffmann, 2004). 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) (Charpentier and Stella, 1986). 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 (Black and Lenzo, 2001). 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 2.4, we 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 3.

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 (Kain and Macon, 1998).

To extract pitch-synchronous frames from a given speech signal, we use the algorithm described in (Goncharoff and Gries, 1998). 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.

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 $\tilde{\omega}(\omega|\xi_1, \xi_2, \ldots); 0 \leq \omega, \tilde{\omega} \leq \pi$, 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.

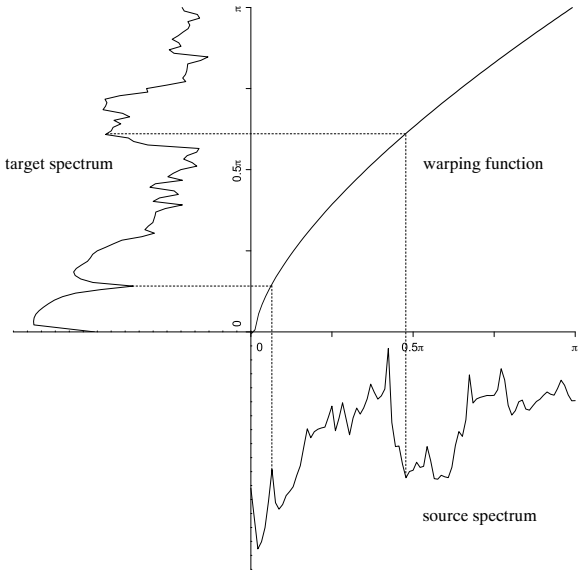

Figure 1: Warping the magnitude spectrum: an example

2.3 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 (Sündermann, Ney, and Höge, 2003). 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.

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. (Sündermann, Ney, and Höge, 2003):

$$\tilde{\omega}(\omega|\omega_1^I, \tilde{\omega}_1^I) = \alpha_i \omega + \beta_i \quad \text{for} \quad \omega_i \leq \omega < \omega_{i+1} \tag{1}$$

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

$$0 = \omega_0 < \omega_1 < \cdots < \omega_I < \omega_{I+1} = \pi,$$

$$\text{for } \tilde{\omega}_i \text{ equivalent}; \quad i = 0 \ldots I.$$

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

parameters	linear	nonlinear
one	• piece-wise linear with two segments – asymmetric (Wegmann et al., 1996) – symmetric (Uebel and Woodland, 1999)	• bilinear (Acero and Stern, 1991) • power (Eide and Gish, 1996) • quadratic (Pitz et al., 2001)
several	• piece-wise linear with several segments (Sündermann, Ney, and Höge, 2003)	• allpass transform (McDonough, Byrne, and Luo, 1998)

Table 1: Categorization of VTLN warping functions

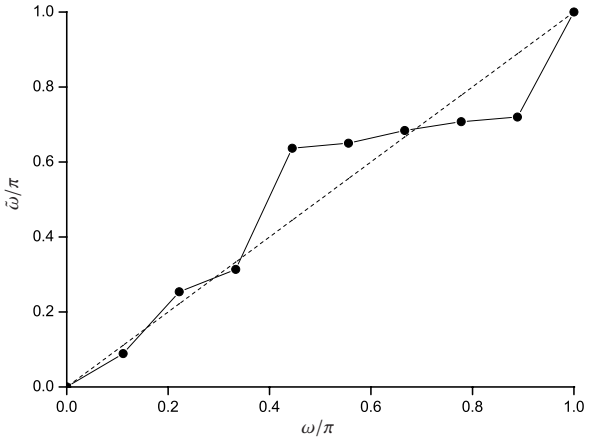

Figure 2: Example of a piece-wise linear warping function

2.4 On the Computational Complexity

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 3.1, $T_f = 101$ 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.

	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

3 Experiments

3.1 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.

3.2 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 (Sündermann et al., 2004) and leads to the following equation:

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

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

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

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

	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

3.3 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.

3.4 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 3.2.
- As Table 4 shows, the overall sound quality of the two compared techniques FD and TD-VTLN is equivalent. The average MOS corresponds to that reported in the literature dealing with VTLN-based voice conversion, cf. (Eichner, Wolff, and Hoffmann, 2004).
- 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 (Sündermann, Ney, and Höge, 2003).

4 Conclusion

This paper addresses the comparison of FD and TD-VTLN in terms of computational complexity and conversion quality. It turns out that the computational costs can be reduced by a factor of about 20 replacing FD by TD-VTLN while keeping the sound quality and the ability of voice identity conversion.

References

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

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

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

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

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

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

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

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

Moulines, E. and Y. Sagisaka. 1995. Voice Conversion: State of the Art and Perspectives. *Speech Communication*, 16(2).

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

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

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

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

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

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