Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis

David Sündermann[1,3], *Guntram Strecha*[2], *Antonio Bonafonte*[1], *Harald Höge*[3], *Hermann Ney*[4]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Dresden University of Technology, Dresden, Germany
[3]Siemens AG, Munich, Germany
[4]RWTH Aachen – University of Technology, Aachen, Germany
david@suendermann.com, guntram.strecha@ias.et.tu-dresden.de,
antonio.bonafonte@upc.edu, harald.hoege@siemens.com, ney@cs.rwth-aachen.de

Abstract

Recently, we demonstrated that vocal tract length normalization (VTLN) can be applied to voice conversion tasks. In particular, when the conversion algorithm is performed in time domain, this technique is very resource-efficient and, consequently, suitable for embedded applications. In this paper, we use VTLN-based voice conversion as a novel feature of a small footprint speech synthesizer running on mobile devices. The characteristics of this feature are investigated by means of extensive subjective tests.

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 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.

The same technique can be used for voice conversion [2], which is the modification of a source speaker's voice in order to sound like another speaker [3]. 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.

In speech recognition, most parts of the signal processing are performed in frequency domain, hence, VTLN is applied to the frequency spectrum. 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) [4]. This leads to the idea of performing the voice conversion in time domain as we demonstrated in a recent publication [5]. It turns out that using time-domain instead of frequency-domain VTLN reduces the computational costs by the factor of 20 without affecting the speech quality. Therefore, TD-VTLN-based voice conversion is particularly suitable for embedded applications.

The next section briefly describes the fundamentals of TD-VTLN-based voice conversion. In Section 3, we apply this technique to a small footprint speech synthesizer running on mobile devices. Then, in Section 4, we control the running time and memory requirements of the technique and present results of an extensive subjective evaluation that investigated the technique's conversion performance and its behavior with respect to the naturalness of the converted voices.

This work has been partially funded by the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation - http://www.tc-star.org

We would like to acknowledge the contribution of the numerous participants in the subjective tests that form the fundament of this work.

2. TD-VTLN-Based Voice Conversion

As mentioned in the introduction, former applications of VTLN to voice conversion were performed in frequency domain [2, 6]. Here, VTLN tries to compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis of the magnitude spectrum, cf. example in Figure 1. In speech recognition literature, several warping functions are proposed, however, our experience shows that their behavior in a voice conversion framework is very similar. Hence, our investigations focused on piece-wise linear warping functions with several segments. It turned out that by exploiting several properties of the discrete Fourier transformation, one can derive a conversion rule that can directly be applied to the time frames of the source speech [5]. Furthermore, a subjective evaluation of the respective techniques showed that the number of warping segments can be reduced to one without degrading neither conversion performance nor speech quality. Consequently, in the following, we limit our considerations to this special case. We

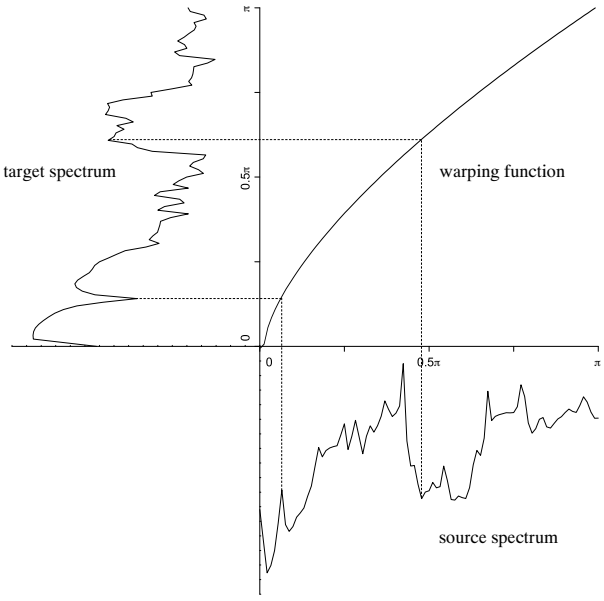

Figure 1: *Warping the magnitude spectrum: an example.*

Table 1: *FD vs. TD-VTLN: Running time and memory requirements.*

	running time / operations			memory / 16bit		
	FD-VTLN	TD-VTLN	TD-VTLN+	FD-VTLN	TD-VTLN	TD-VTLN+
DFT	$4T^2 - 2T$			T		
spline interpolation	$40T$	$40T$		$6T$	$6T$	
linear interpolation			$6T$			T
IDFT	$4T^2 - 2T$			T		
total	$8T^2 + 36T$	$40T$	$6T$	$8T$	$6T$	T

refer to a warping function as $\tilde{\omega}(\omega)$, where ω is a source frequency and $\tilde{\omega}$ is the warped frequency. Now, we want to warp a complex-valued source spectrum X by applying $\tilde{\omega}$ yielding $\tilde{X}$. For the warped spectrum, we have

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

The linear warping function $\tilde{\omega}(\omega) = \alpha\omega$ yields $\tilde{\omega}^{-1}(\omega) = \frac{\omega}{\alpha}$; α is the warping factor. By using the scaling rule of the discrete Fourier transformation, we derive the time correspondence of $\tilde{X}$:

$$\mathcal{F}^{-1}\left\{\tilde{X}(\omega)\right\} = \mathcal{F}^{-1}\left\{X\left(\frac{\omega}{\alpha}\right)\right\} = \alpha x(\alpha t) \,. \qquad (1)$$

This means a simple scaling and stretching of a frame's time signal by the factor α.

3. Voice Conversion as a Module of a Small Footprint Text-to-Speech System

In Figure 2, a typical structure of a text-to-speech (TTS) system is shown. It consists of three blocks: text processing, prosody control, and acoustic module.

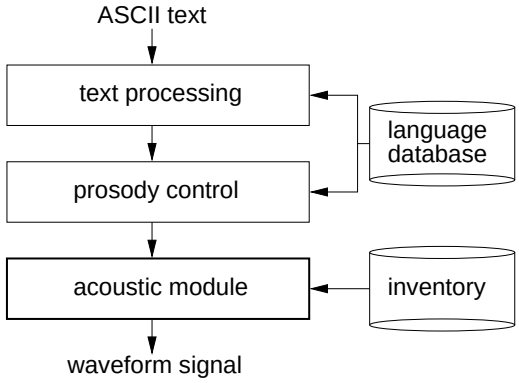

Figure 2: *Scheme of a text-to-speech system.*

The text processing and prosody control of the particular concatenative TTS system used in our investigations are designed for an application in mobile phones, for details, cf. [7].

The acoustic module consists of two submodules: the unit selection that searches for appropriate diphones to match the target phoneme sequence, and the acoustic synthesis, cf. Figure 3. Here, an inventory is required that contains speech segments (diphones) of a certain speaker. As the reduction of the inventory size is one of the essential steps in reducing the system's footprint, these segments may be compressed, e.g. using adaptive differential pulse code modulation or adaptive multirate (narrowband and wideband) [8]. The acoustic synthesis decodes, if necessary, the speech segments determined by the unit selection, transforms them by means of TD-VTLN-based voice conversion as described in Section 2, produces the prosody by applying prosodic targets (fundamental frequency contour, phoneme durations, intensity) and concatenates the selected diphones.

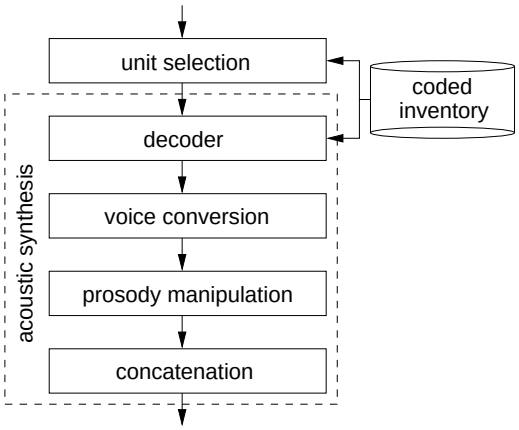

Figure 3: *Scheme of the acoustic module with integrated voice conversion.*

In this work, we used two German inventories, one of a female and one of a male speaker. These inventories consist of 1176 (female) and 1212 (male) different diphones without alternatives. The sizes of the uncompressed 16 kHz, 16 bit inventories are 6.3 MByte and 5.0 MByte, respectively.

4. Evaluation

4.1. Running Time and Memory Requirements

As already mentioned in the introduction, in [5] we showed that performing the VTLN-based voice conversion in time domain instead of in frequency domain essentially accelerates the algorithm. In Table 1, the running time and memory requirements of both algorithms are broken down. Here, T is the number of samples of the considered pitch-synchronous time frame. In our former work, we applied cubic spline interpolation for the resampling according to Eq. 1 whereas for the use in the low footprint speech synthesizer, linear interpolation was used to further reduce the computational costs (in the table referred to as *TD-VTLN+*).

As an example, we look at the experimental speech corpus described in Section 4.2. For the female voice, we have an average frame length of $T = 86$ and for the male $T = 134$. This leads to acceleration factors of 120 and 185, respectively when using TD-VTLN+ instead of FD-VTLN.

4.2. The Experimental Corpus

In order to investigate the properties of the voice conversion algorithm independently of the speech synthesizer, we also performed experiments on real speech data. We chose data that stemmed from the same speakers the inventories described in Section 3 are based on: 38 utterances (152s) of the female and 34 utterances (83s) of the male speaker. The speech signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16kHz.

4.3. The Objective of VTLN-Based Voice Conversion

Former publications on VTLN-based voice conversion [6] stated that this technique is able to change the speaker identity in the way that we suppose to hear a different speaker, whereas, often, it is impossible to generate a certain predefined voice. This leads to the question, how many well-distinguishable voices can be derived from one given voice. For instance, when a speech synthesizer is to be used in a computer game to let computer-animated persons speak, it is helpful when all involved characters have their own voice (possibly correlated to their properties: male vs. female, teen vs. aged, pleasant vs. nasty, etc.).

In the following, we limit the number of parameters describing the VTLN-based voice conversion to the warping factor α (cf. Section 2) and the ratio r between the mean fundamental frequencies after and before the conversion. In our experience, both parameters play an important role when assessing voice identities. To simplify matters, they are combined in the vector $v = (\alpha, r)$.

4.4. On the Naturalness of Converted Voices

It is obvious that only parameter values inside a certain range result in reasonable, i.e. natural sounding voices. E.g., setting $v = (1, 1)$ does not change the voice at all and should result in the maximum naturalness when distortions by the analysis-synthesis system can be neglected. On the other hand, extreme values as $r \to \infty$ produce artificial or even irrecognizable voices.

Hence, at the beginning, we tried to learn a relation between the parameter settings and the voices' naturalness. This was done by performing a subjective test according to [9] where 11 subjects were asked to rate the naturalness of 50 conversion samples derived from the two speaker's databases described in Section 4.2 on a scale between 1 (very artificial) and 5 (very natural). The parameter values were equidistantly distributed along the four lines $v = (x, 1)$, $v = (1, x)$, $v = (1 + ax, 1 + x)$; $a > 0$, and $v = (1 - bx, 1 + x)$; $b > 0$. In doing so, the factors a and b as well as the values of x were determined with the help of informal listening tests. To estimate the naturalness score in the whole v space, we applied a two-dimensional interpolation based on Delaunay triangulation [10]. As an example, Figure 4 shows the results for the female speaker.

4.5. On the Dissimilarity of Converted Voices

The demand for well-distinguishable voices, cf. Section 4.3, leads to the question how the subjective dissimilarity of two voices V' and V'' produced by warping a source voice V depends on the objective difference between the two used parameter vectors $v' = (\alpha', r')$ and $v'' = (\alpha'', r'')$. To describe the dependence between both measures, we introduced the following model:

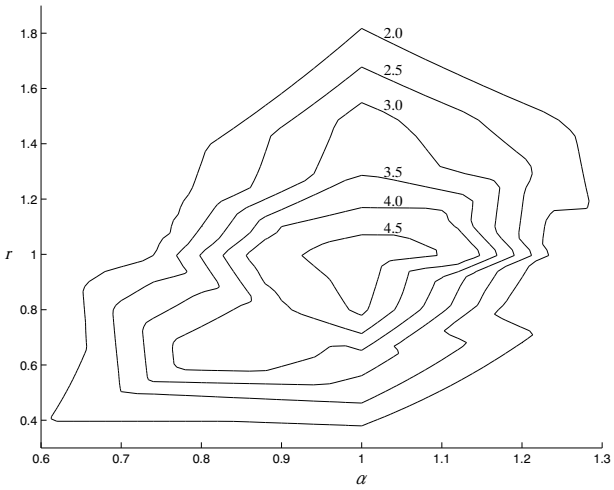

Figure 4: *Dependence of the naturalness of a converted voice on the parameters α and r. The contour lines indicate the levels of naturalness.*

- Since we want to apply the Euclidean distance in the parameter space, we have to make sure that the contributions of the involved parameters are equivalent. E.g., we cannot expect that a listener feels the same dissimilarity when facing voices generated using $v' = (0.9, 1)$ and $v'' = (1.1, 1)$ as when the parameters were $v' = (1, 0.9)$ and $v'' = (1, 1.1)$, although, in both cases, the Euclidean distance is $|v'' - v'| = 0.2$. Therefore, we scale the involved parameters using the weights w and $(1 - w)$, respectively. Besides, we take into account that a logarithmic frequency scale better represents the human perception than the linear one. I.e., by logarithmizing the fundamental frequency ratio, the parameter vectors $v'' = (1, 0.5)$ and $v'' = (1, 2)$ result in the same distance from the vector $v' = (1, 1)$ (one octave).
- Now, we are able to formulate the boundary conditions. In the following, we use the subjective distance D that rates the dissimilarity of two voices on a scale between 1 (definitely identical) and 5 (definitely different), and the objective distance

$$d = \sqrt{w^2(\alpha'' - \alpha')^2 + (1 - w)^2 \log^2\left(\frac{r''}{r'}\right)} \,. \quad (2)$$

  In case, the parameter vectors v' and v'' are identical, i.e. $d = 0$, we expect the lowest dissimilarity ($D = 1$). If the distance between the vectors approaches infinity, we expect the voices to be totally different ($D = 5$). A relation between d and D that fulfills these boundary conditions is the following:

$$D = 5 \cdot \left(1 - \left[\frac{\gamma}{\beta} d + \sqrt[\beta]{\frac{5}{4}}\right]^{-\beta}\right); \quad \gamma, \beta > 0 \,. \quad (3)$$

In order to determine the parameter weight w, we performed another subjective test, where 10 subjects were asked to rate the dissimilarity of 24 pairs of voices derived from the two speakers' databases described in Section 4.2 on the scale defined above. The compared voices were the same as generated for the naturalness experiment in Section 4.4. A pair always

Table 2: *Parameters of the dissimilarity model.*

	female	male
γ	4.7	3.2
w	0.5	0.7
ε	0.3	0.4

consisted of voices derived from opposite parameter vectors, i.e. $v' = (1+x, 1)$ and $v'' = (1-x, 1)$, $v' = (1, 1+x)$ and $v'' = (1, 1-x)$, $v' = (1+ax, 1+x)$ and $v'' = (1-ax, 1-x)$, $v' = (1-bx, 1+x)$ and $v'' = (1+bx, 1-x)$. Averaging over the participants, for each gender, we obtained $I = 12$ scores D_1^I in addition to the corresponding parameter vectors ${v'}_1^I$, and ${v''}_1^I$. By applying Eqs. 2 and 3, we are able to estimate the weight ω as follows:

$$w = \arg\min_{\omega} \min_{\beta,\gamma} \varepsilon(\omega, \beta, \gamma) \quad \text{with}$$

$$\varepsilon(\omega, \beta, \gamma) = \sqrt{\frac{1}{I} \sum_{i=1}^{I} \left[D\left(d_\omega(v_i', v_i''), \beta, \gamma\right) - D_i\right]^2} \,.$$

It turns out that β becomes sufficiently large to approximate Eq. 3 by its limit for $\beta \to \infty$:

$$D = 5 - 4e^{-\gamma d}; \quad \gamma > 0 \,. \tag{4}$$

In Table 2, for both speakers (female and male), the determined parameters are displayed. In order to assess the performance of the model, we also include the (absolute) model error ε:

$$\varepsilon = \min_{\omega,\beta,\gamma} \varepsilon(\omega, \beta, \gamma) \,.$$

4.6. Generating Well-Distinguishable Voices

As we have argued in Section 4.3, we want to use TD-VTLN-based voice conversion to generate a certain number of well-distinguishable voices whose naturalness is above a certain threshold. As an example, from each of the given synthesis voices, we want to create 5 voices with a naturalness score of at least 3.0 and a dissimilarity score of at least 3.0:

- At first, we determine the area of the parameter space that provides a naturalness score greater than 3.0 (cf. contour lines in Figure 4).
- The α axis of the parameter space is scaled by the factor w, the r axis is logarithmized and scaled by $(w-1)$.
- Then, we distribute 5 vectors v_1^5 inside the region so that the minimum distance between two of these vectors $d(v_i, v_j)$ becomes maximal:

$$v_1^5 = \arg\max_{\nu_1^5} \min_{\substack{i,\, j\, =\, 1,\ldots,5 \\ i \neq j}} d(\nu_i, \nu_j) \,.$$

  Figure 5 displays the vectors for the female voice.
- When we take the minimum distance d_{min} between two of the involved vectors and apply it to Eq. 4, we obtain an estimate of the minimal subjective dissimilarity D_{min}. In our case, we have $d_{min} = 0.21$ or $D_{min} = 3.5$ for the female voice and $d_{min} = 0.22$ or $D_{min} = 3.0$ for the male.
- In order to control the fulfillment of the requirement of a minimum dissimilarity score of 3.0, we performed a third subjective test, where we synthesized 40 German sentences using the text-to-speech framework described in Section 3 and asked 13 subjects to rank the dissimilarity of all voice combinations. For the female source voice, we obtained an average score of $\bar{D} = 4.6$ and a minimum score of $D_{min} = 4.4$. For the male voice, the results were $\bar{D} = 4.2$ and $D_{min} = 3.4$, respectively.

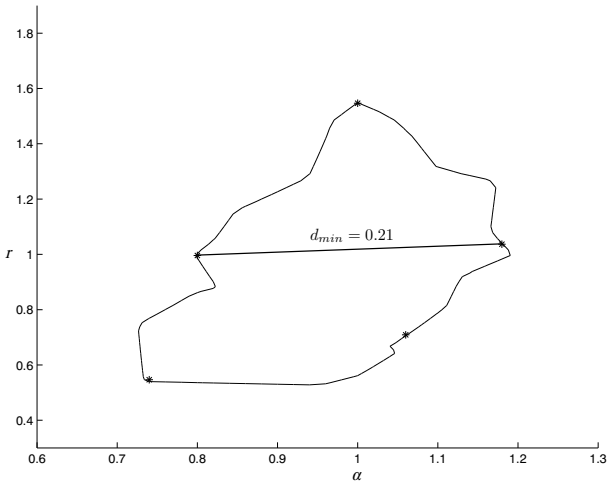

Figure 5: *Distributing five maximally distant vectors in the area of a minimal naturalness score of* 3.0.

5. Conclusion

The subjective test on the dissimilarity of converted synthesized voices proved that the example task of generating five well-distinguishable voices from one source voice succeeded for the female as well as for the male speaker. The dissimilarity model developed in this paper is a relation between distances in the parameter space and subjective dissimilarity scores. However, the model errors of Table 2 and those reported in Section 4.6 show that there is still a need for improvement.

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

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

[4] 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.

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

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

[7] H.-U. Hain, T. Volk, and T. Fingscheidt, "Preprocessing and Prosody Generation for a TTS System with a Very Small Footprint," in *Proc. of the ESSV'03*, Karlsruhe, Germany, 2003.

[8] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha, H. Kruschke, and U. Kordon, "A Multilingual TTS System with Less than 1 Megabyte Footprint for Embedded Applications," in *Proc. of the ICASSP'03*, Hong Kong, China, 2003.

[9] D. Sündermann, A. Bonafonte, H. Duxans, and H. Höge, "TC-STAR: Evaluation Plan for Voice Conversion Technology," in *Proc. of the DAGA'05*, Munich, Germany, 2005.

[10] F. P. Preparata and M. I. Shamos, *Computational Geometry - an Introduction*. New York, USA: Springer, 1985.