VTLN-BASED CROSS-LANGUAGE VOICE CONVERSION

David Sündermann, Hermann Ney

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

Harald Höge

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

ABSTRACT

In speech recognition, vocal tract length normalization (VTLN) is a well-studied technique for speaker normalization. As cross-language voice conversion aims at the transformation of a source speaker's voice into that of a target speaker using a different language, we want to investigate whether VTLN is an appropriate method to adapt the voice characteristics. After applying several conventional VTLN warping functions, we extend the conventional piece-wise linear function to several segments, allowing a more detailed warping of the source spectrum. Experiments on cross-language voice conversion are performed on three corpora of two languages and both speaker genders.

1. INTRODUCTION

Vocal tract length normalization [1] tries to compensate for the effect of speaker dependent vocal tract lengths by warping the frequency axis of the amplitude spectrum. In speech recognition, VTLN aims at the normalization of a speaker's voice in order to remove individual speaker characteristics.

A similar task is voice conversion. It describes the modification of a source speaker's voice such that it is perceived to be spoken by a target speaker [2]. In this paper, we show how VTLN can be applied to this task focusing on cross-language voice conversion.

In Section 2, we delineate the concept of cross-language voice conversion and describe how to find corresponding speech segments respectively artificial phonetic classes in the training material of source and target speaker. These corresponding classes are used to estimate the parameters of class-dependent VTLN warping functions.

Subsequently, in Section 3, we apply this training procedure to conventional warping functions depending on only one parameter.

Often, these conventional functions do not sufficiently model the speakers' characteristics. Therefore, we introduce a piece-wise linear warping function consisting of several linear segments. Another method to augment the number of parameters is the all-pass transform [12]. The greater the parameter number is, the more carefully we must deal with their practical estimation. All these considerations are discussed in Section 4.

Since the parameter estimation for classes with only few observations is very inaccurate and, besides, we do not want the parameters to change abruptly from one class to another, in Section 5, we introduce two parameter smoothing methods.

Finally, in Section 6, we present experimental results on three German and English corpora.

2. VTLN-BASED CROSS-LANGUAGE VOICE CONVERSION

2.1. Motivation

Over the last decade, a lot of scientific effort has been expended to realize a human vision from biblical times: to build a speech-to-speech (S2S) translator. S2S translation systems [3] combine almost every speech processing and natural language processing application, as speech recognition, statistical machine translation, or speech synthesis. It is obvious that sometimes a user of a S2S translator wants his system to speak with his own voice, in particular, when several persons utilize the translator simultaneously. Furthermore, a S2S translation system must be able to cope with a new user and rapidly adapt the utterances of the speech synthesis standard speaker (source speaker S) to the user's voice (target speaker T).

This voice adaptation is performed by the S2S translator's voice conversion module which, in case of being confronted with a new user, must be content with a small amount of training data, namely with only some words, because the translation system is to react with minimum delay. Having seen more data from the target speaker, we are able to refine our models or our model's parameters, respectively.

Most of the training procedures of state-of-the-art voice conversion techniques require training data which have to be the same utterances of both source and target speaker [4]. Besides, these utterances should feature a high degree of

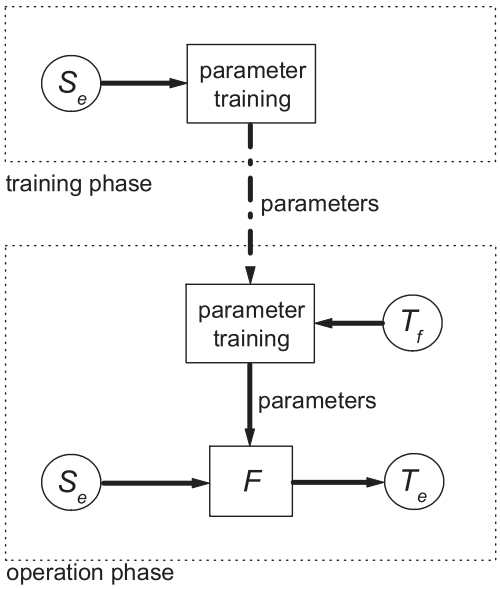

Fig. 1. Cross-Language Voice Conversion for Speech-to-Speech Translation.

natural time alignment and similar pitch contour [5]. However, these claims contradict the conditions of a S2S system processing spontaneous speech of several unknown speakers, and, of course, multiple languages. [4] and [6] report cross-language voice conversion approaches but both expect two bilingual speakers having long experience with the foreign language such that the above requirements for monolingual voice conversion are fulfilled as well.

The concept of *genuine* cross-language voice conversion we propose in [7] performs the complete parameter training in operation phase using T's utterances in language f and those of S in language e. Output are the latter utterances spoken by T using e. Dealing with cross-language voice conversion, one has to take into account that in a real-world application we should search for ways to perform parts of the training beforehand. At least, when we apply cross-language voice conversion to S2S translation, we are familiar with the standard speaker S of the speech synthesis component, hence, we can execute a part of the parameter training off-line. Besides, the training material is not limited to the sparse utterances of S emitted during the translation process but can be extended to the whole database of the speech synthesizer assuming a concatenative synthesis system, v. Figure 1.

2.2. Automatic Segmentation and Mapping

As opposed to monolingual voice conversion or conventional cross-language voice conversion described above, in *genuine* cross-language voice conversion we do not possess corresponding time frames of source and target speaker and, furthermore, language e and f generally use different phoneme sets. In [7], we investigate the following solution of this lack.

At first, we subdivide speech material of speaker S and T into K_S respectively K_T artificial phonetic classes. This is done by clustering the frequency spectra of period-synchronous frames obtained by a pitch tracker. For unvoiced signal parts, pseudo periods are used. Now, for each source class k_S we determine the most similar target class $\hat{k}_T(k_S)$. This class mapping is basis for an arbitrary statistical voice conversion parameter training.

2.3. Statistical Voice Conversion Parameter Training

Let $X_1^I = X_1, \ldots, X_I$ be the spectra belonging to source class k_S and Y_1^J those of the mapped class $\hat{k}_T(k_S)$, we generally estimate the parameter vector ϑ by minimizing the sum of the euclidean distances between all target class spectra and transformed source class spectra. Here, we utilize the spectral conversion function F_ϑ depending on the parameter vector ϑ.

$$\vartheta = \arg\min_{\vartheta'} \sum_{i=1}^{I} \sum_{j=1}^{J} \int_{\omega=0}^{\pi} \left| Y_j(\omega) - F_{\vartheta'}(X_i, \omega) \right|^2 d\omega \quad (1)$$

In Section 2.1, we have argued that for cross-language voice conversion computational resources can be limited because we have to perform a part of the parameter training in operation phase. In conjunction with a suitable smoothing technique, we often can neglect the variety of the classes' observation spectra by introducing a mean approximation without an essential effect on the voice conversion parameters.

$$\vartheta = \arg\min_{\vartheta'} \int_{\omega=0}^{\pi} \left| \bar{Y}(\omega) - F_{\vartheta'}(\bar{X}, \omega) \right|^2 d\omega \quad (2)$$

Here, $\bar{X}$ and $\bar{Y}$ are the source and target classes' average spectra.

3. WARPING FUNCTIONS WITH ONE PARAMATER

In speech recognition, several VTLN warping functions have been proposed whose parameters usually are limited to one variable, the warping factor α. Established warping functions are

- symmetric piece-wise linear function with two segments [8]

$$\tilde{\omega}_\alpha(\omega) = \begin{cases} \alpha\omega & : \quad \omega \leq \omega_0 \\ \alpha\omega_0 + \frac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & : \quad \omega \geq \omega_0 \end{cases} \quad (3)$$

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & : \quad \alpha \leq 1 \\ \frac{7}{8\alpha}\pi & : \quad \alpha \geq 1 \end{cases}$$

- power function [9]

$$\tilde{\omega}_\alpha(\omega) = \left(\frac{\omega}{\pi}\right)^\alpha$$

- quadratic function [10]

$$\tilde{\omega}_\alpha(\omega) = \omega + \alpha\left(\frac{\omega}{\pi} - \left(\frac{\omega}{\pi}\right)^2\right)$$

- bilinear function [11]

$$\tilde{z}_\alpha(z) = \frac{z-\alpha}{1-\alpha z} \quad \text{with} \quad z = e^{i\omega} \tag{4}$$

In order to estimate the class dependent warping factor α, we use Eqs. 1 or 2, where

$$F_\alpha(X,\omega) = X(\tilde{\omega}_\alpha(\omega)). \tag{5}$$

4. WARPING FUNCTIONS WITH SEVERAL PARAMETERS

4.1. Piece-Wise Linear Warping with Several Segments

One of the disadvantageous properties of the conventional warping functions with one parameter is that the whole frequency axis is always warped in the same direction, either to lower or to higher frequencies. Consequently, these functions are not able to model spectral conversions where certain parts of the axis move to higher frequencies, and other parts to lower frequencies, or vice versa. Such functions would require at least one inflection point and would cross the $\tilde{\omega} = \omega$ diagonal.

Applying the VTLN technique to voice conversion, we want to use more exact models than in speech recognition, i. e. warping functions with several parameters, to better describe the individual characteristics of the speakers' vocal tracts.

Assuming there is an ideal warping function for a given class pair $(k_S, \hat{k}_T)$, an obvious model is the interpolation of this function by several linear segments, as a consequence from the simple two-segment linear warping, v. Eq. 3.

$$\tilde{\omega}_{\tilde{\omega}_1^S}(\omega) = \begin{cases} \tilde{\omega}_{0,\tilde{\omega}_1}(\omega) & \text{for} \quad 0 \le \omega \le \frac{1}{S+1}\cdot\pi \\ \vdots & \vdots \\ \tilde{\omega}_{\tilde{\omega}_s,\tilde{\omega}_{s+1}}(\omega) & \text{for} \quad \frac{s}{S+1}\cdot\pi \le \omega \le \frac{s+1}{S+1}\cdot\pi \\ \vdots & \vdots \\ \tilde{\omega}_{\tilde{\omega}_S,\pi}(\omega) & \text{for} \quad \frac{S}{S+1}\cdot\pi \le \omega \le \pi \end{cases} \tag{6}$$

$$\tilde{\omega}_{\tilde{\omega}',\tilde{\omega}''}(\omega) = \tilde{\omega}' + \left(\frac{S+1}{\pi}\cdot\omega - s\right)\cdot(\tilde{\omega}'' - \tilde{\omega}')$$

$$0 \le \tilde{\omega}_1 \le \cdots \le \tilde{\omega}_S \le \pi. \tag{7}$$

This formula describes a piece-wise linear function $\tilde{\omega}(\omega)$ starting at $(0,0)$, ending at (π,π), and connecting S points whose ω values are equidistantly distributed. The corresponding $\tilde{\omega}_s$ are the parameters of the warping function. The resulting function is monotonous according to Eq. 7, as we do not want parts of the frequency axis to be exchanged. For an example, v. Figure 4.

4.2. VTLN with All-Pass Transform

As the piece-wise linear warping function with several segments, the all-pass transform [12] also deals with the claim that we want more flexible warping functions by extending the number of parameters. As opposed to the piece-wise warping, the all-pass transform, in general, results in a non-linear warping function.

In addition to the real warping factor $-1 < \alpha < 1$, a set of complex parameters is introduced: $\{\beta_p, \gamma_p : p = 1,\ldots,P\}$ with $|\beta_p| < 1$ and $|\gamma_p| < 1$.

The all-pass transform is defined as

$$\tilde{z}(z) = \frac{z-\alpha}{1-\alpha z}\prod_{p=1}^{P}\left(\frac{z-\beta_p}{1-\beta_p^* z}\frac{z-\beta_p^*}{1-\beta_p z}\right)\left(\frac{1-\gamma_p^* z}{z-\gamma_p}\frac{1-\gamma_p z}{z-\gamma_p^*}\right)$$

$$\text{with} \quad z = e^{i\omega}.$$

It is obvious that for $P = 0$ respectively $\beta_p = \gamma_p = 0$; $p = 1,\ldots,P$, this general all-pass transform formula passes into the bilinear function, v. Eq. 4.

4.3. Practical Parameter Estimation

In general, augmenting the number of parameters, confronts us with an increasing need of computation time. Particularly, this is the case if the minimization of Eqs. 1 or 2 is performed by calculating the distances for all possible parameter combinations concerning a certain resolution. This estimation method results in an exponential increase of computing time in dependence on the number of considered parameters.

Viewing the definition of the piece-wise linear warping function with several segments, cf. Eq. 6, we note that the integrals used in Eqs. 1 and 2 can be rewritten as (also cp. Eq. 5)

$$\begin{aligned} d_{\tilde{\omega}_1^S} &= \int_{\omega=0}^{\pi} \left|Y(\omega) - X(\tilde{\omega}_{\tilde{\omega}_1^S}(\omega))\right|^2 d\omega \\ &= \sum_{s=0}^{S} \int_{\omega=\frac{s}{S+1}\cdot\pi}^{\frac{s+1}{S+1}\cdot\pi} \left|Y(\omega) - X(\tilde{\omega}_{\tilde{\omega}_s^{s+1}}(\omega))\right|^2 d\omega\,. \end{aligned}$$

This enables us to use dynamic programming for searching the minimum distance and therewith the optimal parameter vector $\tilde{\omega}_1^S$.

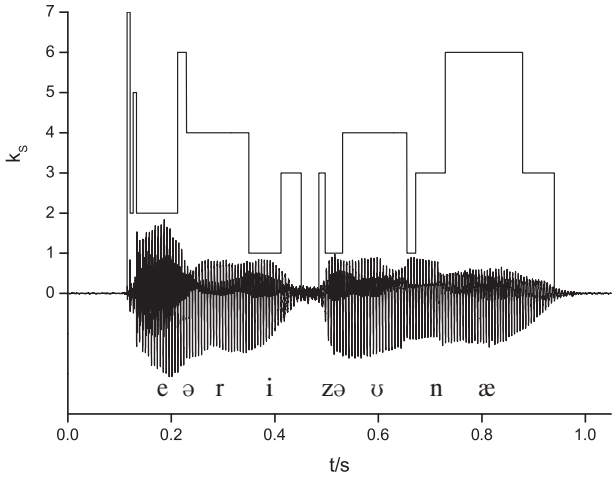

Fig. 2. Automatic Class Segmentation for the Word "Arizona".

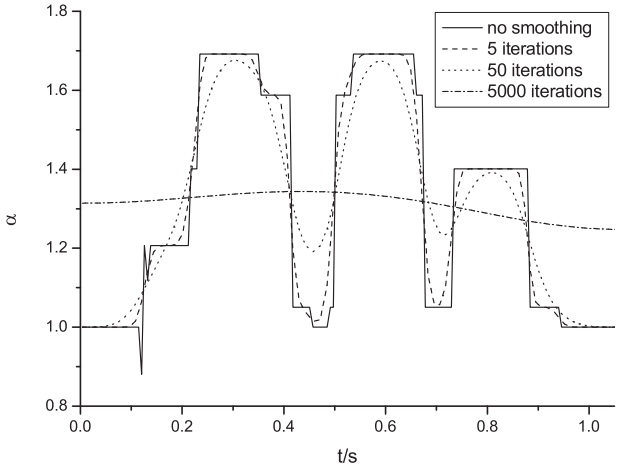

Fig. 3. Iterative Integrating Smoothing for Warping Functions with one Parameter.

Unfortunately, this simplification cannot be applied to the all-pass transform, thus we are forced to find another appropriate minimization technique. For instance, the gradient method determines a local minimum of a multidimensional function next to a given initial vector.

Since we search for the *global* minimum of the distance function, ϑ, we have to ensure that the initial value is in the next neighborhood because we often do not have a concave function of the parameters $\alpha, \beta_1^P, \gamma_1^P$.

We expect ϑ to be in the environment of the parameter vector ϑ_0 which results in the diagonal warping function $\tilde{\omega} = \omega$. Now, we obtain a new initial vector ϑ_1 by determining its components using normally distributed random numbers. Mean vector of this distribution is ϑ_0, the covariance matrix should be diagonal if we expect the parameters to be independent of each other. If the distance which results from a new run of the gradient method is smaller than that for the initial vector ϑ_0, we memorize it as the currently best solution. This procedure is repeated R times.

We notice that smaller variances lead faster to the global minimum if we are sure that it is near ϑ_0, otherwise the variances should be increased. At any rate, each variance greater than zero yields ϑ for $R \to \infty$.

5. PARAMETER SMOOTHING

5.1. Iterative Integrating Smoothing

Basis of the cross-language voice conversion technique delineated in this paper is the automatic class segmentation and mapping described in Section 2.2. In Figure 2, we show the time course of the word "Arizona" and the corresponding classes for $K_S = 8$.

To avoid that the class-dependent voice conversion parameters jump at the class boundaries causing distinctly audible artifacts in the converted speech, we introduce an integrating parameter smoothing which iteratively adapts a parameter vector by adding a weighted mean of the chronologically neighbored vectors. Figure 3 shows the effect of this smoothing technique for 5, 50 and 5000 iterations using the symmetric piece-wise warping function described in Eq. 3. If the number of iterations approaches infinity, we obtain a constant function over the time representing the mean parameter vector.

These considerations are applicable to the piece-wise warping function with several segments as well. In Figure 4, we show the results for 50 iterations. At the mark $t = 0.6s$, we exemplify this warping technique, formally defined in Eq. 6. It is obvious that the particular property of warping functions with several parameters – they can cross the diagonal – is also true for the piece-wise warping (this has been discussed for the all-pass transform in Section 4.2).

5.2. Deviation Penalty

Viewing Figures 2 and 3, we note that for certain classes the obtained parameter values highly deviate from the mean. E. g. for $k_S = 7$ we obtain an α less than 1, whereas the particular voice conversion (female–male) should result in values greater than 1. Considering the mean of $\bar{\alpha} = 1.3$, the parameter values are to be controlled and, if necessary, corrected towards the mean.

This is performed by applying the minimization Eqs. 1 or 2 a second time, having added a penalty term to the enclosed integral. Both addends are normalized by their maximum and then weighted utilizing the real value $0 \leq \lambda \leq 1$ to adjust the penalty strength. Hence, $\lambda = 1$ does not in-

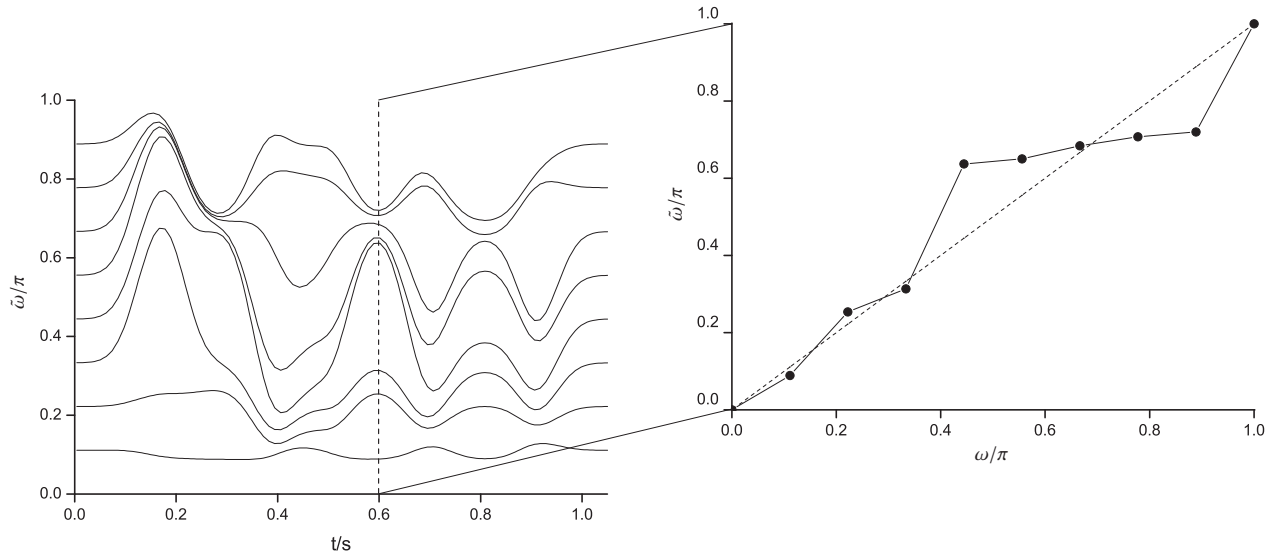

Fig. 4. Iterative Integrating Smoothing for Piece-Wise Warping with Nine Segments ($S = 8$) .

fluence the class parameters at all, whereas $\lambda = 0$ forces all parameters to be equal to their mean $\bar{\vartheta}$. An equilibrium between both terms is to be around $\lambda = 0.5$.

In the following, we assume X and Y to have the unity energy E_0 in order to remove the dependence of the distances on the signal loudness.

$$\begin{aligned} d_\vartheta &= \lambda \frac{\int_{\omega=0}^{\pi} |Y(w) - X(\tilde{\omega}_\vartheta(\omega))|^2 \, d\omega}{\max_{X',Y'} \int_{\omega=0}^{\pi} |Y'(w) - X'(\omega)|^2 \, d\omega} \\ &\quad + (1-\lambda) \frac{\int_{\omega=0}^{\pi} \left(\tilde{\omega}_{\bar{\vartheta}}(\omega) - \tilde{\omega}_\vartheta(\omega)\right)^2 d\omega}{\max_{\bar{\vartheta}',\vartheta'} \int_{\omega=0}^{\pi} \left(\tilde{\omega}_{\bar{\vartheta}'}(\omega) - \tilde{\omega}_{\vartheta'}(\omega)\right)^2 d\omega} \end{aligned}$$

After calculating the maximal distance between arbitrary complex spectra X' and Y' respectively real warping functions $\bar{\vartheta}'$ and ϑ', we obtain

$$\begin{aligned} d_\vartheta &= \int_{\omega=0}^{\pi} \left\{ \frac{\lambda}{4E_0} |Y(\omega) - X(\tilde{\omega}_\vartheta(\omega))|^2 \right. \\ &\quad \left. + \frac{1-\lambda}{\pi^3} \left(\tilde{\omega}_{\bar{\vartheta}}(\omega) - \tilde{\omega}_\vartheta(\omega)\right)^2 \right\} d\omega. \end{aligned}$$

6. EXPERIMENTS

Several experiments have been performed to investigate the properties of VTLN voice conversion with respect to the warping functions discussed in this paper.

As argued in Section 2.1, we used three sparse corpora to investigate our model of cross-language voice conversion:

[A] 3 English sentences of a female speaker,

[B] 10 German sentences of a male speaker (poems),

[C] 3 German sentences of a male speaker (news).

In the following, we report results for three combinations of these corpora:

- F2M: female [A] is converted to male [B],
- M2F: male [B] is converted to female [A],
- M2M: male [C] is converted to male [B].

As error measure, we use the normalized class average distance

$$d_{cad} = \frac{\sum_{k=1}^{K_S} \int_{\omega=0}^{\pi} \left| \bar{Y}_k(\omega) - \bar{X}_k(\tilde{\omega}_{\vartheta_k}(\omega)) \right|^2}{4 K_S E_0} .$$

Again, $\bar{X}$ and $\bar{Y}$ are spectra with unity energy E_0, consequently, we have $0 \leq d_{cad} \leq 1$ (cp. Section 5.2).

In Table 1, we show results for warping functions with one parameter (cf. Section 3). In the third row the results for the trivial solution $\tilde{\omega} = \omega$, i. e. no warping at all, is displayed to assess the absolute d_{cad} values.

We note that the presented warping techniques do not essentially differ, but nevertheless, in our experiments, the power function consistently produced the best outcomes. The most significant effect was achieved for male-to-female voice conversion which is due to the large differences of the

Table 1. *Error Measure for Warping Functions with One Parameter*

warping function	class average distance [%]		
	F2M	M2F	M2M
no warping	8.3	13.2	7.3
piece-wise linear	6.0	**6.4**	**6.2**
power	**5.2**	**6.4**	**6.2**
quadratic	5.4	7.8	**6.2**
bilinear	5.5	6.5	**6.2**

vocal tract. Concerning the above results, the other way around is more complicated. This statement is also supported by our next experiments dealing with the piece-wise warping with several segments, v. Table 2

Table 2. *Error Measure for the Piece-Wise Warping Function with Several Segments*

S	class average distance [%]		
	F2M	M2F	M2M
1	6.7	7.6	6.3
2	6.0	6.1	5.7
4	5.4	5.0	5.1
8	4.9	4.1	4.7
16	4.5	3.4	4.0
32	4.2	2.3	3.0
64	**4.1**	**1.4**	**2.3**

This table conspicuously demonstrates how the number of free parameters affects the warping precision. If S becomes the number of spectral lines of the compared spectra, it passes into a variant of dynamic frequency warping with certain constraints.

Nevertheless, subjective tests have shown, that excessively increasing the number of free parameters, results in an overfitting between source and target spectra and therewith disturbs the naturalness of the output speech.

Future experiments are to investigate the consistency of the above results on other corpora and compare the piecewise warping function with the all-pass transform considering equal numbers of free parameters. Furthermore, the overfitting effect is to be demonstrated using an adequate objective error criterion.

7. CONCLUSION

In this paper, the concept of cross-language voice conversion has been adapted to be used for spontaneous speech-to-speech translation. We show how vocal tract length normalization, well-known from speech recognition, is used as voice conversion technique. Four conventional VTLN warping functions are faced two extended warping models, the piece-wise warping function with several segments and the all-pass transform. Furthermore, we delineate two parameter smoothing approaches and conclude with presenting experimental results on three corpora of two languages.

8. REFERENCES

[1] T. Kamm, G. Andreou, and J. Cohen, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," in *Proc. of the 15th Annual Speech Research Symposium, Baltimore, USA*, 1995.

[2] E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspectives," in *Speech Communication, 16(2)*, 1995.

[3] Y. Gao and A. Waibel, "Speech-to-speech translation," in *Proc. of the ACL'02 Workshop on Speech-to-Speech Translation, Philadelphia, USA*, 2002.

[4] O. Türk, "New methods for voice conversion," in *PhD Thesis, Boğaziçi University, Istanbul, Turkey*, 2003.

[5] A. Kain and M. W. Macon, "Spectral voice transformations for text-to-speech synthesis," in *Proc. of the ICASSP'98, Sydney, Australia*, 1998.

[6] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on gmm and straight," in *Proc. of the EUROSPEECH'01, Aalborg, Denmark*, 2001.

[7] D. Sündermann and H. Ney, "An automatic segmentation and mapping approach for voice conversion parameter training," in *Proc. of the AST'03, Maribor, Slovenia*, 2003.

[8] L. F. Uebel and P. C. Woodland, "An investigation into vocal tract length normalization," in *Proc. of the EUROSPEECH'99, Budapest, Hungary*, 1999.

[9] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of the ICASSP'96, Atlanta, USA*, 1996.

[10] 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.

[11] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space," in *Proc. of the ICASSP'91, Toronto, Canada*, 1991.

[12] J. McDonough, "Speaker normalization with all-pass transforms," in *Technical Report No. 28, Center for Language and Speech Processing, John Hopkins University*, 1998.