An Automatic Segmentation and Mapping Approach for Voice Conversion Parameter Training

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

Abstract

In many applications of voice conversion (VC), we do not possess corresponding training utterances of source and target speaker. In this paper, an automatic phonetic class segmentation and mapping approach based on dynamic frequency warping is presented. After locating corresponding classes of source and target speaker, we are able to apply conventional parameter training methods for VC. As an example, we utilize this approach to estimate the parameters of warping functions for vocal tract length normalization which serves as a simple VC technique.

1 Introduction and Motivation

Voice conversion describes the modification of a source speaker's voice such that it is perceived to be spoken by a target speaker (Moulines and Sagisaka, 1995). One of the most important application fields of VC is speaker adaptation in text-to-speech (TTS) systems (Kain and Macon, 1998), (Tamura et al., 1998), (Tang et al., 2001). Here, the aim is to optimize the system for one or two standard speakers and then adapt the synthesized speech to an arbitrary target speaker. One important advantage of applying VC within TTS systems is the small necessary amount of target speaker training data and the abolition of compilation and parametrization which would be required in case of introducing a new voice into the system.

Over the last decade, text-to-speech systems became a new repute in the context of speech-to-speech (S2S) translation (Gao and Waibel, 2002), e.g. in the *Verbmobil* project, a past research program of Germany's Federal Ministry of Research and Technology. (Kay et al., 1994) write about their vision regarding this project: "*Verbmobil* is a portable simultaneous interpreter. Carry it to a meeting with speakers of other languages and it will translate your spoken words for them." Reading these lines, we understand that also for S2S translation the application of voice conversion is of interest, since each *Verbmobil* user wants his personal voice to be represented, although, in particular, the outcomes of the Verbmobil project have shown that we are still far away from realizing the vision formulated above. Other fields of combinations of S2S translation and VC are: multilingual telephony, movie synchronization, broadcast translation, etc.

We have argued that one of the VC's benefits in comparison with the conventional way of introducing a new voice into a TTS system is the requirement of only little training data. Nevertheless, we need at least a few utterances of source and target speaker. For instance, (Kain and Macon, 1998) have used only 31 short words and 8 sentences, yielding about one minute of speech.

Most of the training procedures of state-of-the-art VC techniques require training data which have to be the same utterances of both source and target speaker (Türk, 2003). Besides, these utterances should feature a high degree of natural time alignment and similar pitch contour (Kain and Macon, 1998). However, these claims extremely contradict

the conditions of a S2S system processing spontaneous speech of multiple unknown speakers, and, of course, multiple languages. (Mashimo et al., 2001) and (Türk, 2003) report cross-language VC approaches but both expect two bilingual speakers having long experience with the foreign language such that the above requirements for mono-lingual VC are fulfilled as well. In Figure 1, we have a schematic view of this conventional cross-language approach. S is the source speaker using in training language f and in operation phase language e, T the target speaker with the same language combination.

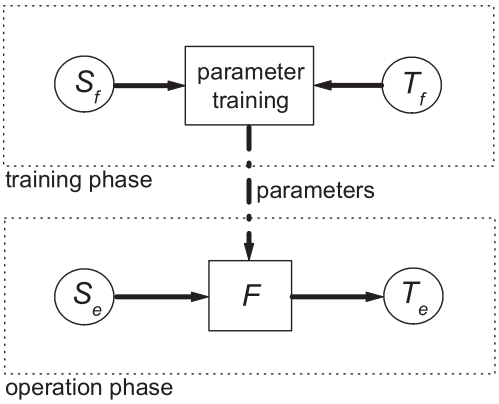

Figure 1: *Conventional Cross-Language Voice Conversion*

Now, we want to assume a *genuine* cross-language voice conversion technique which is familiar only with the standard speaker of a TTS system (as a basic component of a S2S translation system). This standard speaker is the source speaker S of our VC module using language e, cf. Figure 2.

We have mentioned that the conventional VC approach should need an appropriate natural time alignment between the training material of speaker S and T and, preferably, similar pitch contours. This is required in order to support the forced time alignment routine which provides corresponding time frames of S and T. Consequently, we can use common training methods for the VC parameter training to obtain a conversion function F that transforms feature vectors of S to those of T. Such methods are among others

- Vector Quantization with mapping codebooks (Stylianou et al., 1998),

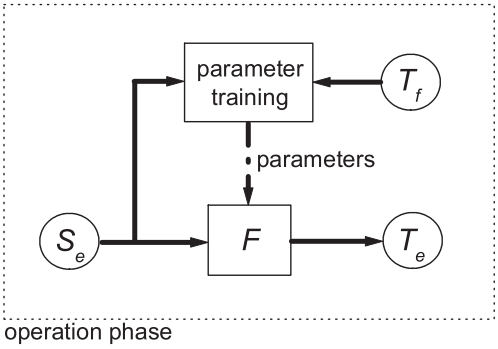

Figure 2: *'Genuine' Cross-Language Voice Conversion*

- Neural Networks (Lee et al., 1996),
- Gaussian Mixture Models (Stylianou et al., 1998).

In case of the *genuine* cross-language VC approach, we do not have the above time frame correspondents. The following considerations deal with remedying this deficiency by formulating an algorithm which automatically determines segments of artificial phonetic classes within the speech material (Section 2) and a mapping between corresponding classes of speaker S and T using dynamic frequency warping (Section 3). Differences in the pitch contour are eliminated by a vector length normalization. As an example for VC parameter training based on automatically determined class mapping, in Section 4, we show how vocal tract length normalization (VTLN) can be applied to voice conversion and how its class-dependent parameters can be estimated. Finally, in Section 5, we present experimental results concerning the automatic class segmentation and mapping algorithm.

2 Segmentation of Speech Material into Artificial Phonetic Classes

We are given a time discrete speech signal x which is apportioned into N pitch-synchronous frames in a way that the initial and terminal samples are located next to zero crossings. This pre-segmentation is done by means of a pitch tracker, cp. e.g. (Hess, 1983), (Talkin, 1995).

Performing DFT (Discrete Fourier Transformation) and removing redundancy from the frequency

representation results in half the number of complex spectral lines as the number of samples in the corresponding frame. In the following, we only consider the amplitude spectrum, because the phase spectrum hardly influences the affiliation to a phonetic class.

As we process pitch-synchronous frames, we obtain a homogeneous spectrum which can be interpolated with the help of cubic splines. On this spectral envelope, we equidistantly distribute P points, P being half the maximum number of frame samples in the speech signal. These length normalized spectral vectors are independent of the basis frequency that hardly affects the phonetic class either.

By removing those spectral vectors which have an energy lower than a certain threshold, we want to prevent unintended interferences by irrelevant signal parts (like background noises) during the following processing steps. To remove the influence of the signal loudness as well, the remaining spectral vectors are normalized in their amplitude such that they have unity energy.

Now, we apply a clustering algorithm on the energy and length normalized frequency vectors (e.g. K-means, hierarchical agglomerative clustering) and obtain K sets of maximally distant vectors: the members of K artificial phonetic classes.

Some clustering methods directly provide the cluster centers, but as mentioned above, we do not have any information about the phase spectrum anymore. Besides, this mean spectrum is not the cluster's unconditionally best representation, because significant signal characteristics can be lost during the averaging operations. Therefore, we compare all members $X_1, \ldots, X_M$ of a certain class k with each other, summing up the euclidean distances to all other class members and finding the minimum of these scores and, therewith, the most central class index

$$\hat{m} = \arg \min_{m=1,\ldots,M} \sum_{\mu=1}^{M} \sum_{p=1}^{P} (X_m(p) - X_\mu(p))^2. \qquad (1)$$

$X_k := X_{\hat{m}_k}$ is to be the corresponding most typical spectrum of class k.

A block diagram of the whole procedure described in this section is displayed in Figure 3.

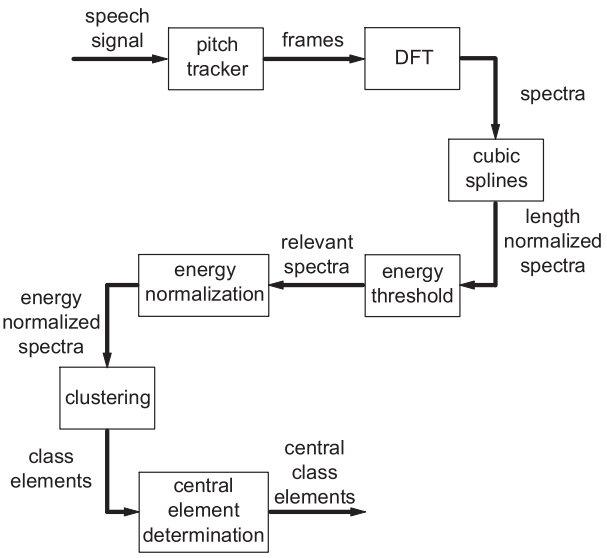

Figure 3: *Automatic Class Segmentation*

3 Class Mapping between Source and Target Speaker

Applying the algorithm for automatic class segmentation as explained above to speech material of source and target speaker, we obtain K_S respectively K_T classes with their most central elements, i.e. most typical spectra X_{k_S} and X_{k_T}. Now, we want to find an optimal mapping between source and target classes.

In the considerations described in Section 2, we have directly compared P points of the length normalized spectrum of one speaker with each other, v. Equation 1. This turns out to be reasonable, as we do not expect significant variations of the spectral envelope within one phonetic class. The situation can change dramatically if we compare the spectra of two different speakers, perhaps with different sexes. Here, another adaptation technique is necessary. Already in the 1980s, dynamic frequency warping (DFW) has been proposed for vowel classification and normalization (Ainsworth et al., 1984), (Matsumoto and Wakita, 1986).

In order to obtain a meaningful warping function, we want to define some constraints. Let us consider a two-dimensional matching space between the amplitude spectra X and Y. Introducing the index $j = 1, \ldots, J$, the warping function is explained by $\{X(C_X(j)), Y(C_Y(j))\}$. Here, $C_X(j)$ and $C_Y(j)$ are index functions so that

- $C_X(1) = C_Y(1) = 1$ (start in the first index),
- $C_X(J) = C_Y(J) = P$ (finish in the maximum index),
- $C(j) \in \{C(j-1), C(j-1)+1\}$ (warping function is monotonous).

In addition to these constraints, we introduce a constant $\Pi \geq 0$ which is to support the *diagonality* of the warping function respectively to suppress irrealistic jumps or retentions similar to a gap penalty (for details about the adjusting of Π cf. Section 5.2). Hence, the distance of the two spectra X and Y is defined as the minimum costs of a warping function between X and Y considering the above constraints (δ denotes the Kronecker Delta).

$$d(X,Y) = \min_{C_X,C_Y,J} \sum_{j=1}^{J} \Big\{ [X(C_X(j)) - Y(C_Y(j))]^2 - \Pi\delta(\Delta C_X(j), \Delta C_Y(j)) \Big\}$$

$$\text{with } \Delta C(j) = C(j) - C(j-1).$$

The optimal mapping of the class k_S is the argument of the minimum distance between the central spectra of k_S and k_T for $k_T = 1, \ldots, K_T$.

$$\hat{k}_T(k_S) = \arg \min_{k_T=1,\ldots,K_T} d(X_{k_S}, X_{k_T})$$

4 VTLN for Voice Conversion

4.1 Parameter Training

Vocal tract length normalization is a well-studied technique in speech recognition (Kamm et al., 1995). It tries to compensate for the effect of speaker dependent vocal tract lengths by warping the frequency axis of the amplitude spectrum. As opposed to the virtual arbitrarity of warping functions provided by DFW (cf. Section 3), the number of different VTLN warping functions is strongly restricted, because, in general, the parameter training is very time-consuming and, furthermore, has to be executed during the operation phase of the speech recognizer, where computing speed is expensive.

In most cases, the number of VTLN parameters is limited to one variable, the warping factor α. Established warping functions are

- piece-wise linear function

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

  - asymmetric (Wegmann et al., 1996)

$$\omega_0 = \frac{7}{8}\pi$$

  - symmetric (Uebel and Woodland, 1999)

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

- power function (Eide and Gish, 1996)

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

- quadratic function (Pitz and Ney, 2003)

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

- bilinear function (Acero and Stern, 1991)

$$\tilde{\omega}_\alpha(\omega) = -i \log \frac{e^{\omega i} + \alpha}{1 + \alpha e^{\omega i}}$$

- all-pass transforms (McDonough et al., 1998)

$$\tilde{\omega}_\alpha(\omega) = -i \log\left(\left(e^{\omega i} - \alpha\right) \sum_{n=0}^{\infty} \alpha^n e^{n\omega i}\right)$$

- and chains of them (Molau et al., 2000).

Assuming we are given a spectrum X, a parameter α', and a warping function $\tilde{\omega}$, we can compute the values of the warped spectrum $\tilde{X}$ for $p = 1, \ldots, P$.

$$\tilde{X}(p, \alpha') = X\left(\frac{P}{\pi}\tilde{\omega}_{\alpha'}\left(\frac{\pi}{P}p\right)\right) \tag{2}$$

Since X is a discrete spectrum accepting only integer arguments $1, \ldots, P$, we estimate $\tilde{X}$ again by cubic spline interpolation, cp. Section 2.

The optimal warping factor α can be determined by minimizing the euclidean distance between target spectrum and warped source spectrum.

$$\alpha_{k_S} = \arg \min_{\alpha'=[\alpha_{min},\alpha_{max}]} \sum_{p=1}^{P} (X_{\hat{k}_T(k_S)}(p) - \tilde{X}_{k_S}(p,\alpha'))^2$$

The range $[\alpha_{min}, \alpha_{max}]$ depends on the warping function type and on the characteristics of the converted speakers, e.g. we can have $[0.7, 1.3]$ for the piecewise linear function and $[-0.5, 0.5]$ for the bilinear function.

4.2 Voice Conversion

All above considerations (Section 2 to 4.1) only attend to the *parameter training*, cp. Figure 2. In our case, the number of *parameters* is limited to one, namely α.

The voice conversion itself (represented in the diagram by VC function F) is composed of three partial conversions:

- spectral conversion (VTLN),
- basis frequency conversion,
- and speaking rate conversion.

Given I complex spectral vectors of the source speaker X_i (output of the DFT, cf. Figure 3), their lengths P_i, the corresponding classes k_i, and the average numbers of frame samples of source and target speaker $\bar{P}_S$ respectively $\bar{P}_T$, we can merge both spectral conversion and basis frequency conversion by adapting Equation 2.

$$\tilde{X}_i(p) = X_i\left(\frac{P}{\pi}\tilde{\omega}_{\alpha_{k_i}}\left(\frac{\pi\bar{P}_S}{P_i\bar{P}_T}p\right)\right)$$

$$\text{for} \quad p = 1, \ldots, \left\lfloor \frac{P_i\bar{P}_T}{\bar{P}_S} \right\rfloor$$

We obtain the spectral warped complex vectors $\tilde{X}_i$ which possess a lower dimensionality in case of a higher mean basis frequency of the source speaker compared to the target speaker, and vice versa. Simply concatenating their IDFTs (Inverse Discrete Fourier Transforms) would result in acceleration or deceleration of the output signal. Besides, we sometimes intend to influence the speaking rate such that we can combine both steps by skipping respectively repeating frames with respect to a ratio consisting of $\bar{P}_S$, $\bar{P}_T$ and a speaking rate manipulation factor.

5 Experiments

Several experiments have been performed to demonstrate the potency and limitations of the presented automatic segmentation and mapping approach.

As argued in the introduction, we used two very discriminative and sparce corpora to investigate our model of *genuine* cross-language Voice Conversion (v. Figure 2):

- 10 German sentences of a male speaker,
- 3 English phrases of a female speaker.

5.1 Automatic Class Segmentation

First, we want to deal with the algorithm for automatic class segmentation proposed in Section 2. We let the algorithm generate $K \in \{1, 2, 4, 8, 16\}$ artificial phonetic classes. Since most of these artificial classes are more or less similar to established phonemes, we specify the IPA symbols of the phonemes which are most similar to the sound corresponding to the central class element. In Table 1 we display the results for the male speaker, in Table 2 those for the female.

Table 1: *Class Segmentation for the Male Speaker: Phoneme Correspondents*

k	IPA				
	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
1	ɜ	ŋ	ɜ	y	ŋ
2		ø	a	ɑ̃	ɑ̃
3			e	e	ø̃
4			ʃ	s	θ
5				œ̃	ɜ
6				a	a
7				ɜ̃	õ
8				ʃ	ç
9					ɜ̃
10					ɔ
11					o
12					s
13					e
14					æ
15					œ
16					ʃ

Table 2: *Class Segmentation for the Female Speaker: Phoneme Correspondents*

k	IPA				
	$K=1$	$K=2$	$K=4$	$K=8$	$K=16$
1	ɜ	ŋ	ɔ	ɛ̃	ø̃
2		a	a	ɒ	ɒ
3			ɜ	ɔ̃	ɛ
4			ʃ	ʃ	ʃ
5				ŋ	ɔ̃
6				ã	ã
7				ɜ̃	ɜ̃
8				ʃ	ʃ
9					œ
10					a
11					ɪ
12					ʃ
13					ŋ
14					e
15					ɔ
16					ʃ

Viewing Tables 1 and 2, in particular, we note two fundamental properties of the proposed segmentation method:

- The artificial phonetic classes represent only stationary phonemes (vowels, nasals, fricatives). This is because we have cut the speech material into small units (frames) losing all context information which is necessary for modeling non-stationary phonemes (in particular plosives). Whether this loss of information affects the Voice Conversion performance is to be investigated in the future.

- Increasing the number of classes, we obtain central class elements which sometimes are very similar. For instance, in Table 2, for $K = 8$, we already have two classes whose central elements sound almost identically [ʃ]. Particularly, this occurs if the analyzed speech data is very sparce because the number of different phonemes in the data is limited. The dependency of the voice conversion quality on the class number K is also an issue of future work.

5.2 Automatic Class Mapping

To investigate the work of the proposed class mapping algorithm, we utilized the results of the experiments in Section 5.1. We want to map the classes of the female speaker k_S to those of the male speaker k_T for $K_S = K_T = 8$. Looking at Tables 1 and 2, we cannot find a clear target correspondent for each k_S. Thus, for the assessment of the mapping quality, a distance table between source and target classes based on the phoneme dissimilarities is introduced, v. Table 3. The bold numbers are the minimum distances in the columns. Following distances are used:

Table 3: *Distance Table of the Phonemes Representing Source and Target Classes for $K_S = K_T = 8$*

	k_S	1	2	3	4	5	6	7	8
k_T		ɛ̃	ɒ	ɔ̃	ʃ	ŋ	ã	ɜ̃	ʃ
1	y	5	4	5	6	5	5	5	6
2	ã	4	4	4	6	**4**	**0**	4	6
3	e	3	4	5	6	5	5	3	6
4	s	6	6	6	4	6	6	6	4
5	œ̃	3	4	**2**	6	**4**	4	2	6
6	a	5	**2**	5	6	5	1	5	6
7	ɜ̃	**2**	5	4	6	**4**	4	**0**	6
8	ʃ	6	6	6	**0**	6	6	6	**0**

0 identical phonemes

1 (de)nasalization

2 related vowels

3 similar vowels or
related vowels and (de)nasalization

4 different vowels/fricatives or
similar vowels and (de)nasalization or
nasal vowel vs. nasal

5 different vowels and (de)nasalization or
nonnasal vowel vs. nasal

6 different phoneme types

When we sum up the column minimum distances for $k_S = 1, \ldots, K_S$ we obtain the optimum mapping score 10.

In the following experiment, we study how far we can approach this optimum score, and how the mapping quality depends on the diagonality constant Π (cp. Section 3) which has important influence on the DFW and, consequently, on the mapping score. This score is defined as the sum of the distances between k_S and $\hat{k}_T(k_S)$ taken from Table 3. The dependence between mapping score and Π is displayed in Figure 4.

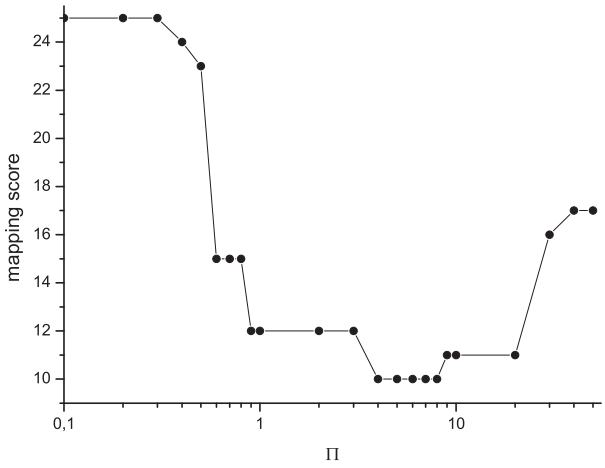

Figure 4: *Dependence of the Mapping Quality on the DFW Diagonality Constant* Π

This diagram shows that at least in our experiment the optimum score is reached by choosing the diagonality constant in the area of $\Pi = [4, 8]$. Further experiments should deal with the question if the optimum adjusting of Π depends on other parameters (like signal energy) or characteristics of the speakers to be compared.

6 Conclusions

In this paper, we have presented an automatic class segmentation and mapping technique for artificial phonetic classes. It is designed for parameter training of cross-language voice conversion. As an example, a simple voice conversion method with one trained parameter based on vocal tract length normalization has been explicated. Finally, experimental results of the proposed class segmentation and mapping technique have been shown.

7 Acknowledgements

This research was supported by a scholarship of the company Siemens AG in Munich, Germany. We particularly thank Harald Höge for his contribution to this work.

References

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

W. A. Ainsworth, K. K. Paliwal, and H. M. Foster. 1984. Problems with Dynamic Frequency Warping as a Technique for Speaker-Independent Vowel Classification. *Proc. of the Institute of Acoustics, 6(4)*.

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

Y. Gao and A. Waibel. 2002. Speech-to-Speech Translation. *Proc. of the ACL'02 Workshop on Speech-to-Speech Translation*. Philadelphia, USA.

W. Hess. 1983. Pitch Determination of Speech Signals. Springer Verlag. New York, USA.

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

T. Kamm, G. Andreou, and J. Cohen. 1995. Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. In *Proc. of the 15th Annual Speech Research Symposium*. Baltimore, USA.

M. Kay, J. M. Gawron, and P. Norvig. 1994. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI Publications. Stanford, USA.

K. S. Lee, D. H. Youn, and I. W. Cha. 1996. A New Voice Transformation Method Based on Both Linear and Nonlinear Prediction Analysis. In *Proc. of the ICSLP'96*. Philadelphia, USA.

M. Mashimo, T. Toda, K. Shikano, and N. Campbell. 2001. Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT. In *Proc. of the EUROSPEECH'01*. Aalborg, Denmark.

H. Matsumoto and H. Wakita. 1986. Vowel Normalization by Frequency Warped Spectral Matching. *Speech Communication, 5(2)*.

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

S. Molau, S. Kanthak, and H. Ney. 2000. Efficient Vocal Tract Normalization in Automatic Speech Recognition. In *Proc. of the ESSV'00*. Cottbus, Germany.

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

M. Pitz and H. Ney. 2003. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. Submitted to *IEEE Trans. on Speech and Audio Processing*.

Y. Stylianou, O. Cappé and E. Moulines. 1998. Continuous Probabilistic Transform for Voice Conversion. *IEEE Trans. on Speech and Audio Processing, 6(2)*.

D. Talkin. 1995. A Robust Algorithm for Pitch Tracking (RAPT). *Speech Coding and Synthesis*. Elsevier Science. Amsterdam, Netherlands.

M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. 1998. Speaker Adaptation for HMM-Based Speech Synthesis System Using MLLR. In *Proc. of the 3th ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

M. Tang, C. Wang, and S. Seneff. 2001. Voice Transformations: From Speech Synthesis to Mammalian Vocalizations. In *Proc. of the EUROSPEECH'01*. Aalborg, Denmark.

O. Türk. 2003. New Methods for Voice Conversion. *PhD Thesis*. Boğaziçi University, Istanbul, Turkey.

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

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