Voice Conversion Using Exclusively Unaligned Training Data

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: Although all conventional voice conversion approaches require equivalent training utterances of source and target speaker, several recently proposed applications call for breaking this demand. In this paper, we present an algorithm which finds corresponding time frames within unaligned training data. The performance of this algorithm is tested by means of a voice conversion framework based on linear transformation of the spectral envelope. Experimental results are reported on a Spanish cross-gender corpus utilizing several objective error measures.

Keywords: voice conversion, unaligned training data, linear transformation of the spectral envelope

1 Introduction

Voice conversion is the adaption of the characteristics of a source speaker's voice to those of a target speaker (Moulines and Sagisaka, 1995). Over the last few years, the interest in voice conversion has risen immensely. This is due to its application to the individualization of text-to-speech systems, whose voices, in general, have to be created in a rather time-consuming way requiring human assistance (Kain and Macon, 1998).

Conventional voice conversion techniques demand equivalent utterances of source and target speaker as training material which can be automatically aligned by dynamic time warping (Stylianou, Cappé, and Moulines, 1995). This procedure is necessary since the training algorithms require corresponding time frames for feature extraction.

Even more complicated is applying voice conversion to speech-to-speech translation, nowadays one of the most challenging tasks of speech and language processing (Gao and Waibel, 2002). Here, the aim is that the standard voice of the text-to-speech module speaking a target language is converted to that of the input speaker using a source language. Hence, for training, one of them has to utter the training sentences in the other's language and, for testing, we even need bilingual databases of both speakers (Mashimo et al., 2001).

The pre-condition of having equivalent utterances is inconvenient and, often, results in expensive manual work, since, new speech material must be recorded or bilingual speakers are required.

Therefore, in Section 2, we propose a new algorithm which finds corresponding time frames within unaligned training data. As an example, this algorithm is embedded into a well-studied voice conversion framework based on linear transformation of the spectral envelope (Stylianou, Cappé, and Moulines, 1995). This technique is briefly described in Section 3. Finally, in Section 4, experimental results are reported on a Spanish cross-gender corpus utilizing several objective error measures.

2 *On Finding Corresponding Time Frames within Unaligned Speech Data*

In conventional voice conversion training, we need equivalent utterances of source and target speaker that should feature a high degree of natural time alignment and a similar pitch contour (Kain and Macon, 1998). Through applying dynamic time warping, we finally obtain a reasonable mapping between the time frames of the speech data, which means that corresponding frames represent equivalent phonetic units.

In case we do not have this time alignment but distinct utterances, we are able to find corresponding artificial phonetic classes by means of a straight-forward approach proposed by Sündermann, Ney, and Höge (2003). As this technique only provides one frame pair per phonetic class, it is only helpful if a small number of parameters is to be estimated. The authors utilized it to determine up to 64 parameters for describing the warping function of VTLN-based voice conversion, but they stated that the naturalness of the output speech suffers for parameter numbers greater than eight.

Describing the characteristics of a speaker's voice more exactly seems to require essentially more degrees of freedom than in the case of VTLN-based voice conversion (Türk, 2003). For instance, Toda et al. (2000) reported for their voice conversion system based on a Gaussian mixture model (GMM) and linear transformation in cepstral space up to 64 GMM components, 40-dimensional feature vectors and full covariance matrices. This large number of parameters could only be reliably estimated by being provided about 64 sentences of time-aligned training data.

Consequently, the baseline algorithm for finding corresponding artificial phonetic classes needs to be extended in order to obtain frame pairs which are comparable to the time alignment paradigm concerning their number and reliability.

In the following, we describe the preprocessing of the speech data and its segmentation into artificial phonetic classes, the mapping between classes of source and target speaker and the extraction of corresponding time frames.

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). However, as we have argued in the introduction, the extraction of pitch marks should not be done neither supervised nor utilizing additional equipment as, for instance, a laryngograph. Therefore, we used the fully automatic pitch mark extractor developed by Goncharoff and Gries (1998). In order to assess the performance of this algorithm, we tested its output in comparison to manually corrected pitch marks which were generated using the laryngograph signal, cf. Section 4.2.

After extracting the pitch marks of a given speech signal, we split it at these marks obtaining frames of different lengths. In voiced regions, the frame lengths depend on the fundamental frequency, in unvoiced regions, the pitch extraction algorithm utilizes a mean approximation.

By applying discrete Fourier transformation to the frames, we obtain complex-valued spectra which still have distinct lengths. Since the algorithms described in this paper require spectra of uniform dimensionality, we normalize the spectrum lengths by means of complex cubic spline interpolation to the maximum spectrum length of all frames (Unser, Aldroubi, and Eden, 1993). In the following, these unidimensional complex spectra are referred to as X.

2.2 Automatic Segmentation

Now, we are ready to distribute the set of unidimensional spectra among K well-distinct classes which can be regarded as artificial phonetic classes. This is done by clustering the amplitude spectra with the help of the k-means algorithm using the squared Euclidean distance as discrimination criterion (Spath, 1985). K-means delivers the class members as well as their centroid spectra $\bar{X}_k$. In this step, the phase information is neglected as it does not seem to be of importance assigning the spectra to the respective classes.

2.3 Class Mapping

During training, we first preprocess and segment the given speech material of source and target speaker as explained above. We get the source centroids $\bar{X}_k$ and the target cen-

troids $\bar{Y}_l$. Now, for each target class l, we want to know the corresponding source class $k(l)$. When comparing spectral vectors of different speakers, it is helpful to compensate for the effect of speaker-dependent vocal tracts. In particular, this compensation is important for cross-gender speech comparisons. This is done by using dynamic frequency warping and, afterwards, we are allowed to assess the similarity of two classes by means of the Euclidean distance (Matsumoto and Wakita, 1986):

$$k(l) = \arg \min_{\kappa=1,\ldots,K} D_{\mathrm{DFW}}(\bar{X}_\kappa, \bar{Y}_l) \ .$$

Here, D_{DFW} is the distance between the frequency-aligned spectra derived from $\bar{X}_\kappa$ and $\bar{Y}_l$ by dynamic frequency warping.

2.4 Extracting Corresponding Time Frames

Once we have mapped one source cluster to each target cluster, we can shift the latter in such a way that each centroid $\bar{Y}$ coincide with the corresponding source centroid $\bar{X}$. Finally, for each shifted target cluster member $Y' = Y - \bar{Y} + \bar{X}$, we determine the nearest member of the mapped source class, X, using the Euclidean distance. The desired spectrum pairs consist of the respective unshifted target spectra Y and the determined corresponding source spectra X:

$$X = \arg \min_{\chi} |\chi - Y - \bar{X} + \bar{Y}| \ .$$

3 Voice Conversion Based on Linear Transformation

Already in the middle of the 90s, Stylianou, Cappé, and Moulines (1995) presented a method for statistical learning of the correspondence between spectral parameters measured from two different speakers uttering the same text. This approach and its extension by Kain and Macon (1998) has been adopted by most people dealing with voice conversion nowadays, cf. e.g. (Mashimo et al., 2001) or (Ye and Young, 2003).

In the following, we briefly explain the basic idea of linear-transformation-based voice conversion and describe how we get from time to feature space and vice versa.

3.1 The Main Concept

Let x_1^M be a sequence of M training feature vectors (whose nature is to be explained in Section 3.2) which characterizes speech of the source speaker and y_1^M the equivalent of the target speaker. Then, we use the combination of these sequences $z_1^M = \binom{x_1}{y_1}, \ldots, \binom{x_M}{y_M}$ to estimate the parameters of a GMM $(\alpha_i, \mu_i, \Sigma_i)$ with I components for the joint density $p(x, y)$ (Kain and Macon, 1998).

In the operation phase, a target feature vector y is derived from a source vector x by the conversion function which minimizes the mean squared error between the converted source and target vectors processed in training:

$$y = \sum_{i=1}^{I} p(i|x) \cdot (\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx\,-1} (x - \mu_i^x)) \ , \quad (1)$$

$$\text{where} \quad p(i|x) = \frac{\alpha_i N(x|\mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{I} \alpha_j N(x|\mu_j^x, \Sigma_j^{xx})} \quad \text{and}$$

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} ; \ \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} .$$

3.2 From Time to Feature Space

As explained in Section 2.1, we consider the spectra derived from pitch-synchronous time frames to be unidimensional. In general, the dimensionality of the latter is too high (> 200) to be directly processed by the above training algorithms. This is due to problems estimating the full covariance matrices.

In literature, we find several feature representations which reduce the number of dimensions to between 15 and 40 features, e.g. line spectral frequencies (Kain and Macon, 1998) or mel frequency cepstral coefficients (MFCC) (Toda et al., 2000). A recently proposed feature set is based on a spectral interpolation based on cubic splines whose interpolation points are mel-frequency-distributed (Ye and Young, 2003). The authors stated that this representation outperforms the MFCC approach. Since our experiments confirmed this outcome, in the following, we will utilize the mel frequency spline interpolation of the amplitude spectrum. Here, the phase spectrum is neglected.

3.3 From Feature to Time Space

In operation phase, the linear transformation described in Eq. 1 delivers a sequence of con-

verted vectors. This sequence can be transformed to the spectral domain by reapplying cubic spline interpolation.

Computing the features from the complex-valued spectra removed the phase information which is significant for the perceptive sound quality, cf. above. A trick to generate the output phase is to simply add the input phase spectrum, as, often, phase manipulation deteriorates the naturalness of the converted speech.

Once we have produced the unidimensional output spectra, we want to deal with the transformation to the time domain.

During training, we were able to derive the mean fundamental frequency (f_0) ratio by comparing the lengths of the time frames of source and target speaker. In operation phase, we take the f_0 trajectory of the source utterance and divide it by this ratio obtaining a simple approximation of the target speaker's f_0 trajectory.

Then, we adapt the length of the corresponding unidimensional spectrum accordingly by again using cubic spline interpolation, cf. Section 2.1. Finally, we apply frequency domain pitch-synchronous overlap and add (FD-PSOLA) to return to time space, taking into account that frames must be skipped or repeated, respectively, in order to preserve the speaking rate (Kleijn and Paliwal, 1995).

4 Experiments

After describing the characteristics of our experimental corpus, we control the performance of the algorithm for pitch mark extraction which is requirement for unsupervised pitch-synchronous speech modification, cf. Section 2.1. Then, we address ourselves to several objective error criteria which, finally, are used to assess the voice conversion performance using aligned and unaligned training data.

4.1 The Experimental Corpus

The corpus utilized in this work contains several hundred Spanish sentences uttered by a female and a male speaker. The speech and laryngograph signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16 kHz. The supervised pitch labeling was done by phoneticians using the pitch tracker developed by Talkin (1989) and, in addition, the laryngograph signal. These manually corrected pitch marks serve as reference for the investigations regarding the automatic pitch mark extraction.

gross error [%]	**male**	**female**
Goncharoff and Gries	3.8	3.4
Bagshaw et al.	6.9	3.5

Table 1: Evaluation of the pitch mark extraction algorithm.

4.2 Automatic Pitch Mark Extraction

Pitch tracking and pitch mark extraction are well-studied fields of signal processing (Hess, 1983). From our experience in speech synthesis we know that pitch segmentation errors often result in distortions or artifacts. Therefore, it is important to assess the accuracy of the utilized pitch mark extraction algorithm to avoid these effects from the very beginning.

The considered *algorithm for accurately marking pitch pulses in speech signals* from Goncharoff and Gries (1998) is a very fast and easily implementable algorithm. However, its performance in terms of accuracy does not seem to be tested adequately. Therefore, we used the provided manually controlled pitch mark data as reference material for evaluation. As error measure, we used the acknowledged gross detection error rate, which equals the percentage of time frames whose fundamental frequency deviates more than 20% from the reference frequency (Rabiner et al., 1976).

The test was performed on 100 sentences of the male speaker and of the female speaker, respectively. Compared with the outcomes of an evaluation of seven pitch trackers, the utilized algorithm's accuracy seems to be state-of-the-art (Bagshaw, Hiller, and Jack, 1993). Table 1 shows the gross error for the implemented algorithm and the average results of the reference. Of course, this is not a hard evaluation as the underlying test data are distinct.

4.3 Objective Error Measures

In the literature dealing with voice conversion, several objective error measures are used. They require reference speech data of the target speaker which is aligned to the source test utterances by dynamic time warping.

The most common measure is the relative spectral distortion D which compares the distance between the converted speech (represented by the vector sequence $\tilde{x}_1^N$) and the reference (y_1^N) with that between source (x_1^N) and reference. From this general definition, one has derived several sub-categories including measuring distances between the feature vectors (Tamura et al., 1998), the magnitude spectra (Sündermann, Ney, and Höge, 2003), or the log spectra (Ye and Young, 2003). These relative distortions are 1.0 for a system which directly passes the source speech to the output without converting it at all. In the case of producing the perfect output, i.e. the reference speech, they are 0. In addition, (Kain and Macon, 1998) have argued that a trivial linear-transformation-based voice conversion system could always predict the mean of the target vectors. This leads to an expression for the spectral distortion with the distance between reference speech and mean target vectors as denominator.

Since the magnitude spectra as well as the spline interpolation features depend on the signal loudness, the spectral distortion varies depending on the signal level of the compared vectors. To avoid this effect, we normalize their energies. However, through this step, deviations in low-energy regions are counted in the same way like those in high energy regions. Therefore, finally, we apply a weighted mean to compute the average spectral distortion. The weights $w_n : n = 1, \ldots, N$ are the normalized geometric means of the compared vectors' signal energies ($E(x)$: signal energy of x; $d(x, y)$: vector distance, cf. Table 2):

$$D = \frac{\sum_{n=1}^{N} w_n(\tilde{x}_1^N, y_1^N) d(\frac{\tilde{x}_n}{\sqrt{E(\tilde{x}_n)}}, \frac{y_n}{\sqrt{E(y_n)}})}{\sum_{n=1}^{N} w_n(x_1^N, y_1^N) d(\frac{x_n}{\sqrt{E(x_n)}}, \frac{y_n}{\sqrt{E(y_n)}})}$$

$$\text{with} \quad w_n(x_1^N, y_1^N) = \frac{\sqrt{E(x_n)E(y_n)}}{\sum_{\nu=1}^{N} \sqrt{E(x_\nu)E(y_\nu)}}.$$

4.4 Comparative Evaluation

Due to the novelty of the presented algorithm for voice conversion parameter training using unaligned data, we only present results which consider a limited amount of training data, namely ten sentences of the male speaker and of the female speaker, respectively. For testing, we used the same amount of data but, of course, a set of different sentences.

	compared vectors	$d(x, y)$
Tamura	spline features	$\sqrt{E(x-y)}$
Kain	spline features	$E(x-y)$
Sündermann	magnitude spectra	$E(x-y)$
Ye	magnitude spectra	$E(\ln x - \ln y)$

Table 2: Objective error measures: Vector distances

Kain and Macon (1998) demonstrated that increasing the number of GMM components does not show a positive effect for relatively sparse training data. Our experiments have shown that even using two components, in most cases, deteriorate the outcomes. Hence, in this paper we use $I = 1$; consequently, the expression Gaussian mixture model is actually not correct for the present parameter settings. However, in the future, we want to extend the amount of training data and therewith the number of GMM components.

To assess the effects which are caused by the feature representation, at the beginning, we measured the distortion which results from transforming the reference speech to feature space and back and then regarding the result as being the converted speech (for our experiments, we used 20 features). Although one would expect to obtain a zero distortion at least for the error criteria based on feature vectors, we noted that the multitude of executed spline interpolations, f_0 adaption, Hamming windowing (as a part of the FD-PSOLA technique) cause considerable distortions, cf. Sections 3.2 and 3.3. In Table 3, they are referred to as *initial* distortions.

4.5 Interpretation

The outcomes of these initial experiments show that

- the results of the voice conversion technique using aligned data are comparable with those reported in the literature, cf. e.g. (Kain and Macon, 1998). In other words, our baseline system shows state-of-the-art performance.
- The relative deterioration by using exclusively unaligned training data is

D [%]	**Tamura**	**Sünder-mann**	**Ye**	**Kain**
initial	0.13	0.12	0.20	0.02
aligned	0.71	0.55	0.58	0.42
unaligned	0.78	0.66	0.60	0.53
initial	0.18	0.14	0.31	0.02
aligned	0.75	0.56	0.92	0.38
unaligned	0.87	0.71	1.00	0.50

Table 3: Comparative evaluation between voice conversion using aligned and unaligned training data. Top: male-to-female; bottom: female-to-male

around 15% for male-to-female conversion and around 25% for female-to-male. Nevertheless, as a starting point, these results are rather satisfactory since, so far, we have used only a simple implementation which is to be optimized and developed further in the future. For instance, we intuitively chose $K = L = 8$ source and target classes for the k-means clustering without controlling the significance of this decision.

Besides, utilizing the pure spectra for the clustering used for our segmenation and mapping algorithm might not be the ideal choice. For instance, MFCCs could be a better representation of the phonetic content of the compared spectra. Another improvement can be expected by using a probabilistic model like a GMM instead of the hard k-means clustering method.

- The most distinctive error measure seems to be that of Kain and Macon (1998). It reports only two percent initial distortion, which is rather closed to the expected zero distortion. The relative differences between initial distortion and that of the aligned training method and that between both training methods are the highest in comparison with the other criteria.

5 Conclusion

In this paper, an algorithm for voice conversion parameter training which finds corresponding time frames within exclusively unaligned training data is presented. It is tested in comparison with the conventional method of using equivalent training utterances. The outcomes show an relative deterioration of around 15% for male-to-female voice conversion and 25% for the other direction. These initial results are satisfactory because of the importance of voice conversion applications where aligned training data is not available. The presented system is not optimized yet and serves as a good starting point for intensive investigations regarding its accuracy in the future.

References

Bagshaw, P. C., S. M. Hiller, and M. A. Jack. 1993. Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching. In *Proc. of the Eurospeech'93*, Berlin, Germany.

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

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

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

Kleijn, W. B. and K. K. Paliwal. 1995. *Speech Coding and Synthesis*. Elsevier Science B.V., Amsterdam, Netherlands.

Mashimo, M., 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.

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

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

Rabiner, L. R., M. J. Cheng, A. E. Rosenberg, and A. McGonegal. 1976. A Comparative Study of Several Pitch Detection Algorithms. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 23.

Spath, H. 1985. *Cluster Dissection and Analysis*. Halsted Press, New York, USA.

Stylianou, Y., O. Cappé, and E. Moulines. 1995. Statistical Methods for Voice Quality Transformation. In *Proc. of the Eurospeech'95*, Madrid, Spain.

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

Talkin, D. 1989. Looking at Speech. *Speech Technology*, 4.

Tamura, M., 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.

Toda, T., J. Lu, H. Saruwatari, and K. Shikano. 2000. Straight-Based Voice Conversion Algorithm Based on Gaussian Mixture Model. In *Proc. of the ICSLP'00*, Beijing, China.

Türk, O. 2003. *New Methods for Voice Conversion*. Ph.D. thesis, Boğaziçi University, Istanbul, Turkey.

Unser, M., A. Aldroubi, and M. Eden. 1993. B-Spline Signal Processing. *IEEE Trans. on Signal Processing*, 41(2).

Ye, H. and S. Young. 2003. Perceptually Weighted Linear Transformations for Voice Conversion. In *Proc. of the Eurospeech'03*, Geneva, Switzerland.