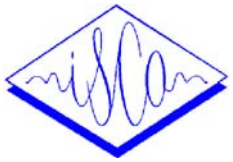
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A First Step Towards Text-Independent Voice Conversion

David Sündermann, Antonio Bonafonte

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
C/ Jordi Girona, 1 i 3, 08034 Barcelona, Spain
{suendermann,antonio}@gps.tsc.upc.es

Hermann Ney

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

Harald Höge

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

Abstract

So far, all conventional voice conversion approaches are text-dependent, i.e., they need equivalent training utterances of source and target speaker. Since several recently proposed applications call for renouncing this requirement, in this paper, we present an algorithm which finds corresponding time frames within text-independent 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.

1. Introduction

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker. 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 [1].

Conventional voice conversion techniques are text-dependent. I.e., they need equivalent utterances of source and target speaker as training material which can be automatically aligned by dynamic time warping [2]. This procedure is necessary since the training algorithms require corresponding time frames for feature extraction.

An even more challenging task is voice conversion for speech-to-speech translation, nowadays one of the most challenging tasks of speech and language processing [3]. Here, the aim is the conversion of the standard voice of the text-to-speech module speaking a target language to the voice 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 utterances of both speakers [4].

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

Therefore, in Section 2, we propose an algorithm which finds corresponding time frames within text-independent training data. As an example, this algorithm is embedded into a well-studied voice conversion framework based on linear transformation of the spectral envelope [2]. 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 [1]. 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 corresponding 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 in [5]. 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. For instance, [4] 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 co-

variance 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 text-dependent paradigm concerning their number and reliability.

In the following, we describe the preprocessing of the speech signal 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 [1].

To extract pitch-synchronous frames from a given speech signal, we use the algorithm described in [6]. 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 without zero padding to the frames, we obtain complex-valued spectra with distinct numbers of spectral lines. Since the algorithms described in this paper require spectra of the same number of spectral lines, we normalize them by means of complex cubic spline interpolation to the maximum number of spectral lines of all frames. In the following, these spectra are referred to as X.

2.2. Automatic Segmentation

Now, we are ready to distribute the set of spectra among K well-distinct classes which can be regarded as artificial phonetic classes. This is done by clustering the magnitude spectra with the help of the k-means algorithm using the squared Euclidean distance as discrimination criterion. K-means delivers the class members as well as their centroid spectra $\bar{X}_k$.

2.3. Class Mapping

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

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

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:

$$X = \arg \min_{\chi} |\chi - Y - \bar{X} + \bar{Y}| \,. \tag{2}$$

3. Voice Conversion Based on Linear Transformation

Already in the middle of the 90s, [2] 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 [1] has been adopted by most people dealing with voice conversion nowadays, cf. e.g. [4] or [7].

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 $z_1^M = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_M \\ y_M \end{pmatrix}$ to estimate the parameters of a GMM $(\alpha_i, \mu_i, \Sigma_i)$ with I components for the joint density $p(x, y)$ [1].

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:

$$y = \sum_{i=1}^{I} p(i|x) \cdot (\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx\,-1} (x - \mu_i^x)) \,, \tag{3}$$

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

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

3.2. From Time to Feature Space

As explained in Section 2.1, we consider the spectra derived from pitch-synchronous time frames to have the same number of spectral lines. In general, the dimensionality of these spectral vectors 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 [1] or mel frequency cepstral coefficients (MFCC) [4]. A recently proposed feature set is based on a spectral interpolation by means of cubic splines whose interpolation points are mel-frequency-distributed [7]. 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 magnitude spectrum. Here, the phase spectrum is neglected.

3.3. From Feature to Time Space

In operation phase, the linear transformation described in Eq. 3 produces a sequence of converted 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 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 voiced 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 number of spectral lines 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 [8].

4. Experiments

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 signals were recorded in an acoustically isolated environment and sampled at a sample frequency of 16 kHz.

4.2. 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 [9], the magnitude spectra [5], or the log spectra [7]. 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, [1] 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.

	compared vectors	$d(x,y)$
Tamura et al. [9]	spline features	$\sqrt{E(x-y)}$
Kain and Macon [1]	spline features	$E(x-y)$
Sündermann et al. [5]	magnitude spectra	$E(x-y)$
Ye and Young [7]	magnitude spectra	$E(\ln x - \ln y)$

Table 1: *Objective error measures: Vector distances.*

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 1):

$$D = \frac{\sum_{n=1}^{N} w_n(\tilde{x}_1^N, y_1^N) d(\frac{\tilde{x}_n}{\sqrt{E(\tilde{x}_n)}}, \frac{y_n}{\sqrt{E(y_n)}})}{\sum_{n=1}^{N} w_n(x_1^N, y_1^N) d(\frac{x_n}{\sqrt{E(x_n)}}, \frac{y_n}{\sqrt{E(y_n)}})} \quad (4)$$

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

4.3. Comparative Evaluation

In our experiments, we investigated the influence of the amount of training data (one to 64 sentences) and that of the number of GMM components (one to eight) on the performance of the conventional training approach using text-dependent training data and that based on text-independent data, cf. Figure 1. For testing, ten sentences were used, which were, of course, distinct from the training material.

To assess the effects which are caused by the feature representation, at the beginning, we measured the *initial* 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), cf. Table 2. Although one

D	Tamura	Kain	Sündermann	Ye
initial distortion	0.13	0.02	0.12	0.20
text-dependent	0.68	0.39	0.57	0.49
text-independent	0.75	0.47	0.65	0.55
initial distortion	0.18	0.02	0.14	0.31
text-dependent	0.76	0.38	0.58	0.76
text-independent	0.87	0.49	0.74	0.92

Table 2: *Comparative evaluation between voice conversion using text-dependent and text-independent training data.* The table contains the best results of the respective technique i.e., 64 training sentences and eight GMM components for text-dependent training data; two training sentences and one GMM component for the text-independent case. Top: male-to-female; bottom: female-to-male

would expect to obtain an initial distortion of zero, this cannot even be achieved for the error criteria based on feature vectors, as 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.

4.4. Interpretation

These initial experiments show that

• the results of the voice conversion technique using text-dependent data are comparable with those reported in the literature, cf. e.g. [1]. In other words, our baseline system shows state-of-the-art performance.

• The relative deterioration by using text-independent training data is around 14% for male-to-female conversion and around 23% for female-to-male, cf. Table 2. Besides, in Figure 1, we note that already for two training sentences, the text-independent training technique produces a saturation effect; and using more than one GMM component does not improve the performance at all. This might be due to the fixed number of $K = L = 8$ source and target classes for the k-means clustering. 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.

• The most distinctive error measure seems to be that of [1]. It reports only two percent initial distortion, which is rather closed to the expected zero distortion. Besides, the relative differences between initial distortion and that of the text-dependent 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 text-independent 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 14% for male-to-female voice conversion and 23% for the other direction. These initial results are satisfactory because of the importance of voice conversion applications where text-dependent 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.

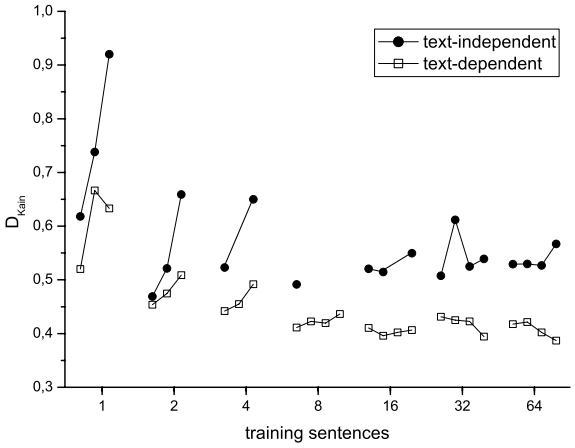


Figure 1: *Dependency of the voice conversion performance on the number of training sentences and that of the GMM components.* The figure shows results for male-to-female conversion based on the error measure according to [1]. For each set of training sentences, from the left to the right, the number of GMM components is 1, 2, 4, 8.[1]

6. References

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

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical Methods for Voice Quality Transformation," in *Proc. of the Eurospeech'95*, Madrid, Spain, 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] 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.

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

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

[7] H. Ye and S. J. Young, "Perceptually Weighted Linear Transformations for Voice Conversion," in *Proc. of the Eurospeech'03*, Geneva, Switzerland, 2003.

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

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

[1] Especially for small numbers of training sentences, we face numerical problems estimating the full covariance matrices. Therefore, some of the curve points are missing.