May 2004

Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation

Athanasios Mouchtaris
University of Pennsylvania

Jan Van der Spiegel
University of Pennsylvania, jan@seas.upenn.edu

Paul Mueller
Corticon, Inc.

Recommended Citation

Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, "Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation", . May 2004.

Copyright 2004 IEEE. Reprinted from *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP 2004)* Volume 1, pages I-1 - I-4.
Publisher URL:http://ieeexplore.ieee.org/xpl/tocresult.jsp?isNumber=29343

Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation

Abstract

The objective of voice conversion methods is to modify the speech characteristics of a particular speaker in such manner, as to sound like speech by a different target speaker. Current voice conversion algorithms are based on deriving a conversion function by estimating its parameters through a corpus that contains the same utterances spoken by both speakers. Such a corpus, usually referred to as a parallel corpus, has the disadvantage that many times it is difficult or even impossible to collect. Here, we propose a voice conversion method that does not require a parallel corpus for training, *i.e.* the spoken utterances by the two speakers need not be the same, by employing speaker adaptation techniques to adapt to a particular pair of source and target speakers, the derived conversion parameters from a different pair of speakers. We show that adaptation reduces the error obtained when simply applying the conversion parameters of one pair of speakers to another by a factor that can reach 30% in many cases, and with performance comparable with the ideal case when a parallel corpus is available.

Keywords

voice conversion, gaussian mixture model, text-to-speech synthesis, speaker adaptation

Comments

Copyright 2004 IEEE. Reprinted from *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP 2004)* Volume 1, pages I-1 - I-4.
Publisher URL:http://ieeexplore.ieee.org/xpl/tocresult.jsp?isNumber=29343

NON-PARALLEL TRAINING FOR VOICE CONVERSION BY MAXIMUM LIKELIHOOD CONSTRAINED ADAPTATION

*Athanasios Mouchtaris, Jan Van der Spiegel**

Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104

Paul Mueller

Corticon Inc.
155 Hughes Rd.
King of Prussia, PA 19406

ABSTRACT

The objective of voice conversion methods is to modify the speech characteristics of a particular speaker in such manner, as to sound like speech by a different target speaker. Current voice conversion algorithms are based on deriving a conversion function by estimating its parameters through a corpus that contains the same utterances spoken by both speakers. Such a corpus, usually referred to as a parallel corpus, has the disadvantage that many times it is difficult or even impossible to collect. Here, we propose a voice conversion method that does not require a parallel corpus for training, *i.e.* the spoken utterances by the two speakers need not be the same, by employing speaker adaptation techniques to adapt to a particular pair of source and target speakers, the derived conversion parameters from a different pair of speakers. We show that adaptation reduces the error obtained when simply applying the conversion parameters of one pair of speakers to another by a factor that can reach 30% in many cases, and with performance comparable with the ideal case when a parallel corpus is available.

1. INTRODUCTION

Voice conversion methods attempt to modify the characteristics of speech by a given source speaker, so that it sounds as if it was spoken by a different target speaker. Applications for voice conversion include “personalization” of a Text-To-Speech (TTS) synthesis system so that it “speaks” with the voice of a particular person, as well as creating new voices for a TTS system without the need of retraining the system for every new voice. A number of different approaches has been proposed for achieving voice conversion (see [1, 2, 3] and the references therein).

The common characteristic of these approaches is that they focus on the short-term spectral properties of the speech signals, which they modify according to a conversion function designed during the training phase. During training, the parameters of this conversion function are derived based on minimizing some error measure. In order to achieve this however, a speech corpus is needed that contains the same utterances (words, sentences, *etc.*) from both the source and target speakers. The disadvantage of this method is that for many cases it is difficult or even impossible to collect such a corpus. If, for example, the desired source or target speaker is not a person directly available, it is evident that collecting such a corpus would probably be impossible, especially since a large number of data are needed in order to obtain meaningful results. Recently, an algorithm that attempted to address this issue was proposed [4], by concentrating on the phonemes spoken by the two speakers. The objective was to derive a conversion function that can transform the phonemes of the source speaker into the corresponding phonemes of the target speaker, thus not requiring a parallel corpus for training. However, accurately recognizing the phonemes spoken by the two speakers during training, as well as the phonemes spoken by the source speaker during conversion, is essential for this algorithm to operate correctly, and this can be a difficult requirement to meet in practice.

Here we propose a conversion algorithm that relaxes the constraint of using a parallel corpus during training. Our approach is to adapt the conversion parameters for a given pair of source and target speakers, to the particular pair of speakers for which no parallel corpus is available. Referring to Fig. 1, we assume that a parallel corpus is available for speakers A and B (in the left part of the diagram), and for this pair a conversion function is derived by employing one of the conversion methods that are given in the literature [3]. For the particular pair that we focus on, speakers C and D (in the right part of the diagram), a non-parallel corpus is available for training. Our approach is to adapt the conversion function derived for speakers A and B to speakers C and D, and use this new adapted conversion function for these speakers. Adaptation is achieved by relating the non-parallel corpus to the parallel corpus, as shown in the diagram and detailed in the following sections.

2. SPECTRAL CONVERSION

Voice conversion is essentially achieved by spectral conversion. The objective of spectral conversion is to derive a function that can convert the short-term spectral properties of a reference waveform into those of a desired signal. A training dataset is created from the existing reference and the target speech waveforms by applying a short sliding window and extracting the parameters that model the short-term spectral envelope (in this paper we use the line spectral frequencies - LSF’s - due to their desirable interpolation properties [3]). This procedure results in two vector sequences, $[\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n]$ and $[\boldsymbol{y}_1\boldsymbol{y}_2\ldots\boldsymbol{y}_n]$, of reference and target spectral vectors respectively. A function $\mathcal{F}(\cdot)$ can be designed which, when applied to vector $\boldsymbol{x}_k$, produces a vector close in some sense to vector $\boldsymbol{y}_k$. Recent results have clearly demonstrated the superiority of the algorithms based on Gaussian mixture models (GMM’s) for the voice conversion problem [2, 3].

According to GMM-based algorithms, a sequence of spectral vectors $\boldsymbol{x}_k$ as above can be considered as a realization of a random vector $\boldsymbol{x}$ with probability density function (pdf) as GMM

$$\mathrm{g}(\boldsymbol{x}) = \sum_{i=1}^{M} p(\omega_i)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx}), \qquad (1)$$

where, $p(\omega_i)$ is the prior probability of class ω_i, and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The parameters of the GMM, *i.e.* the mean vectors, covariance matrices and priors, can be estimated using the expectation maximization (EM) algorithm [5].

*This research has been funded by the Catalyst Foundation.

The analysis that follows focuses on the conversion method of [3], which offers great insight as to what the conversion parameters represent. Assuming that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian for each class ω_i, then, in mean-squared sense, the optimal choice for the function $\mathcal{F}$ is

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}_k) &= \mathrm{E}(\boldsymbol{y}|\boldsymbol{x}_k) \\ &= \sum_{i=1}^{M} p(\omega_i|\boldsymbol{x}_k)\left[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\left(\boldsymbol{x}_k - \boldsymbol{\mu}_i^x\right)\right], \end{aligned} \tag{2}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operator and the conditional probabilities $p(\omega_i|\boldsymbol{x}_k)$ are given from

$$p(\omega_i|\boldsymbol{x}_k) = \frac{p(\omega_i)\mathcal{N}(\boldsymbol{x}_k; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{M} p(\omega_j)\mathcal{N}(\boldsymbol{x}_k; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \tag{3}$$

If the source and target vectors are concatenated, creating a new sequence of vectors $\boldsymbol{z}_k$ that are the realizations of the random vector $\boldsymbol{z} = [\boldsymbol{x}^T\boldsymbol{y}^T]^T$ (where ^{T} denotes transposition), then all the required parameters in the above equations can be found by estimating the GMM parameters of $\boldsymbol{z}$. Then,

$$\boldsymbol{\Sigma}_i^{zz} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \ \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}. \tag{4}$$

The EM algorithm is applied to $\boldsymbol{z}$. Since this method estimates the desired function based on the joint density of $\boldsymbol{x}$ and $\boldsymbol{y}$, it is denoted as the JDE method. Note that in order to estimate the GMM of $\boldsymbol{z}$, it is required to correctly align vectors $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ during training, and this can only be achieved when a parallel corpus is used.

The JDE spectral conversion algorithm can be implemented with the covariance matrices having no structural restrictions or restricted to be diagonal, denoted as full and diagonal conversion respectively. Full conversion is of prohibitive complexity when combined with the adaptation algorithm for the non-parallel corpus conversion problem examined in the next section, thus here we concentrate on diagonal conversion. Note that the covariance matrix of $\boldsymbol{z}$ for the JDE method cannot be diagonal because this method is based on the cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$ which is found from (4). This will be zero if the covariance of $\boldsymbol{z}$ is diagonal. Thus, in order to obtain an efficient structure, we must restrict *each* of the matrices $\boldsymbol{\Sigma}_i^{xx}$, $\boldsymbol{\Sigma}_i^{yy}$, $\boldsymbol{\Sigma}_i^{xy}$, and $\boldsymbol{\Sigma}_i^{yx}$ in (4) to be diagonal. For achieving this restriction, the EM algorithm for full conversion must be modified accordingly, and the details can be found in [6].

3. ML CONSTRAINED ADAPTATION

The majority of spectral conversion methods that have been described so far in the literature, including the GMM-based methods, assume a parallel speech corpus for obtaining the spectral conversion parameters for every pair of reference and target speakers. Our objective here is to derive an algorithm that relaxes this constraint. In other words, we propose in this section an algorithm that derives the conversion parameters from a speech corpus in which the reference and target speakers do not necessarily utter the same words or sentences. In order to achieve this result, we apply the maximum-likelihood constrained adaptation method [7], which offers the advantage of a simple probabilistic linear transformation leading to a mathematically tractable solution.

In addition to the pair of speakers for which we intend to derive the non-parallel training algorithm, we also assume that a parallel speech corpus is available for a *different* pair of speakers. From this latter corpus, we obtain a joint GMM model, derived as explained in Section 2. In the following, the spectral vectors that correspond to the reference speaker are considered as realizations of random vector $\boldsymbol{x}$, while $\boldsymbol{y}$ corresponds to the target speaker of the parallel corpus. From the non-parallel corpus, we also obtain a sequence of spectral vectors, considered as realizations of random vector $\boldsymbol{x}^{'}$ for the reference speaker and $\boldsymbol{y}^{'}$ for the target speaker. We then attempt to relate the random variables $\boldsymbol{x}^{'}$ and $\boldsymbol{x}$, as well as $\boldsymbol{y}^{'}$ and $\boldsymbol{y}$, in order to derive a conversion function for the non-parallel corpus based on the parallel corpus parameters.

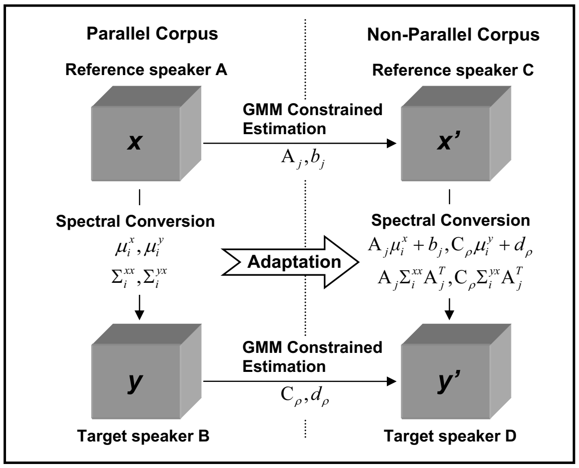

Parallel Corpus
Non-Parallel Corpus
Reference speaker A
Reference speaker C
x
GMM Constrained Estimation
A_j, b_j
x'
Spectral Conversion
μ_i^x, μ_i^y
$\Sigma_i^{xx}, \Sigma_i^{yx}$
Adaptation
Spectral Conversion
$\mathrm{A}_j\mu_i^x + b_j, \mathrm{C}_\rho\mu_i^y + d_\rho$
$\mathrm{A}_j\Sigma_i^{xx}\mathrm{A}_j^T, \mathrm{C}_\rho\Sigma_i^{yx}\mathrm{A}_j^T$
y
GMM Constrained Estimation
C_ρ, d_ρ
y'
Target speaker B
Target speaker D

Fig. 1. Block diagram outlining spectral conversion for a parallel and non-parallel corpus. In the latter case, spectral conversion is preceded by adaptation of the derived parameters from the parallel corpus to the non-parallel corpus.

We assume that the target random vector $\boldsymbol{x}^{'}$ is related to reference random vector $\boldsymbol{x}$ by a probabilistic linear transformation

$$\boldsymbol{x}^{'} = \begin{cases} \mathbf{A}_1\boldsymbol{x} + \boldsymbol{b}_1 & \text{with probability } p(\lambda_1|\omega_i) \\ \mathbf{A}_2\boldsymbol{x} + \boldsymbol{b}_2 & \text{with probability } p(\lambda_2|\omega_i) \\ \quad\vdots & \qquad\qquad\vdots \\ \mathbf{A}_N\boldsymbol{x} + \boldsymbol{b}_N & \text{with probability } p(\lambda_N|\omega_i). \end{cases} \tag{5}$$

This equation corresponds to the GMM constrained estimation that relates $\boldsymbol{x}^{'}$ with $\boldsymbol{x}$ in the block diagram of Fig. 1. In the above equation, $\mathbf{A}_j$ denotes a $K \times K$ dimensional matrix (K is the number of components of vector $\boldsymbol{x}$), and $\boldsymbol{b}_j$ is a vector of the same dimension with $\boldsymbol{x}$. Each of the component transformations j is related with a specific Gaussian i of $\boldsymbol{x}$ with probability $p(\lambda_j|\omega_i)$ satisfying

$$\sum_{j=1}^{N} p(\lambda_j|\omega_i) = 1, \quad i = 1, \ldots, M, \tag{6}$$

where M is the number of Gaussians of the GMM that corresponds to the joint vector sequence of the parallel corpus. Clearly,

$$\mathrm{g}(\boldsymbol{x}^{'}|\omega_i, \lambda_j) = \mathcal{N}(\boldsymbol{x}^{'}; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T), \tag{7}$$

resulting in the pdf of $\boldsymbol{x}^{'}$

$$\mathrm{g}(\boldsymbol{x}^{'}) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(\omega_i)p(\lambda_j|\omega_i)\mathcal{N}(\boldsymbol{x}^{'}; \mathbf{A}_j\boldsymbol{\mu}_i^x + \boldsymbol{b}_j, \mathbf{A}_j\boldsymbol{\Sigma}_i^{xx}\mathbf{A}_j^T). \tag{8}$$

In similar manner, we relate the random vectors $\boldsymbol{y}^{'}$ and $\boldsymbol{y}$ by another probabilistic linear transformation

$$\boldsymbol{y}^{'} = \begin{cases} \mathbf{C}_1\boldsymbol{y} + \boldsymbol{d}_1 & \text{with probability } p(\kappa_1|\omega_i) \\ \mathbf{C}_2\boldsymbol{y} + \boldsymbol{d}_2 & \text{with probability } p(\kappa_2|\omega_i) \\ \quad\vdots & \qquad\qquad\vdots \\ \mathbf{C}_L\boldsymbol{y} + \boldsymbol{d}_L & \text{with probability } p(\kappa_L|\omega_i). \end{cases} \tag{9}$$

The above equation corresponds to the GMM constrained estimation that relates $\boldsymbol{y}^{'}$ with $\boldsymbol{y}$ in the block diagram of Fig. 1. The matrices $\mathbf{A}_j$ and $\mathbf{C}_\rho$, the vectors $\boldsymbol{b}_j$ and $\boldsymbol{d}_\rho$, and the probabilities $p(\omega_i)$, $p(\lambda_j|\omega_i)$, and $p(\kappa_\rho|\omega_i)$, can be estimated by use of the non-parallel corpus and the GMM of the parallel corpus, using maximum likelihood estimation techniques. The EM algorithm can be applied to this case in a similar manner to estimating the parameters of a GMM from observed data. In essence, it is a linearly constrained maximum-likelihood estimation of the GMM parameters. Note that classes ω_i are the same for $\boldsymbol{x}$ and $\boldsymbol{y}$ by design in Section 2. Under this assumption and given the linearity of the transformations (5) and (9), $\boldsymbol{x}^{'}$ and $\boldsymbol{y}^{'}$ will also be jointly Gaussian for a particular class ω_i, λ_j, and κ_ρ, and the pdf of $\boldsymbol{y}^{'}$ will have a similar form with (8). It is now possible to derive the conversion function for the non-parallel training problem, based entirely on the parameters derived from a parallel corpus of a different pair of speakers. Based on the aforementioned assumptions, it holds that

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}^{'}|\boldsymbol{x}^{'}_k, \omega_i, \lambda_j, \kappa_\rho) &= \boldsymbol{\mu}_i^{y^{'}} + \boldsymbol{\Sigma}_i^{y^{'}x^{'}} \boldsymbol{\Sigma}_i^{x^{'}x^{'}-1} \left(\boldsymbol{x}^{'}_k - \boldsymbol{\mu}_i^{x^{'}}\right) \\ &= \mathbf{C}_\rho \boldsymbol{\mu}_i^y + \boldsymbol{d}_\rho + \mathbf{C}_\rho \boldsymbol{\Sigma}_i^{yx} \boldsymbol{\Sigma}_i^{xx^{-1}} \mathbf{A}_j^{-1} \\ &\quad \left(\boldsymbol{x}^{'}_k - \mathbf{A}_j \boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right), \end{aligned} \tag{10}$$

since

$$\boldsymbol{\Sigma}_i^{y^{'}x^{'}} = \mathbf{C}_\rho \boldsymbol{\Sigma}_i^{yx} \mathbf{A}_j^T, \; \boldsymbol{\Sigma}_i^{x^{'}x^{'}} = \mathbf{A}_j \boldsymbol{\Sigma}_i^{xx} \mathbf{A}_j^T, \tag{11}$$

and

$$\boldsymbol{\mu}_i^{y^{'}} = \mathbf{C}_\rho \boldsymbol{\mu}_i^y + \boldsymbol{d}_\rho, \; \boldsymbol{\mu}_i^{x^{'}} = \mathbf{A}_j \boldsymbol{\mu}_i^x + \boldsymbol{b}_j. \tag{12}$$

Finally, the conversion function for the non-parallel case becomes

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}^{'}_k) &= \mathrm{E}(\boldsymbol{y}^{'}|\boldsymbol{x}^{'}_k) \qquad (13) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{\rho=1}^{L} p(\omega_i|\boldsymbol{x}^{'}_k) p(\lambda_j|\boldsymbol{x}^{'}_k, \omega_i) p(\kappa_\rho|\omega_i) \\ &\quad \Big[\mathbf{C}_\rho \boldsymbol{\mu}_i^y + \boldsymbol{d}_\rho + \mathbf{C}_\rho \boldsymbol{\Sigma}_i^{yx} \boldsymbol{\Sigma}_i^{xx^{-1}} \mathbf{A}_j^{-1} \\ &\quad \left(\boldsymbol{x}^{'}_k - \mathbf{A}_j \boldsymbol{\mu}_i^x - \boldsymbol{b}_j\right)\Big], \end{aligned}$$

where

$$p(\omega_i|\boldsymbol{x}^{'}_k) = \frac{p(\omega_i) \sum_{j=1}^{N} p(\lambda_j|\omega_i) \mathrm{g}(\boldsymbol{x}^{'}_k|\omega_i, \lambda_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} p(\omega_i) p(\lambda_j|\omega_i) \mathrm{g}(\boldsymbol{x}^{'}_k|\omega_i, \lambda_j)}, \tag{14}$$

$$p(\lambda_j|\boldsymbol{x}^{'}_k, \omega_i) = \frac{p(\lambda_j|\omega_i) \mathrm{g}(\boldsymbol{x}^{'}_k|\omega_i, \lambda_j)}{\sum_{j=1}^{N} p(\lambda_j|\omega_i) \mathrm{g}(\boldsymbol{x}^{'}_k|\omega_i, \lambda_j)}, \tag{15}$$

and $\mathrm{g}(\boldsymbol{x}^{'}|\omega_i, \lambda_j)$ is given from (7). Thus, all the parameters of the conversion function (13) are known.

4. RESULTS AND DISCUSSION

The spectral conversion method for the case of a non-parallel training corpus that was derived in the previous paragraph is evaluated in this section. As mentioned previously, the spectral vectors used here are the LSF's (22nd order) due to their favorable interpolation properties. It is important to mention that the corpus used is the *VOICES* corpus, available from OGI's CSLU [8, 9]. This is a parallel corpus and is used for both the parallel and non-parallel training cases that are examined in this section, in a manner explained in the next paragraph. The error measure used in this section is the mean-squared error normalized by the initial distance between the reference and target speakers, *i.e.*

$$\mathcal{E} = \frac{\frac{1}{n} \sum_{k=1}^{n} \|\boldsymbol{y}_k - \mathcal{F}(\boldsymbol{x}_k)\|^2}{\frac{1}{n} \sum_{k=1}^{n} \|\boldsymbol{y}_k - \boldsymbol{x}_k\|^2},$$

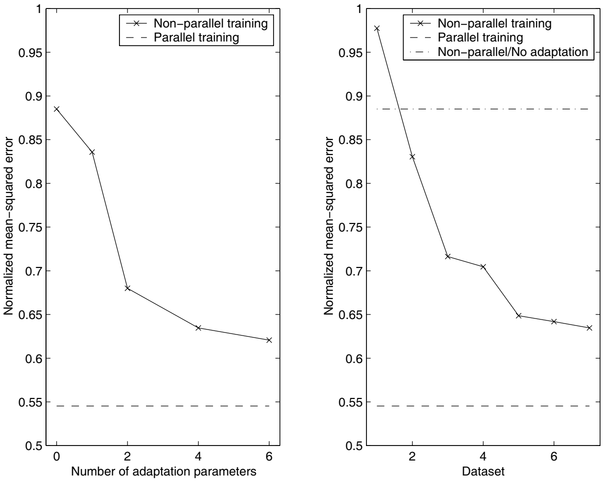
where $\boldsymbol{x}_k$ is the reference vector at instant k, $\boldsymbol{y}_k$ is the target vector at instant k, and $\mathcal{F}(\cdot)$ denotes the conversion function used, which can be the one of (2) or (13) depending whether training is performed in a parallel or non-parallel manner. For all results given in this section, the number of GMM classes for the parameters obtained from the parallel corpus is 16, while the number of vectors for the parallel and the non-parallel training corpus is about 19,000 (denoted here as full corpus), which corresponds to 40 out of the 50 sentences available in the corpus. The results given in this section are the averages of the remaining 10 sentences.

Conversion Method	Normalized Error			
	Test1		Test2	
	None	Adapt.	None	Adapt.
Case 1-A	0.8882	0.6809	1.0264	0.6980
Case 2-A	0.7307	0.6761	0.8342	0.7073
Case 1-B	0.8512	0.6368	1.0371	0.7462
Case 2-B	0.7252	0.6169	0.8850	0.6346
Parallel	0.5221		0.5453	

Table 1. Normalized error for 2 different pairs of parameters derived from a parallel corpus, when applied to 2 different speaker pairs of a non-parallel corpus.

In Table 1, the normalized mean-squared error is given for two different pairs of non-parallel reference and target speakers (Test 1 and Test 2 in the table) for two different adaptation cases (*i.e.* two different pairs of speakers in parallel training, Cases 1-2). The column denoted as "None" in this table corresponds to no adaptation, *i.e.* when the derived parameters from the parallel corpus are directly applied to the speaker pair from the non-parallel corpus, while the column "Adapt." corresponds to the conversion function of (13), for 4 adaptation parameters for both the reference and the target speaker ($L = N = 4$). The last row of the table gives the error when the conversion parameters are derived by parallel training (*i.e.* the ideal case). This table shows the performance of our algorithm for two different choices of the training corpus. For the first one (Cases 1-A and 2-A), the corpus for the parallel pair (speakers A and B) is chosen to be sentences 1-10 of the full corpus, while for adaptation, sentences 11-25 for relating speaker C with speaker A and sentences 26-40 for relating speaker D with speaker B (see Fig. 1). This means that all sentences are different for the different tasks. For the second choice of corpus (Cases 1-B and 2-B), the full training corpus is used for all tasks. Inevitably for this latter case, the sentences in parallel and non-parallel training will be the same. In parallel training, the fact that the same sentences are used is essential since the reference and target vectors are aligned, and this vector-to-vector correspondence is required during training. On the contrary, for non-parallel training the corpus is used as explained here for adaptation of the spectral conversion parameters, thus the fact that the corpus was created in a parallel manner is not exploited and is not expected to influence the results. The results 1-A and 2-A, derived with different sentences as explained, are included in order to further support this argument. We tested the performance of the algorithm with a variety of speaker pairs, using 10 out of the 12 speakers in the corpus, but here only some representative results are given due to space limitations.

It is apparent from Table 1 that the adaptation methods proposed result in a large error decrease compared to simply applying the conversion parameters of a given pair to a different pair of speakers. This improvement can reach the level of 30% when the initial distance is large, which is exactly what is desired. This is



(a) (b)

Fig. 2. Normalized error (a) when using different number of adaptation parameters (0 corresponds to no adaptation), and (b) for various choices of training dataset (see Table 2). The dashed line corresponds to the error when a parallel corpus is used for training. The dashed-dotted line corresponds to no adaptation.

true *both* when the sentences are different or the same (Cases 1-A & 2-A *vs.* 1-B & 2-B) and this supports our previous argument. The performance for the latter cases is on the average better compared to the former, due to the fact that when the full corpus is used for adaptation, more vectors are available and adaptation is more accurate (40 *vs.* 15 sentences, see also Fig. 2(b) discussed later in this section). The performance that we obtain when the conversion parameters are derived by parallel training is always better, compared with non-parallel training (although in most cases the two are comparable). This is an expected and intuitive result since in parallel training we exploit a particular advantage of the speech corpus which is not available in a non-parallel corpus. The methods proposed here intend to address the lack of a parallel corpus and are suitable only for this case. The error does not seem to display any particular patterns when no adaptation is performed, but it interesting that in most cases we examined the initial distance is decreased (*i.e.* error less than one). In future work we intend to further analyze this issue using a larger number of data.

In Fig. 2(a), the performance of the algorithm for a different number of adaptation parameters is shown, using the full corpus both for parallel and non-parallel training. The number of adaptation parameters that is given is the same for the adaptation of the reference speaker and that of the target speaker, although a different number can be used for each case. Adaptation of 0 parameters in this figure corresponds to the case when no adaptation of the parameters is performed. From this figure it is evident that, as expected, there is a significant error decrease when increasing the number of adaptation parameters, since this corresponds to a more accurate modeling of the statistics of the spectral vectors. On the other hand, when increasing the number of adaptation parameters above 4, the error remains approximately constant, concluding that this number of parameters is sufficient to model the statistics of the spectral vectors and further increase does not offer any advantage.

Dataset	1	2	3	4	5	6	7
kVectors	0.25	0.5	1	2.5	5	10	19

Table 2. Number of vectors (thousands) in non-parallel training for the datasets in Fig. 2(b).

In Fig. 2(b), the performance of the algorithm is given for different sizes of the non-parallel corpus, using the full corpus for parallel training, and 4 adaptation parameters for both the reference and target speaker. The dataset numbers in the figure correspond to the numbers of vectors given in Table 2. The error when no adaptation is used (dashed-dotted line), as well as when the corpus is used in a parallel manner (dashed line), is also shown. From this figure we can see that there is a significant error decrease when the size of the corpus is increased. As is the case for the parallel corpus [3], the error decrease is less significant when the size of the corpus increases above 5,000 - 10,000 vectors.

5. CONCLUSIONS

Current voice conversion algorithms require a parallel speech corpus that contains the same utterances from the source and target speakers for deriving a conversion function. Here, we proposed an algorithm that relaxes this constraint and allows for the corpus to be non-parallel. It was shown that the proposed method performs quite favorably and the conversion error is low and comparable with the error obtained with parallel training. We intend to demonstrate the satisfying performance of this method subjectively as well, which is clearly indicated by our initial listening tests. Note that if the parallel corpus is made in different conditions compared to the non-parallel corpus, then it is possible that the adaptation algorithm described here might not result in significant improvement, due to reasons such as microphone quality, reverberation *etc*. This is an issue we intend to further explore.

6. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (New York, NY), pp. 655–658, April 1988.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142, March 1998.

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Seattle, WA), pp. 285–289, May 1998.

[4] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts," in *IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Hong Kong), pp. 720–723, April 2003.

[5] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[6] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Maximum likelihood constrained adaptation for multichannel audio synthesis," in *Conf. Record of the Thirty-Sixth Asilomar Conf. Signals, Systems and Computers*, vol. 1, (Pacific Grove, CA), pp. 227–232, November 2002.

[7] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic-transformation adaptation of Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 177–187, March 1999.

[8] A. Kain, *High Resolution Voice Transformation*. PhD thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.

[9] http://www.cslu.ogi.edu/corpora/voices/.