NON-PARALLEL VOICE CONVERSION USING JOINT OPTIMIZATION OF ALIGNMENT BY TEMPORAL CONTEXT AND SPECTRAL DISTORTION

H. Benisty, D. Malah, and K. Crammer

Technion, Israel Institute of Technology
Department of Electrical Engineering
Haifa, 32000, Israel
{hadasbe@tx,malah@ee,koby@ee}.technion.ac.il

ABSTRACT

Many voice conversion systems require parallel training sets of the source and target speakers. Non-parallel training is more complicated as it involves evaluation of source-target correspondence along with the conversion function itself. INCA is a recently proposed method for non-parallel training, based on iterative estimation of alignment and conversion function. The alignment is evaluated using a simple nearest-neighbor search, which often leads to phonetic miss-matched source-target pairs. We propose here a generalized approach, denoted as Temporal-Context INCA (TC-INCA), based on matching temporal context vectors. We formulate the training stage as a minimization problem of a joint cost, considering both context-based alignment and conversion function. We show that TC-INCA reduces the joint cost and prove its convergence. Experimental results indicate that TC-INCA significantly improves the alignment accuracy, compared to INCA. Moreover, subjective evaluations show that TC-INCA leads to improved quality of the synthesized output signals, when small training sets are used.

***Index Terms*—** Non-Parallel Voice Conversion, INCA, Gaussian Mixture Model (GMM), Spectral Distance

1. INTRODUCTION

The goal of voice conversion systems is to modify spoken sentences of a source speaker to sound as if a target speaker had said them. Such transformation can be used for personalizing synthesized output signals of Text-To-Speech (TTS) systems used in cases of vocal pathology, or in automatic dialogue systems, and also for entertainment purposes such as online roll playing.

Classical spectral conversion is based on statistical modeling of the spectral feature vectors as a Gaussian Mixture, so the resulting trained conversion is a mixture of linear conversions [1,2]. Several other GMM-based approaches have been suggested since [3–7]. The classical GMM-based method and its variants are trained using parallel sets where the source and target speakers say the same text. Their training process is based on a one-to-one correspondence between the source and target spectral feature vectors.

In a non-parallel setup, no assumptions are made regarding the content of the training sentences. The source-target correspondence is not straightforward as in the parallel case, thus presenting a greater challenge. Some non-parallel methods bypass this problem by modeling the two speakers separately and perform alignment or adaptation of the model parameters [8, 9]. Some train a conversion using an additional parallel set and adapt its parameters to the desired target speaker [10, 11].

A different approach for non-parallel training called Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment method (INCA), was recently proposed [12]. This approach provides a framework for applying parallel training techniques using non-parallel training sets. It is based on an iterative evaluation of an auxiliary conversion function and matching functions between the source and target vectors. Convergence of this process was demonstrated using empirical evaluations, but, as indicated by the authors of INCA, the alignment process is prone to phonetic mismatch. To smooth these errors they train their auxiliary conversion function using the classical GMM-based method, which is known to have over-smoothing characteristics. Still, the importance of correct time alignment was recently demonstrated as having a large influence on the quality of the synthesized converted speech [13].

In this paper we formulate the non-parallel training process as a minimization problem of a joint cost, considering temporal-context alignment and conversion function. We propose a generalization of INCA, denoted here Temporal-Context INCA (TC-INCA), based on matching sequences of vectors (rather than single vectors), according to their original temporal context. We show that TC-INCA (and hence also INCA) are, in fact, alternating minimization steps of the joint cost, and prove they converge.

Fig. 1 illustrates the main difference between TC-INCA, which is based on matching temporal context vectors, and INCA's, which is based on matching single vectors.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

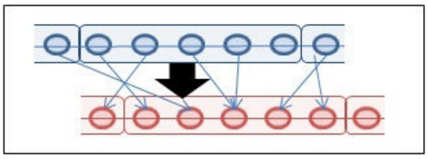


Fig. 1. The proposed alignment process of matching sequences of vectors, used in TC-INCA - thick arrow, as opposed to matching feature vectors used in INCA - thin arrows.

We present objective and subjective evaluations comparing the proposed TC-INCA to INCA. Our method significantly increases the amount of correctly matched pairs and leads to improved synthesized quality and similarity to the target.

2. INCA

We begin by briefly describing the (symmetric variant of) INCA [12]. Let $X = \{\mathbf{x}_k\}_{k=1}^{N_x}, Y = \{\mathbf{y}_j\}_{j=1}^{N_y} \in \mathbb{R}^d$ be two (non-parallel) training sets of feature vectors related to source and target speakers. The training process is based on an iterative evaluation of a parallel auxiliary conversion function, $\mathcal{F}(\cdot)$, its inverse, and two matching functions between the source and target vectors:

$$\begin{cases} p(k) = j & \text{if } \mathbf{x}_k \text{ matches } \mathbf{y}_j \\ q(j) = k & \text{if } \mathbf{y}_j \text{ matches } \mathbf{x}_k. \end{cases} \tag{1}$$

where $k = 1, ..., N_x$ and $j = 1, ..., N_y$.

The iterative process begins by initializing at $t = 0$ an auxiliary conversion function to be the identity function: $\mathcal{F}_0(\mathbf{x}) = \mathbf{x}$. In each iteration, the two matching functions, $p_t(\cdot)$ and $q_t(\cdot)$, are evaluated using a nearest neighbor search between converted source vectors and the target vectors, and vice versa, based on the previous auxiliary function $\mathcal{F}_{t-1}$:

$$\begin{aligned} p_t(k) &= \underset{j}{\text{argmin}} \ \|\mathcal{F}_{t-1}(\mathbf{x}_k) - \mathbf{y}_j\|^2 \\ q_t(j) &= \underset{k}{\text{argmin}} \ \left\|\mathbf{x}_k - \mathcal{F}_{t-1}^{-1}(\mathbf{y}_j)\right\|^2, \end{aligned} \tag{2}$$

These matching functions define a parallelized training set, $\left\{(\mathbf{x}_k, \mathbf{y}_{p(k)}), (\mathbf{x}_{q(j)}, \mathbf{y}_j)\right\}$, which reduces the training process of the auxiliary function, $\mathcal{F}_t$, to the parallel case. The simple nearest neighbor search defined in eqn. (2) often leads to alignment errors, where vectors related to different phonemes are matched. To reduce the influence of miss-aligned vectors, the classical GMM-based conversion, known for its smoothing characteristics, is used to train the auxiliary function.

Convergence is measured via the mean squared-error (MSE) between the converted sets and the original sets:

$$\begin{aligned} D_t &= \frac{1}{N_x + N_y}\left(\sum_{k=1}^{N_x} \left\|\mathcal{F}_t(\mathbf{x}_k) - \mathbf{y}_{p_t(k)}\right\|^2 \right. \\ &\quad \left. + \sum_{j=1}^{N_y} \left\|\mathbf{x}_{q_t(j)} - \mathcal{F}_t^{-1}(\mathbf{y}_j)\right\|^2\right). \end{aligned} \tag{3}$$

Erro et al. [12] show that this measure converges empirically. Once convergence is achieved, the last conversion function is used for conversion. Alternatively, any other parallel conversion function may be trained, based on the parallelized set using the final matching functions.

3. TC-INCA

3.1. Joint Cost

In this section we formulate the conversion stage of a non-parallel conversion as a minimization problem of a joint cost, considering both conversion and context-based matching functions. We define a set of context vectors $\{\mathbf{X}_k\}_{k=1}^{\tilde{N}_x} \in \mathbb{R}^{d(T+1)}$, $\{\mathbf{Y}_k\}_{k=1}^{\tilde{N}_y} \in \mathbb{R}^{d(T+1)}$ obtained by concatenating $T/2$ (T is even) successive vectors before and after each training vector:

$$\begin{aligned} \mathbf{X}_k &\triangleq \left(\mathbf{x}_{k-T/2}^\top, ..., \mathbf{x}_k^\top, ...\mathbf{x}_{k+T/2}^\top\right)^\top \\ \mathbf{Y}_j &\triangleq \left(\mathbf{y}_{j-T/2}^\top, ..., \mathbf{y}_j^\top, ...\mathbf{y}_{j+T/2}^\top\right)^\top, \end{aligned} \tag{4}$$

where $\tilde{N}_x(\leq N_x), \tilde{N}_y(\leq N_y)$ are the number of the source and target context vectors, respectively. We assume that the non-parallel source and target sets are extracted from several continuous utterances (words, sentences). To simplify the notation we also assume that the indices k and j reflect their temporal ordering, meaning that $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, for example, are extracted from consecutive time frames.

Given a spectral conversion function, its inverse, $\mathcal{F}^{-1}(\cdot)$, and two matching functions $p(\cdot)$ and $q(\cdot)$ - pairing each source context vector to a target context vector and vice versa, we write a joint cost function, similar to eqn. (3):

$$\mathcal{L} = \sum_{k=1}^{\tilde{N}_x} \left\|\mathcal{F}(\mathbf{X}_k) - \mathbf{Y}_{p(k)}\right\|^2 + \sum_{j=1}^{\tilde{N}_y} \left\|\mathbf{X}_{q(j)} - \mathcal{F}^{-1}(\mathbf{Y}_j)\right\|^2, \tag{5}$$

where the converted context vectors $\mathcal{F}(\mathbf{X}_k)$ are obtained by applying the conversion function on each feature vector:

$$\mathcal{F}(\mathbf{X}_k) \triangleq \left(\mathcal{F}(\mathbf{x}_{k-T/2})^\top, ..., \mathcal{F}(\mathbf{x}_k)^\top, ...\mathcal{F}(\mathbf{x}_{k+T/2})^\top\right)^\top, \tag{6}$$

and similarly for $\mathcal{F}^{-1}(\mathbf{Y}_j)$.

The cost presented in eqn. (5) is the empirical squared-error between the source and target sequences and their estimated versions (using the conversion function), according to the two alignment functions, p and q. Therefore, we regard the training stage as an optimization problem, aiming to minimize this cost:

$$\{\mathcal{F}^*, p^*, q^*\} = \underset{\{\mathcal{F},p,q\}}{\operatorname{argmin}} \mathcal{L}(p, q, \mathcal{F}). \tag{7}$$

In the parallel case, alignment is obtained by using DTW and phonetic labeling (if available). Assuming, w.l.o.g., that the source and target training vectors are ordered so that $\mathbf{x}_k$ matches $\mathbf{y}_k, \quad \forall k = 1, ..., N$, the matching functions become identity functions: $p(k) = q(k) = k$. Substituting eqn. (6) in eqn. (5) and neglecting the ends, our cost becomes:

$$\mathcal{L}_{para} = T\left(\sum_{k=1}^{N} \|\mathcal{F}(\mathbf{x}_k) - \mathbf{y}_k\|^2 + \sum_{j=1}^{N} \left\|\mathbf{x}_j - \mathcal{F}^{-1}(\mathbf{y}_j)\right\|^2\right), \tag{8}$$

which is a symmetric generalization of the empirical loss minimized in the training process of the classical GMM-based conversion [1], up to a constant T.

3.2. Iterative Minimization

In this section we present an iterative approach, for reducing the joint cost defined in eqn. (7), similar to the iterative process of INCA [12]. Applying standard minimization techniques such as gradient descent is rather problematic considering the non trivial dependency of the joint cost with respect to the matching functions. Alternating minimization is a well known iterative technique for minimizing cost functions depending on more than one variables [14]. Applying this method for minimizing the joint cost, reduces eqn. (7) to two minimization problems solved iteratively for $t = 1, 2, ...$:

$$\begin{aligned} \{p_t, q_t\} &= \underset{\{p,q\}}{\operatorname{argmin}} \mathcal{L}(p, q, \mathcal{F}_{t-1}) & (9) \\ \mathcal{F}_t &= \underset{\mathcal{F}}{\operatorname{argmin}} \mathcal{L}(p_t, q_t, \mathcal{F}), & (10) \end{aligned}$$

Lemma 3.1. *The series* $\{\mathcal{L}_t\}$ *converges to a (local) minimum.*

Proof. According to eqns. (9) and (10), the solutions $\{\mathcal{F}_t, p_t, q_t\}$ sustain:

$$\begin{aligned} \mathcal{L}_t &\triangleq \mathcal{L}(p_t, q_t, \mathcal{F}_t) \leq \mathcal{L}(p_t, q_t, \mathcal{F}_{t-1}) \\ &\leq \mathcal{L}(p_{t-1}, q_{t-1}, \mathcal{F}_{t-1}) \triangleq \mathcal{L}_{t-1}. \end{aligned} \tag{11}$$

The series $\{\mathcal{L}_t\}$ is non-increasing and obviously bounded by zero, therefore converges to a (local) minimum. □

Convergence to a global minimum, or even existence of a single minimum is not guarantied since the original minimization problem stated in eqn. (5) is not convex.

Given a conversion function, the joint cost is separable in p and q, leading to a two-step solution of eqn. (9):

$$\begin{aligned} p_t &= \underset{p}{\operatorname{argmin}} \sum_{k=1}^{\tilde{N}_x} \left\|\mathcal{F}_{t-1}(\mathbf{X}_k) - \mathbf{Y}_{p(k)}\right\|^2 \\ q_t &= \underset{q}{\operatorname{argmin}} \sum_{j=1}^{\tilde{N}_y} \left\|\mathbf{X}_{q(j)} - \mathcal{F}_{t-1}^{-1}(\mathbf{Y}_j)\right\|^2 \end{aligned} \tag{12}$$

We apply a nearest-neighbor search, similar to the one applied for INCA, but instead of using single spectral feature vectors, we use the context vectors defined in eqn. (4):

$$\begin{aligned} p_t(k) &= \underset{j}{\operatorname{argmin}} \|\mathcal{F}_{t-1}(\mathbf{X}_k) - \mathbf{Y}_j\|^2 \\ q_t(j) &= \underset{k}{\operatorname{argmin}} \left\|\mathbf{X}_k - \mathcal{F}_{t-1}^{-1}(\mathbf{Y}_j)\right\|^2. \end{aligned} \tag{13}$$

According to our preliminary experiments, an optimal exhaustive search for the exact solutions of (12) yields a negligible improvement compared to a nearest-neighbor search.

Substituting eqn. (6) into eqn. (10) and neglecting the ends, the minimized term takes a similar form to the parallel symmetrical cost presented in eqn. (8):

$$\begin{aligned} \mathcal{F}_t = \underset{\mathcal{F}}{\operatorname{argmin}} \Bigg\{ &\sum_{k=1}^{\tilde{N}_x} \left\|\mathcal{F}(\mathbf{x}_k) - \mathbf{y}_{p_t(k)}\right\|^2 + \\ &+ \sum_{j=1}^{\tilde{N}_y} \left\|\mathbf{x}_{q_t(j)} - \mathcal{F}^{-1}(\mathbf{y}_j)\right\|^2 \Bigg\}. \end{aligned} \tag{14}$$

Consequently, any parallel conversion method minimizing this squared error can be used as an auxiliary function, using the parallelized training set - $\left\{\left(\mathbf{x}_k, \mathbf{y}_{p_t(k)}\right), \left(\mathbf{x}_{q_t(j)}, \mathbf{y}_j\right)\right\}$. The classical GMM-based conversion, for example, can fit this description since its parameters are evaluated using Least Squares minimization of the MSE between the converted and target vectors [1].

The TC-INCA algorithm, is summarized in Table 1. We note that if no context frames are considered, meaning $T = 0$, TC-INCA essentially becomes identical to INCA.

Table 1. Joint Cost Optimization Using TC-INCA.

Input: a non-parallel training set of context vectors $\{X, Y\}$
Initialization: set the initial conversion function to identity: $\mathcal{F}_0(\mathbf{X}) = \mathbf{X}$
Main Iteration: for $t = 1, 2...$ perform the following steps:
1. Evaluate the matching functions, p_t, q_t, using eqn. (13).
2. Train an auxiliary conversion function using eqn. (14).
3. Evaluate the cost function $\mathcal{L}(p_t, q_t, \mathcal{F}_t)$ using eqn. (5) and check convergence.
Output: conversion and matching functions $p, q, \mathcal{F}$.

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

Three U.S. English speakers (two females and one male) taken from the CMU ARCTIC database [15] were used for our objective and subjective evaluations in two directions - female to female (F2F) and female to male (F2M). Analysis, synthesis and extraction of 24 MFCCs were performed using an available toolkit [16][1], based on the Harmonic Plus Noise Model (HNM) [17, 18].

We used both parallel and non-parallel sets for training, consisting of $(5, 10, 50, 100)$ sentences, and an additional set of 50 parallel sentences for testing, all sampled at 16kHz. The pitch was converted using a simple linear function using the mean and the standard deviation values of the source and target speakers.

4.2. Objective Evaluations

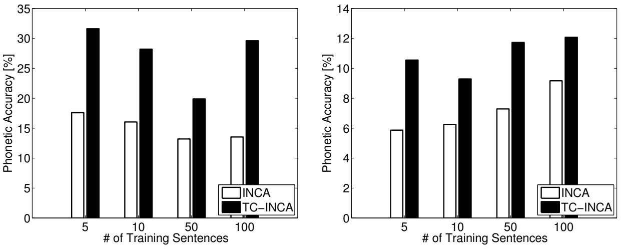
We used two objective criteria to evaluate the performance of the trained matching and conversion functions: phonetic accuracy, measured by the percentage of training vectors having the same phonetic label as their matches, and Normalized Distance (ND), measured using the test set, as the Log-Spectral Distortion (LSD) between the converted and target spectra, divided by the LSD between the source and target spectra.

We used the classical GMM method for training the auxiliary and final conversion functions using full covariance matrices and $(1, 2, 3, 4)$ mixtures for both methods. TC-INCA was trained using several context lengths, $T = (2, 4, 8, 10, 14, 18, 24, 26)$. The number of mixtures (for both methods) and context length (for TC-INCA) were tuned for F2F and F2M and for every training set size, so that maximal (training) accuracy would be attained. Generally, the best accuracy was obtained using longer context $T \in [14, 24]$ for the parallel sets, than for the non-parallel sets $T \in [2, 10]$. Also, as more training sentences were used, more mixtures were preferred.

Fig. 2 presents the accuracy values attained by TC-INCA compared to INCA, averaged over both examined directions (F2F and F2M). TC-INCA leads to significantly higher phonetic accuracy, using either parallel or non-parallel training sets. The ND values achieved by both methods are very similar ($\pm 1\%$), in the range of 0.7-0.75 for the parallel sets and 0.75-0.8 for the non-parallel sets. Nevertheless, the improvement in accuracy has a great influence on the perceived quality and similarity to the target, as presented in the next section.

4.3. Subjective Evaluations

We carried out two preference tests comparing TC-INCA to INCA. Quality tests - in which the listeners were asked to indicate the sentence of better quality, and identity tests, in which the listeners were asked which sentence is more similar to a reference signal (the target speaker). Following Helander et. al [19], we allowed the listeners to answer ”equal”, if they felt they could not decide between the two options. In each test (quality and identity), 10 different randomly ordered (pairs or triplets, correspondingly) were examined by 10 listeners, all 20-30 years old non-experts. For these evaluations we used non-parallel training sets consisting of 5 sentences. Table 2 presents the overall results, averaged over both F2F and F2M conversions. The advantage of TC-INCA is well demonstrated; most listeners marked it as having a higher quality and as more similar to the target speaker, than INCA.



(a) Parallel training sets

(b) Non-parallel sets

Fig. 2. Maximal accuracy [%] (39 phonemes) vs. training set size obtained by TC-INCA and INCA.

Table 2. *Subjective Preference Evaluations.*

	INCA [%]	TC-INCA [%]	Equal [%]
Quality	20 ± 2	$\mathbf{73 \pm 2}$	7 ± 1
Identity	33 ± 2	$\mathbf{54 \pm 2}$	13 ± 1

5. CONCLUSION

We presented a non-parallel training process as a minimization problem of a joint cost, considering both temporal-context alignment and conversion functions. We propose TC-INCA (a generalization of INCA) for iteratively performing this minimization. We show that TC-INCA reduces the joint cost (and therefore INCA too) and prove its convergence.

Objectively, TC-INCA leads to a considerable increase of alignment accuracy and to similar spectral distance values, compared to INCA. Subjective evaluations demonstrate the great influence of accuracy improvement: TC-INCA was rated higher, both in terms of quality and similarity to the target speaker.

6. REFERENCES

[1] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE*

[1] We gratefully thank D. Erro for providing us the analysis/synthesis tools.

Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proc. ICASSP, IEEE*, 1998, vol. 1, pp. 285–288.

[3] A. Kain and M. W. Macon, “Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction,” in *Proc. ICASSP, IEEE*, 2001, vol. 2, pp. 813–816.

[4] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum,” in *Proc. ICASSP, IEEE*, 2001, vol. 2, pp. 841–844.

[5] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] D. Erro, A. Moreno, and A. Bonafonte, “Voice conversion based on weighted frequency warping,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[8] M. Zhang, J. Tao, J. Tian, and X. Wang, “Text-independent voice conversion based on state mapped codebook,” in *Proc. ICASSP, IEEE*, 2008, pp. 4605–4608.

[9] P. Song, W. Zheng, and L. Zhao, “Non-parallel training for voice conversion based on adaptation method,” in *Proc. ICASSP, IEEE*, 2013.

[10] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, “Non-parallel training for voice conversion by maximum likelihood constrained adaptation,” in *Proc. ICASSP, IEEE*, 2004, vol. 1, pp. I–1.

[11] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” in *Proc. ICSLP*, pp. 2446–2449.

[12] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from non-parallel corpora,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[13] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, “On the impact of alignment on voice conversion performance.,” in *Proc. INTERSPEECH*, 2008, pp. 1453–1456.

[14] I. Csiszar and G. Tusnady, “Information geometry and alternating minimization procedures,” *Statistics and Decisions*, vol. 1, pp. 205–237, 1984.

[15] J. Kominek and A. W. Black, “CMU ARCTIC databases for speech synthesis,” 2003.

[16] *http://aholab.ehu.es/users/derro/software.html.*

[17] Y. Stylianou, J. Laroche, and E. Moulines, “High-quality speech modifcation based on a harmonic + noise model,” in *Proc. EUROSPEECH*, 1995.

[18] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[19] E. Helander, J. Nurminen, and M. Gabbouj, “LSF mapping for voice conversion with very small training sets,” in *Proc. ICASSP, IEEE*, 2008, pp. 4669–4672.