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

Hadas Benisty, David Malah, and Koby Crammer

EE Department, Technion – Israel Institute of Technology, Haifa, Israel

• Introduction

❑ **Voice Conversion** - transform a sentence said by a source speaker, to sound as if a target speaker had said it

❑ **Parallel** training sets of the source and target speakers are required by many voice conversion systems, but are not always available for a given target speaker

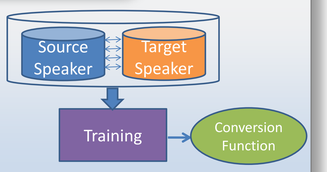
❑ **Non-parallel** systems are more complicated – involve evaluation of **source-target correspondence** along with the conversion function itself

❑ **We propose TC-INCA – a non-parallel conversion method based on temporal context, aiming to increase the alignment accuracy and to reduce the source-target distortion**

• Parallel Voice Conversion

❑ **A Parallel training set** - recorded sentences of the source and target speakers saying the same text

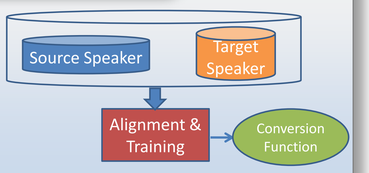
❑ Source-Target correspondence is obvious



• Non-Parallel Voice Conversion

❑ No assumptions are made regarding the uttered text

❑ **Alignment evaluation is needed**



• INCA [Erro et. al., 2010]

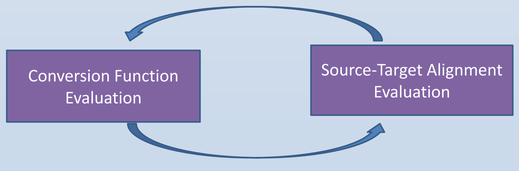
Iterative combination of a **N**earest **N**eighbor search step and a **C**onversion step **A**lignment method

❑ Based on iterative estimation of alignment and conversion function.

❑ The alignment is evaluated using a simple nearest neighbor search between spectral feature vectors related to **single frames**

❑ An auxiliary conversion function is evaluated using a parallel method, based on the evaluated alignment

❑ Often leads to **phonetic miss-matched** source-target pairs



• INCA - Formulation

❑ **Input** - A non-parallel training set $\{\{\mathbf{x}_k\},\{\mathbf{y}_j\}\}$

❑ **Initialization –** set the initial conversion $F(\mathbf{x})=\mathbf{x}$

Main Iteration: for t = 1, 2...

1. Evaluate the matching functions:

$$p_t(k)=\arg\min_j\left\|F_{t-1}(\mathbf{x}_k)-\mathbf{y}_j\right\|^2,\quad q_t(j)=\arg\min_k\left\|\mathbf{x}_k-F_{t-1}^{-1}(\mathbf{y}_j)\right\|^2$$

2. Train an auxiliary conversion function using the parallelized set:

$$\left\{\left(\mathbf{x}_k,\mathbf{y}_{p(k)}\right),\left(\mathbf{x}_{q(j)},\mathbf{y}_j\right)\right\}$$

3. Evaluate the mean squared-error between the converted sets and the original sets and check convergence:

$$\sum_k\left\|F_t(\mathbf{x}_k)-\mathbf{y}_{p_t(k)}\right\|^2+\sum_j\left\|\mathbf{x}_{q_t(j)}-F_t^{-1}(\mathbf{y}_j)\right\|^2$$

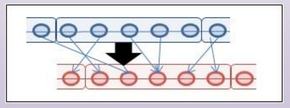
Output: conversion and matching functions p,q,F

• Temporal-Context INCA (TC-INCA)

❑ A generalized approach based on matching sequences of vectors according to their original temporal context

❑ Formulated as a minimization problem **of a joint cost,** considering temporal-context alignment and conversion function

❑ Proved to **converge**



- Thick arrow – TC-INCA matching sequences
- Thin arrows – INCA - matching feature vectors

• TC-INCA-Formulation

❑ **Context vectors** - concatenating T/2 (T is even) successive vectors before and after each training vector:

$$\mathbf{X}_k\triangleq\left(\mathbf{x}_{k-T/2}^T,\ldots,\mathbf{x}_k^T,\ldots,\mathbf{x}_{k+T/2}^T\right)^T,\quad \mathbf{Y}_k\triangleq\left(\mathbf{y}_{k-T/2}^T,\ldots,\mathbf{y}_k^T,\ldots,\mathbf{y}_{k+T/2}^T\right)^T$$

❑ **Joint Cost** - given a spectral conversion function and two matching functions p (·) and q (·) :

$$L_t=\sum_k\left\|F_t(\mathbf{X}_k)-\mathbf{Y}_{p_t(k)}\right\|^2+\sum_j\left\|\mathbf{X}_{q_t(j)}-F_t^{-1}(\mathbf{Y}_j)\right\|^2$$

Where $F_t(\mathbf{X}_k)\triangleq\left(F_t(\mathbf{x}_{k-T/2})^T,\ldots,F_t(\mathbf{x}_k)^T,\ldots,F_t(\mathbf{x}_{k+T/2})^T\right)^T$

❑ The training stage is regarded as an optimization problem:

$$\left\{F^*,p^*,q^*\right\}=\arg\min_{\{F,p,q\}}L$$

❑ Applying alternating minimization:

$$\left\{p_t,q_t\right\}=\arg\min_{\{p,q\}}L\left(F_{t-1},p,q\right)$$

$$F_t=\arg\min_F L\left(F,p_t,q_t\right)$$

❑ Convergence – easily proven since $0\le L_t\le L_{t-1}$

❑ p_t,q_t - evaluated using nearest neighbor search based on the context vectors

❑ F_t - evaluated using the parallelized set $\left\{\left(\mathbf{x}_k,\mathbf{y}_{p(k)}\right),\left(\mathbf{x}_{q(j)},\mathbf{y}_j\right)\right\}$

• Experimental Results

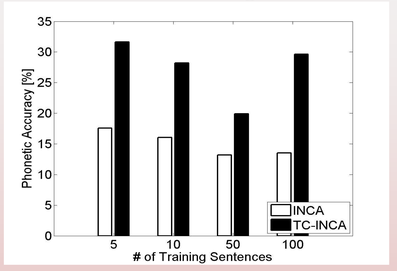
❑ **Evaluated Methods**
- INCA [Erro et. al., 2010]
- The proposed TC-INCA

❑ **Objectively**
- TC-INCA leads to significantly higher phonetic accuracy , using either parallel/non-parallel training sets, and to similar spectral distance values

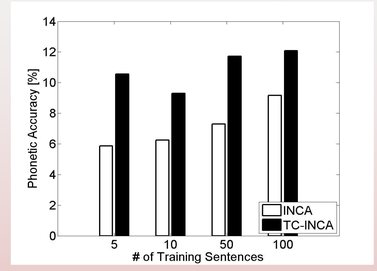
❑ **Subjectively**
- TC-INCA was selected by the majority of listeners as **better** than INCA, both in terms of **quality** and **similarity** to the target

Objective Evaluations

Parallel Training Set



Non-Parallel Training Sets



Subjective Preference

Conversion Method	INCA [%]	TC-INCA [%]	Equal [%]
Quality	20±2	73±2	7±1
Identity	33±2	54±2	13±1