Introduce
- **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 Nearest Neighbor search step and a Conversion step Alignment 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 \( \{ \{ x_i \}, \{ y_j \} \} \)
- **Initialization** – set the initial conversion \( F(x) = x \)
- **Main Iteration** for \( t = 1, 2, ... \)
  1. Evaluate the matching functions:
     \[
     p(k) = \arg \min_j F_j(x_k) - y_j \quad q(j) = \arg \min_i F_i(x_k) - y_j
     \]
  2. Train an auxiliary conversion function using the parallelized set:
     \[
     (x_k, y_{p(k)}), (x_{q(j)}, y_j)
     \]
  3. Evaluate the mean squared error of the converted sets and the original sets and check convergence:
     \[
     \sum |F_j(x_k) - y_{p(k)}|^2 + \sum |x_{q(j)} - F_{q(j)}(y_j)|^2
     \]
- **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

Objective Evaluations
- **Parallel Training Set**
- **Non-Parallel Training Sets**

Subjective Preference

**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