Non-parallel Voice Conversion Using i-Vector PLDA:

Towards Unifying Speaker Verification and Transformation

UNIVERSITY OF EASTERN FINLAND

Aalto University

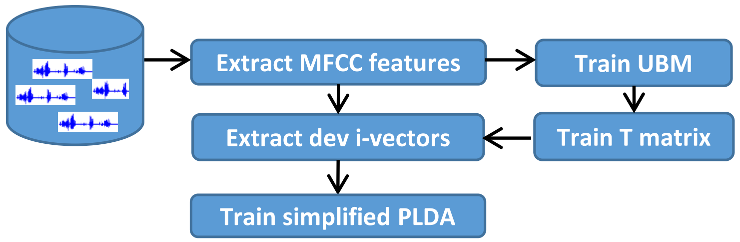
NII

Tomi Kinnunen, Lauri Juvela, Paavo Alku, Junichi Yamagishi

Fully non-parallel system: no frame-level alignment, parallel reference speaker or parallel source/target data required.

Source-to-target **conversion function trained on utterance-level speaker features** instead of low-level frame features.

1. OFFLINE TRAINING



An **i-vector** is a single low-dimensional 'feature vector' of a speech utterance. We model i-vector distributions with **simplified probabilistic linear discriminant analysis** (PLDA) model:

$$\phi_{i,j} = b + \mathbf{S} y_i + \varepsilon_{i,j}$$

- $\phi_{i,j}$: j:th i-vector of i:th speaker
- b: bias
- $\mathbf{S}$: Speaker factor matrix
- y_i: Speaker variable. Convert this
- $\varepsilon_{i,j}$: Residual: within-speaker variation related to content, microphone, etc.

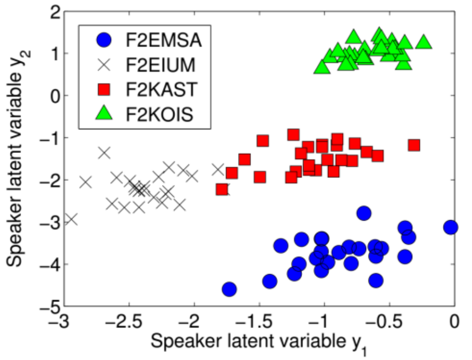
2. TRAINING: extract latent speaker features with PLDA

$\Phi_{\text{src}} = \{\phi_n\}$ ➔ MAP estimate of y_{src}

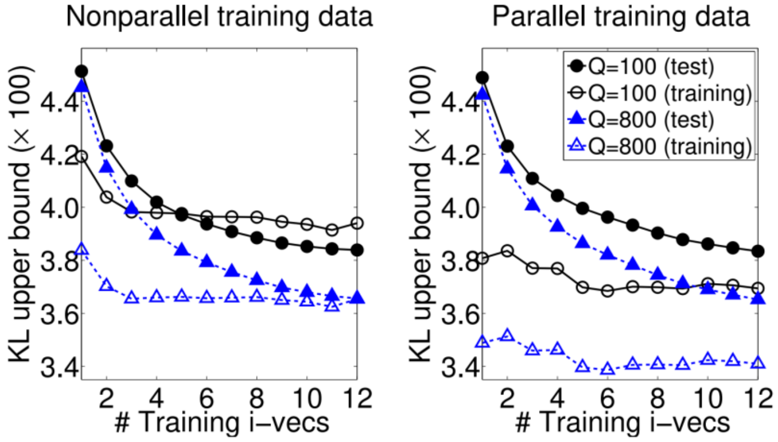
$\Phi_{\text{tar}} = \{\phi_m\}$ ➔ MAP estimate of y_{tar}

No 'pairing': independent extraction of the latent speaker features for each speaker. Number of utterances can also vary.

Datapoint = a speech utterance



$$\hat{y}_i^{\text{MAP}} = \arg\max_{y_i} \prod_{j=1}^{N_i} \mathcal{N}\left(\phi_{i,j} | b + \mathbf{S} y_i, \Lambda\right) \mathcal{N}(y_i | \mathbf{0}, \mathbf{I})$$



3. VOICE CONVERSION: predict target speaker i-vector (i.e. GMM)

PLDA decomposition of source speaker test utterance:

$$\phi_{\text{src}} = b + \mathbf{S}\hat{y}_{\text{src}} + e_{\text{src}}$$

$$\hat{\phi}_{\text{tar}} = b + \mathbf{S}\hat{y}_{\text{tar}} + e_{\text{src}} = \phi_{\text{src}} + \mathbf{S}(\hat{y}_{\text{tar}} - \hat{y}_{\text{src}})$$

Replace source latent identity variable with that of the target speaker.

The predicted target speaker i-vector then defines target GMM means and frame-level conversion:

$$\mu_c^{\text{tar}} = m_c + \mathbf{T}_c \hat{\phi}_{\text{tar}} \quad ➔ \quad \hat{y}_t = x_t + \sum_{c=1}^{C} P(c|x_t)\left(\mu_c^{\text{tar}} - \mu_c^{\text{src}}\right)$$