Novel Metric Learning for Non-Parallel Voice Conversion

Nirmesh J. Shah and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India-382007

Introduction

- Obtaining aligned spectral feature-pairs in non-parallel VC.
- Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) [1].
- Limitation: Euclidean distance may not correlate well with the perceptual distance [2].
- Propose to learn distance metric: Large Margin Nearest Neighbor (LMNN) technique.
- Learned metric: for finding the Nearest Neighbor (NN) pairs in INCA.
- Subjective and objective evaluation of VC systems.

Motivation for Metric Learning

- **INCA Algorithm:** Iteratively repeat three steps, namely, Initialization, Nearest Neighbor Search Step and Transformation Step until the convergence.
- Lower Phonetic Accuracy (PA).
- *t*-stochastic neighbor embedding (t-SNE) visualization of acoustic space.

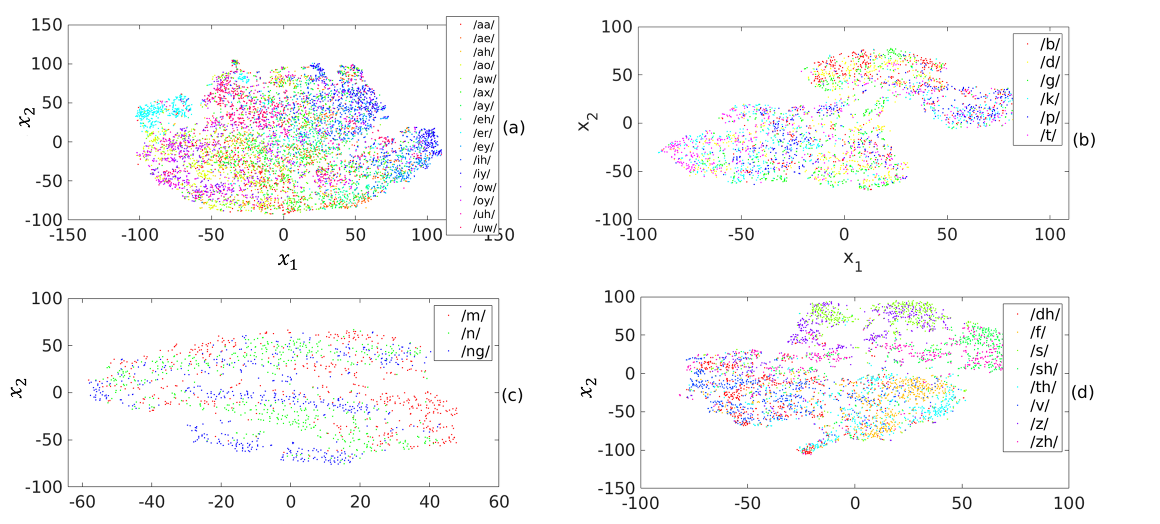

Figure 1: Acoustic features space visualization in 2-D using t-SNE for different speech sound classes, such as (a) vowel, (b) stop, (c) nasal, and (d) fricative.

- Same phoneme uttered by the two speakers does not lie in the neighborhood in Euclidean space.
- Acoustic space $\neq$ Euclidean Space.
- Motivation for defining new metric.

Acknowledgements

The authors would like to thank MeitY, Govt. of India and the authorities of DA-IICT, Gandhinagar, India.
We also acknowledge Samsung R&D Institute, Bangalore.

Metric Learning

- Learning: distance function for a particular task.
- Metric: $d : X \times X \to \mathbb{R}$ should satisfy following four conditions [2]:
  - $d(x_i, x_j) \geq 0$ (non-negativity),
  - $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$ (identity of indiscernible),
  - $d(x_i, x_j) = d(x_j, x_i)$ (symmetry),
  - $d(x_i, x_j) \leq d(x_i, x_r) + d(x_r, x_j)$, where $\forall x_i, x, x_r \in X$ (triangle inequality).
- In general, a distance metric is defined as [2]:

$$d_A(x, y) = (x - y)^T A(x - y). \quad (1)$$

- A must be positive-semidefinite (PSD).
- If A is PSD, $A = G^T G \to d_A(x, y) = ||Gx - Gy||_2^2$.
- Hence, Metric Learning = Learning of global linear transformation.
- Goal: Metric should give minimum squared distance for the pairs $(x_i, x_j) \in \mathcal{S}$.
- The objective function [2]:

$$\arg\min_A \sum_{(x_i,x_j)\in\mathcal{S}} ||x_i - x_j||_A^2, \quad (2)$$

$$\text{subject to} \sum_{(x_i,x_j)\in\mathcal{D}} ||x_i - x_j||_A^2 \geq 1,\ A \succeq 0. \quad (3)$$

- where S and D are set of similar and dissimilar pairs.
- Large Margin Nearest Neighbor (LMNN) [3]:

$$\arg\min_{A\succeq 0} \sum_{(i,j)\in\mathcal{S}} d_A(x_i, x_j) + \lambda \sum_{(i,j,k)\in\mathcal{R}} [1 + d_A(x_i, x_j) - d_A(x_i, x_k)], \quad (4)$$

- where $\mathcal{R}$: set of all triplets (i, j, k) such that x_i and x_j are the target neighbors and x_k is the impostor.

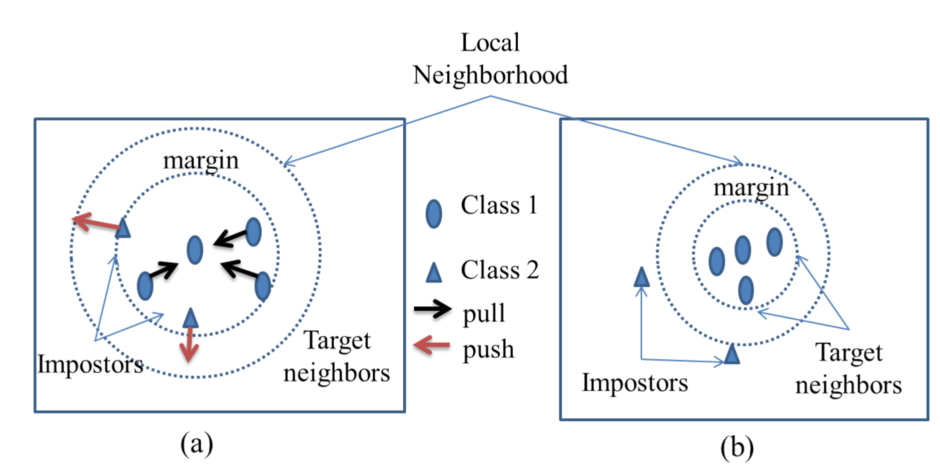

Figure 2: Schematic representation of LMNN technique (a) before and (b) after applying the LMNN technique.

Experimental Results

- TIMIT database for learning metric.
- CMU-ARCTIC database for VC system developments.

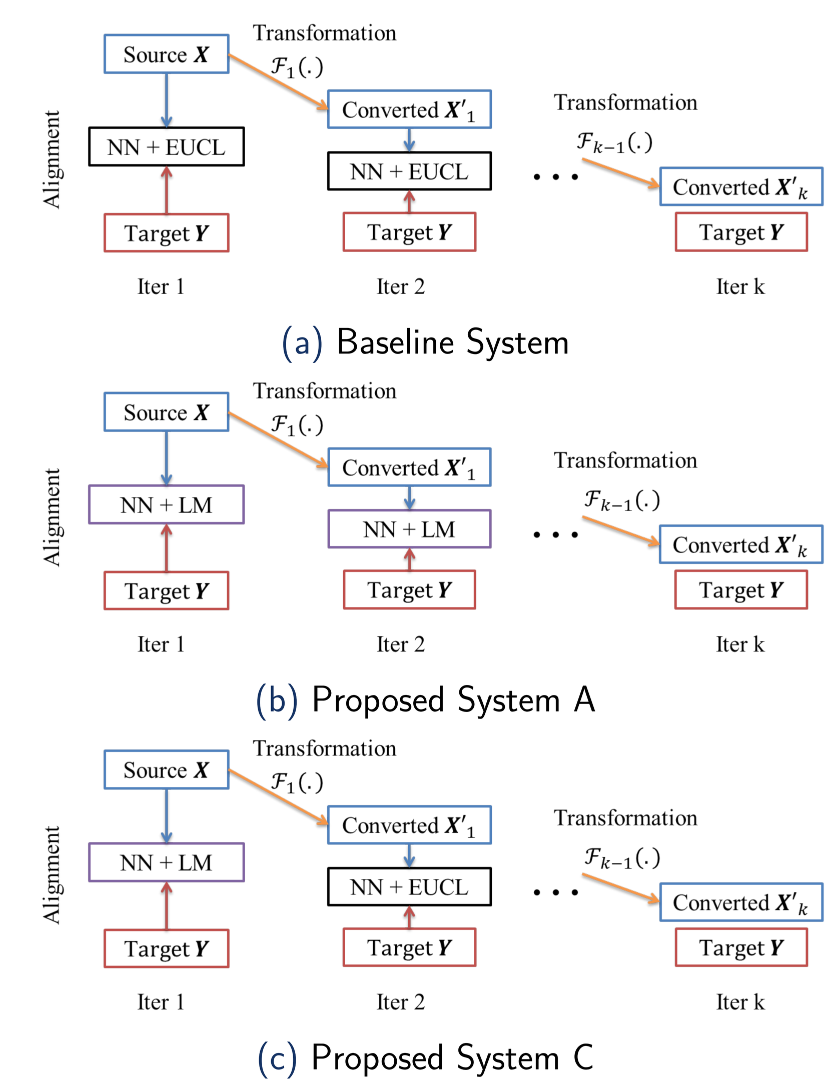

Figure 3: Schematic representation of (a) baseline, (b) proposed system A, and (c) proposed system C. Proposed system B is not shown here, since it applies the baseline technique to the transformed features obtained via the LM, and hence, similar to (a). EUCL: Euclidean metric, LM: Learned metric.

Analysis of Phonetic Accuracy

- Propose technique C is performing consistently better (with on an average 7.93 % relative improvement in PA) than the INCA..

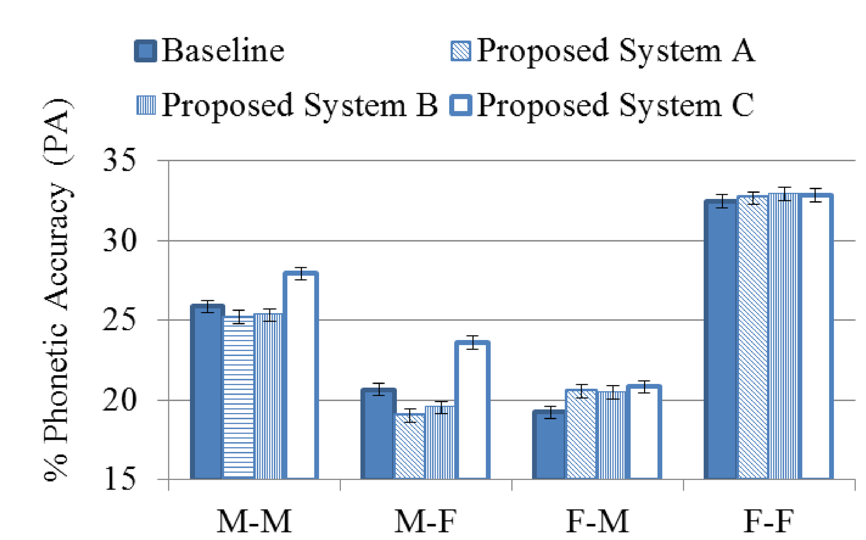

Figure 4: PA of different initialization techniques for non-parallel VC systems.

Evaluations

- Subjective Evaluation: *16* subjects (*5* females and *11* males.

Table 1: MOS analysis for the naturalness of converted voices. Number in the bracket indicates a margin of error corresponding to the 95 % confidence intervals for VC systems

	M-M	M-F	F-M	F-F
Baseline	3.06 (0.27)	2.41 (0.29)	**2.66 (0.28)**	3.5 (0.26)
Proposed System C	**3.31 (0.29)**	**2.81 (0.22)**	2.53 (0.21)	**3.5 (0.25)**

- Objective Evaluation: Mel Cepstral Distortion (MCD)

Table 2: MCD analysis. Number in bracket indicates the margin of error corresponding to the 95 % confidence intervals

	M-M	M-F	F-M	F-F
Baseline	6.53 (0.34)	6.95 (1)	8.02 (1.29)	6.06 (0.93)
Proposed System C	**6.41 (0.09)**	**6.76 (0.26)**	**7.85 (0.34)**	**6.02 (0.24)**

- Pearson Correlation Coefficient (PCC)
- Better phonetic accuracy lead to better MOS.

Table 3: PCC of % PA and MCD with the subjective score

PCC	MOS	SS
PA	0.96	0.37
MCD	-0.3	0.10

Conclusion

- Proposed to exploit metric learning technique for finding NN in the INCA.
- Proposed to use our learned metric only for the initial iteration of INCA since the metric is learned for the actual acoustic features.
- Improvement (in terms of PA) obtained due to proposed system C is clearly reflected in the MOS scores with the PCC of 0.96.

Selected References

[1] D. Erro et al., "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[2] Eric P Xing et al., "Distance metric learning with application to clustering with side-information," in *NIPS*, Vancouver, Canada, 2002, vol. 15, pp. 12–20.

[3] Kilian Q Weinberger and Lawrence K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

Presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12-17, 2019.