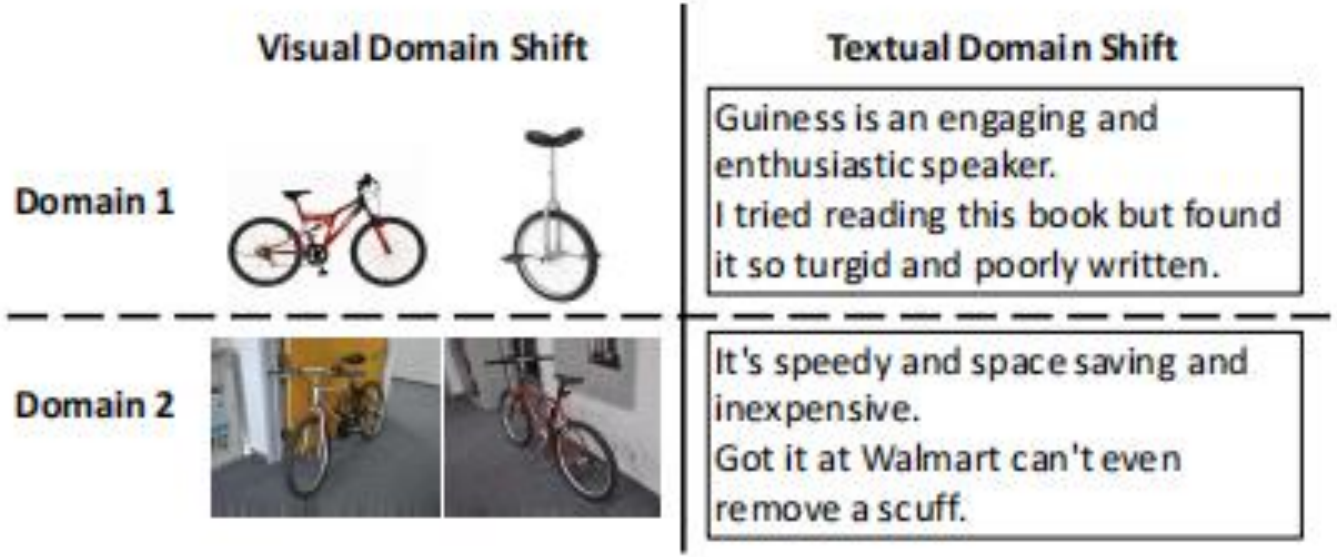
*Weekly Report*

# Cross-Domain Speaker Recognition

- Cross-Domain scenario
    - Cross-lingual
    - Time-drifting
    - Transpathology
    - Cross-device
    - Cross-environmental
    - Cross-directional

Return of Frustratingly Easy Domain Adaptation

- CORrelation ALignment (CORAL)

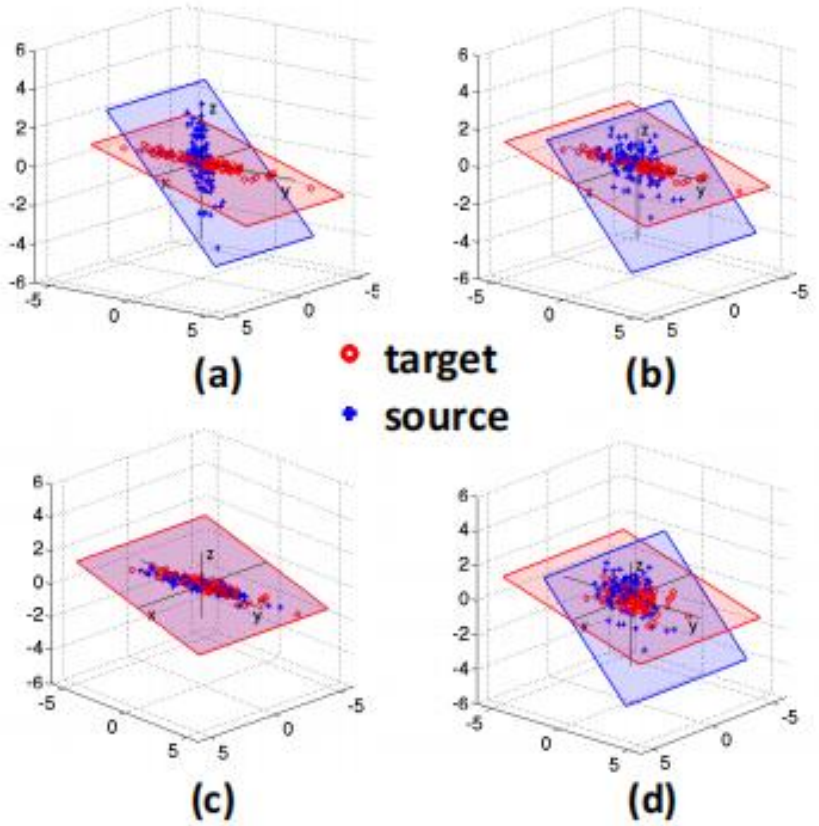



$D_S = \{\vec{x}_i\},\ \vec{x} \in \mathbb{R}^D \qquad L_S = \{y_i\},\ y \in \{1, ..., L\}.$

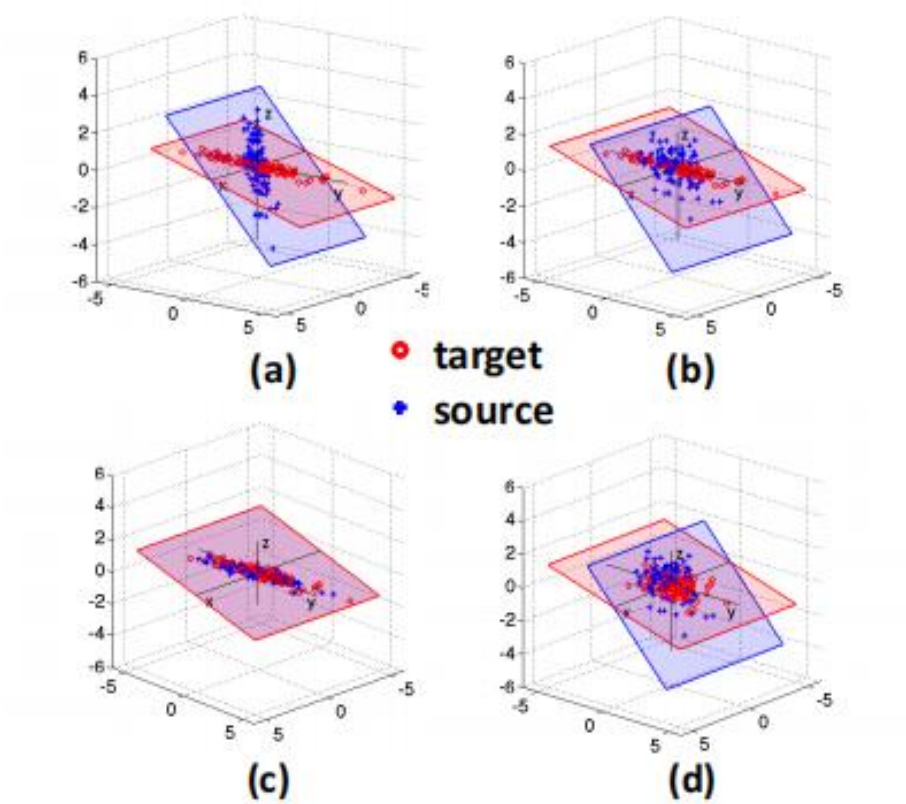
$D_T = \{\vec{u}_i\},\ \vec{u} \in \mathbb{R}^D$

μt = μs = 0 after feature normalization while $C_S \neq C_T$.

$$\min_A \|C_{\hat{S}} - C_T\|_F^2 = \min_A \|A^\top C_S A - C_T\|_F^2$$



$$\min_A \|C_{\hat{S}} - C_T\|_F^2 = \min_A \|A^\top C_S A - C_T\|_F^2$$




$$A^* = U_S E$$
$$= (U_S \Sigma_S^{+\frac{1}{2}} U_S{}^\top)(U_{T[1:r]} \Sigma_{T[1:r]}{}^{\frac{1}{2}} U_{T[1:r]}{}^\top), \quad (2)$$

$C_S^{\frac{-1}{2}}$ $C_T^{\frac{1}{2}}$

**Algorithm 1** CORAL for Unsupervised Domain Adaptation

**Input:** Source Data $D_S$, Target Data $D_T$
**Output:** Adjusted Source Data $D_s^*$
$C_S = cov(D_S) + eye(size(D_S, 2))$
$C_T = cov(D_T) + eye(size(D_T, 2))$
$D_S = D_S * C_S^{\frac{-1}{2}}$ % whitening source
$D_S^* = D_S * C_T^{\frac{1}{2}}$ % re-coloring with target covariance

| | A→C | A→D | A→W | C→A | C→D | C→W | D→A | D→C | D→W | W→A | W→C | W→D | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NA | 35.8 | 33.1 | 24.9 | 43.7 | 39.4 | 30.0 | 26.4 | 27.1 | 56.4 | 32.3 | 25.7 | 78.9 | 37.8 |
| SVMA | 34.8 | 34.1 | 32.5 | 39.1 | 34.5 | 32.9 | 33.4 | 31.4 | 74.4 | 36.6 | 33.5 | 75.0 | 41.0 |
| DAM | 34.9 | 34.3 | 32.5 | 39.2 | 34.7 | 33.1 | 33.5 | 31.5 | 74.7 | 34.7 | 31.2 | 68.3 | 40.2 |
| GFK | 38.3 | 37.9 | 39.8 | 44.8 | 36.1 | 34.9 | 37.9 | 31.4 | 79.1 | 37.1 | 29.1 | 74.6 | 43.4 |
| TCA | 40.0 | **39.1** | **40.1** | 46.7 | **41.4** | 36.2 | 39.6 | 34.0 | 80.4 | **40.2** | 33.7 | 77.5 | 45.7 |
| SA | 39.9 | 38.8 | 39.6 | 46.1 | 39.4 | 38.9 | **42.0** | **35.0** | 82.3 | 39.3 | 31.8 | 77.9 | 45.9 |
| CORAL | **40.3** | 38.3 | 38.7 | **47.2** | 40.7 | **39.2** | 38.1 | 34.2 | **85.9** | 37.8 | **34.6** | **84.9** | **46.7** |

Table 1: Object recognition accuracies of all 12 domain shifts on the Office-Caltech10 dataset (Gong et al. 2012) with SURF features, following the protocol of (Gong et al. 2012; Fernando et al. 2013; Gopalan, Li, and Chellappa 2011; Kulis, Saenko, and Darrell 2011; Saenko et al. 2010).

| | C→I | C→S | I→C | I→S | S→C | S→I | AVG |
|---|---|---|---|---|---|---|---|
| NA | 66.1 | 21.9 | 73.8 | 22.4 | 24.6 | 22.4 | 38.5 |
| SA | 43.7 | 13.9 | 52.0 | 15.1 | 15.8 | 14.3 | 25.8 |
| GFK | 52 | 18.6 | 58.5 | 20.1 | 21.1 | 17.4 | 31.3 |
| TCA | 48.6 | 15.6 | 54.0 | 14.8 | 14.6 | 12.0 | 26.6 |
| CORAL | **66.2** | **22.9** | **74.7** | **25.4** | **26.9** | **25.2** | **40.2** |

Table 4: Object recognition accuracies of all 6 domain shifts on the Testbed Cross-Dataset (Tommasi and Tuytelaars 2014) dataset with DECAF-fc7 features, using the *"full training"* protocol.

# THE CORAL+ ALGORITHM FOR UNSUPERVISED DOMAIN ADAPTATION OF PLDA

- Out of-domain (OOD) features to match the in-domain (InD)

- Probabilistic LDA

$$p(\phi|\mathbf{h},\mathbf{x}) = \mathcal{N}(\phi|\mu, \mathbf{F}\mathbf{h} + \mathbf{G}\mathbf{x} + \mathbf{\Sigma}) \qquad l(\phi_1,\phi_2) = \frac{p(\phi_1,\phi_2)}{p(\phi_1)\,p(\phi_2)}$$

while F and G are the speaker and channel loading matrices, and the diagonal matrix Σ models the residual variances

$$p(\phi) = \mathcal{N}(\phi|\mu, \mathbf{\Phi}_b + \mathbf{\Phi}_w) \qquad \mathbf{\Phi}_b = \mathbf{F}\mathbf{F}^T \quad \mathbf{\Phi}_w = \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma}$$

$$\phi' = \mathbf{C}_I^{\frac{1}{2}}\mathbf{C}_o^{-\frac{1}{2}}\phi$$

$$\mathbf{C}_o' = \mathbf{A}^T\mathbf{C}_o\mathbf{A} = \mathbf{A}^T\mathbf{\Phi}_{w,o}\mathbf{A} + \mathbf{A}^T\mathbf{\Phi}_{b,o}\mathbf{A}$$

- Model-level adaptation

$$\mathbf{\Phi}_b^+ = (1-\beta)\mathbf{\Phi}_{b,o} + \beta\mathbf{A}^T\mathbf{\Phi}_{b,o}\mathbf{A}$$
$$\mathbf{\Phi}_w^+ = (1-\gamma)\mathbf{\Phi}_{w,o} + \gamma\mathbf{A}^T\mathbf{\Phi}_{w,o}\mathbf{A}$$

$$\mathbf{\Phi}_b^+ = \mathbf{\Phi}_{b,o} + \beta\mathbf{B}_b^{-T}(\mathbf{E}_b - \mathbf{I})\mathbf{B}_b^{-1}$$
$$\mathbf{\Phi}_w^+ = \mathbf{\Phi}_{w,o} + \gamma\mathbf{B}_w^{-T}(\mathbf{E}_w - \mathbf{I})\mathbf{B}_w^{-1}$$

simultaneous diagonalization

**Algorithm 1:** The CORAL+ algorithm for unsupervised adaptation of PLDA.

**Input** Out-of-domain PLDA matrices $\{\mathbf{\Phi}_{w,o}, \mathbf{\Phi}_{b,o}\}$
In-domain data $X_{InD}$
Adaptation hyper-parameters $\{\gamma, \beta\}$
**Output** Adapted covariance matrices $\{\mathbf{\Phi}_w^+, \mathbf{\Phi}_b^+\}$

Estimate empirical covariance matrix from in-domain data $X_{InD}$
$\mathbf{C}_I = \mathrm{Cov}(X_{InD})$

Compute out-of-domain covariance matrix
$\mathbf{C}_o = \mathbf{\Phi}_{w,o} + \mathbf{\Phi}_{b,o}$

**for each** $\mathbf{\Phi}$ **in** $\{\mathbf{\Phi}_{w,o}, \mathbf{\Phi}_{b,o}\}$ **do**

Compute Pseudo In-domain covariance matrix
$\mathbf{S} = \mathbf{C}_I^{1/2}\mathbf{C}_o^{-1/2}\mathbf{\Phi}\mathbf{C}_o^{-1/2}\mathbf{C}_I^{1/2}$

Find $\{\mathbf{B}, \mathbf{E}\}$ via simultaneous diagonalization of $\mathbf{\Phi}$ and $\mathbf{S}$
$\{\mathbf{Q}, \mathbf{\Lambda}\} \leftarrow \mathrm{EVD}(\mathbf{\Phi})$
$\{\mathbf{P}, \mathbf{E}\} \leftarrow \mathrm{EVD}(\mathbf{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{S}\mathbf{Q}\mathbf{\Lambda}^{-1/2})$
$\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{P}$

Regularized adaptation of PLDA, $\alpha \in \{\gamma, \beta\}$
$\mathbf{\Phi}^+ = \mathbf{\Phi} + \alpha\mathbf{B}^{-T}\max(0, \mathbf{E} - I)\mathbf{B}^{-1}$

**Notation** $\mathrm{EVD}(.)$ returns a matrix of eigenvectors and the corresponding eigenvalues in a diagonal matrix.

**Table 1**. *Performance comparison on SRE'16 (CMN). The dimension of x-vector after LDA is* 150 *and* 200. *Boldface denotes the best performance for each column.*

| | LDA 150 | | LDA 200 | |
|---|---|---|---|---|
| | EER (%) | MinCost | EER (%) | MinCost |
| `OOD PLDA` | 9.69 | 0.783 | 9.94 | 0.813 |
| `Kaldi PLDA` | 6.82 | 0.552 | 6.57 | 0.558 |
| `CORAL PLDA` | **6.50** | **0.539** | 6.31 | **0.543** |
| `CORAL+ PLDA` | 6.62 | 0.540 | **6.30** | 0.553 |
| `w/o reg` | 6.93 | 0.544 | 6.51 | 0.547 |

**Table 2**. *Performance comparison on SRE'18 (CMN2). The dimension of x-vector after LDA is* 150 *and* 200. *Boldface denotes the best performance for each column.*

| | LDA 150 | | LDA 200 | |
|---|---|---|---|---|
| | EER (%) | MinCost | EER (%) | MinCost |
| `OOD PLDA` | 7.19 | 0.538 | 7.47 | 0.569 |
| `Kaldi PLDA` | 6.25 | 0.435 | 6.48 | 0.466 |
| `CORAL PLDA` | 6.22 | 0.449 | 6.42 | 0.482 |
| `CORAL+ PLDA` | **5.95** | **0.421** | **5.80** | **0.438** |
| `w/o reg` | 6.49 | 0.441 | 6.33 | 0.460 |

- Motivation
  - methods based on neutral spoken voice suffer performance degradation with varying speaker style and effort
  - the performance degradation on spoken voice accompanied by the modest increase in performance on singing voice indicates fine-tuning as a sub-optimal solution

- The increase in the sound range
- Deviation of F1 and F2 resonants

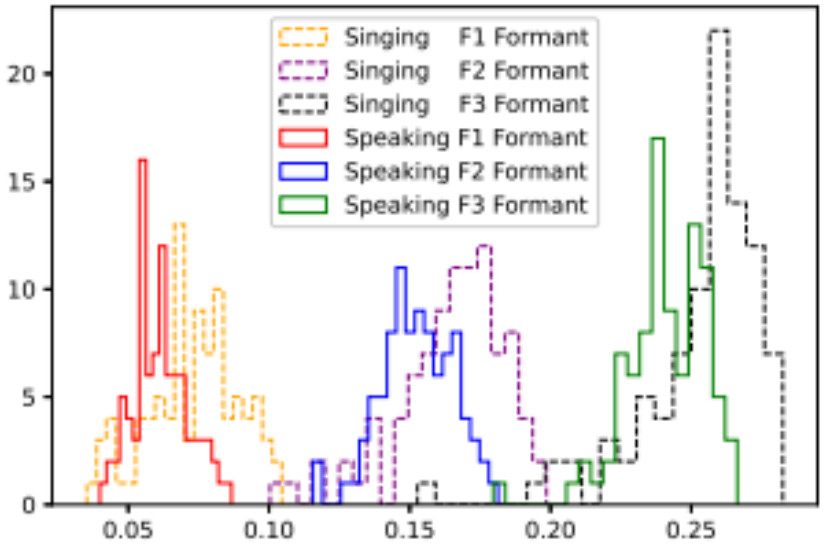



**Fig. 3**: Histogram plots of the first three formants (F1-F3) of spoken and singing speech from the JukeBox-V2 dataset

- 1D-CNN framework

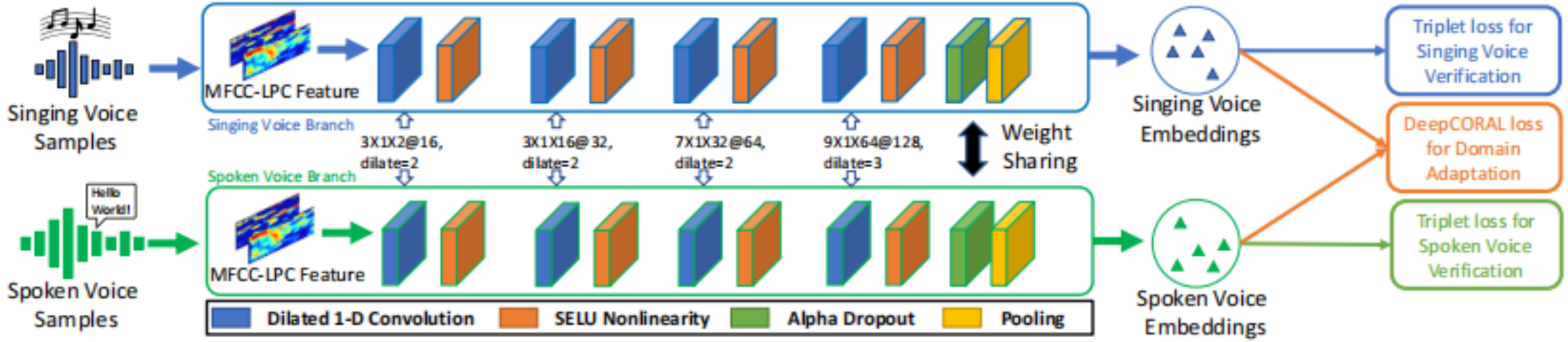



**Fig. 1**: A visual representation of the domain-adaptation-based 1D- CNN framework proposed in Section 3.

Triplet Loss:

$$L(S_a, S_p, S_n) = \sum_{a,p,n}^{N} \cos(g(S_a), g(S_n)) - \cos(g(S_a), g(S_p)) + \alpha_{margin}$$

DA loss:

$$L_{DA}(g(x_{si}), g(x_{sp})) = \frac{1}{4d^2} \|(C_{si} - C_{sp})\|_F^2$$

$\|\cdot\|_F^2$ is the squared matrix Frobenius norm

$$L = \alpha_1 L_{si} + \alpha_2 L_{sp} + \beta L_{DA}$$

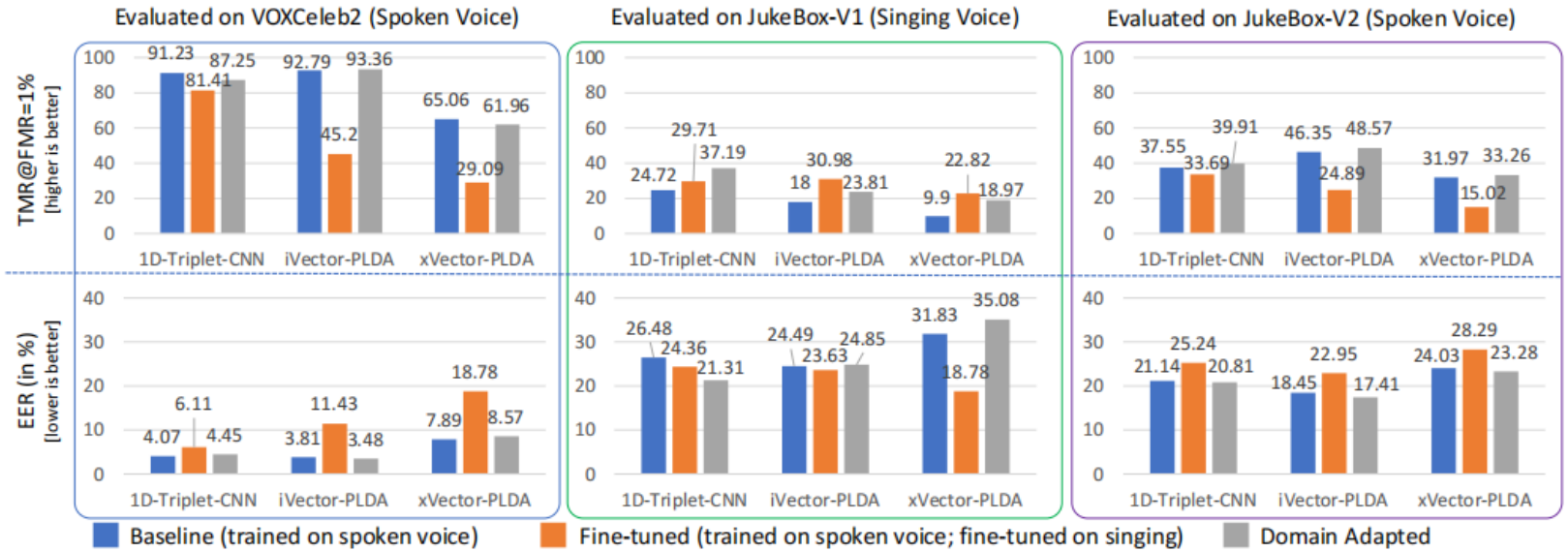



**Fig. 2**: Summary of verification performance (Top: TMR@FMR=1%, Bottom: EER (in%)) across different evaluation conditions. Note the increase in *singer* recognition performance in both fine-tuned (orange bars) and domain adapted (grey bars) models and increase in speaker recognition performance in domain adaptation over fine-tuning.

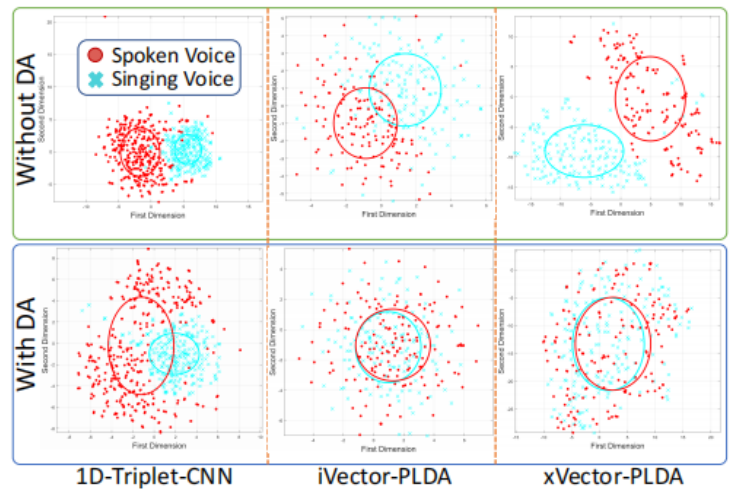



**Fig. 4**: t-SNE plots of the speech embeddings (with and without DA) of singing and spoken voice from the JukeBox-V2 dataset. The circles were added to indicate the apparent cluster boundaries. After DA, the apparent domain gap is reduced leading to overlap of the circle as shown in the lower row. The without DA methods refer to the baseline models.