UNSUPERVISED LEARNING OF ACOUSTIC FEATURES VIA DEEP CANONICAL CORRELATION ANALYSIS

Weiran Wang[1] **Raman Arora**[2] **Karen Livescu**[1] **Jeff A. Bilmes**[3]

[1]TTI-Chicago [2]Johns Hopkins University [3]University of Washington

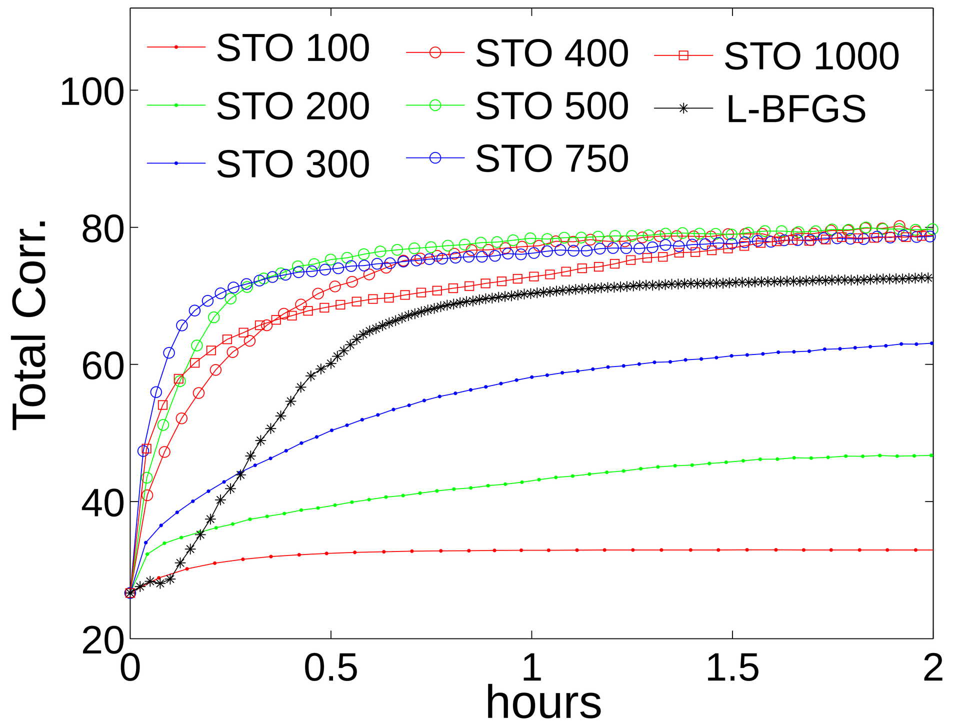
Presented by Sadaoki Furui

1 Overview

- **Background**
  - Can we learn better acoustic features if we have access to multi-view data external to recognizer training/test data?
  - Here the views are acoustics and articulation.
  - Learned transformations are applied to **acoustics-only** data for recognizer training and testing.
  - Previous work: Improved recognition with transformations learned via canonical correlation analysis (CCA) and kernel CCA [1,2].
  - Intuition:
    - 2^{nd} view helps isolate signal from noise.
    - Like articulatory inversion, but using latent articulatory space.
- **This work**
  - We apply deep CCA (DCCA), where the feature mapping is a deep neural network (DNN) [3].
  - DCCA significantly improves phone recognition, without access to test speakers' articulatory data.
  - New stochastic optimization algorithm for large-scale DCCA.

4 Stochastic Optimization

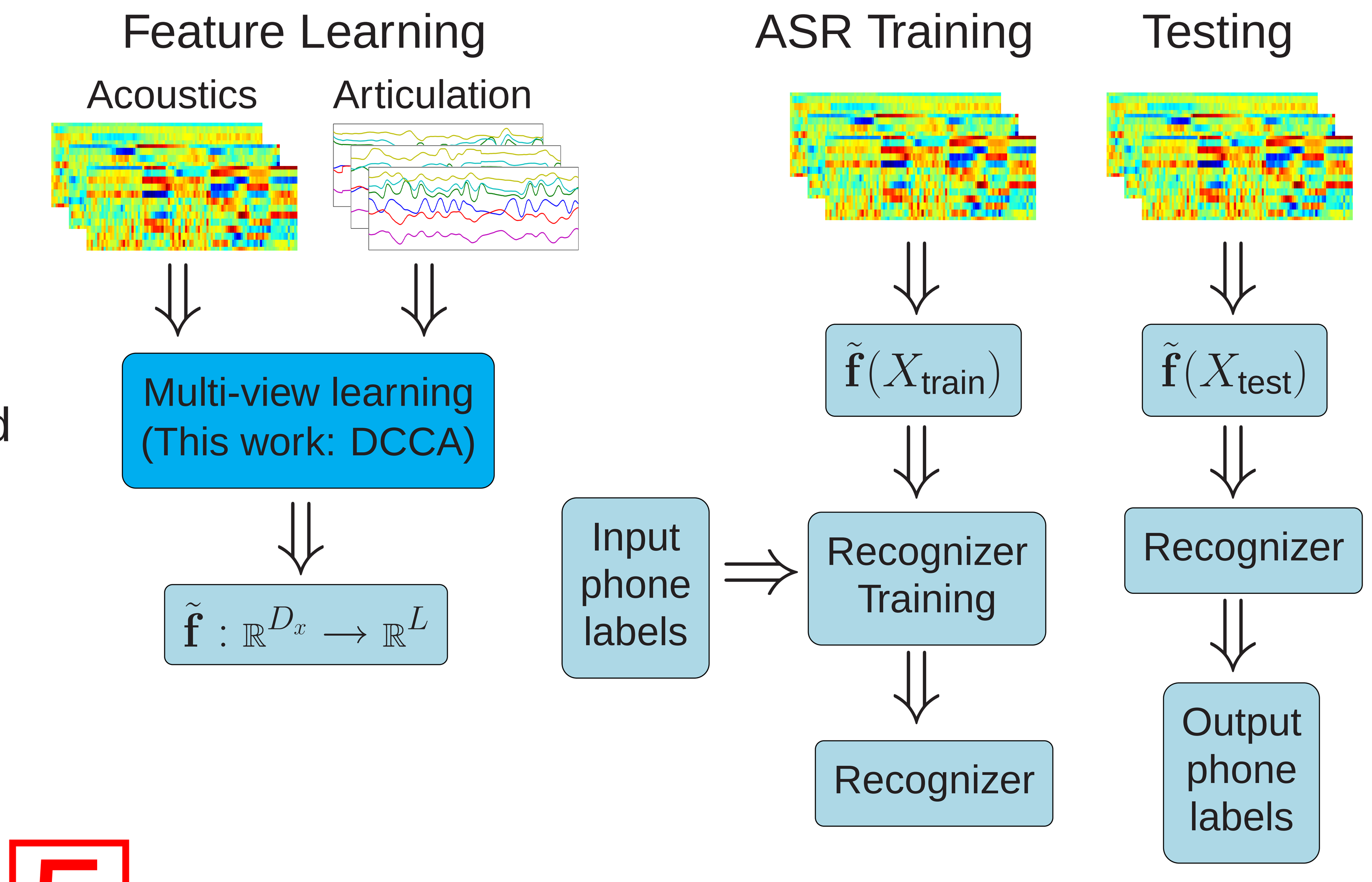
- DCCA objective is a constrained loss that does not decompose over the training samples $\Rightarrow$ not a good fit for stochastic gradient descent, but batch training is very slow.
- We use a minibatch stochastic approach with large minibatches for stable estimates of covariance matrices and gradient.



Learning curves (total correlation vs. training time) of DCCA on the 'JW11' set of [3]. Maximum correlation is the projection dimensionality ($L = 112$).

"STO n" = stochastic optimization with minibatch size n.

2 Multi-View Acoustic Feature Learning

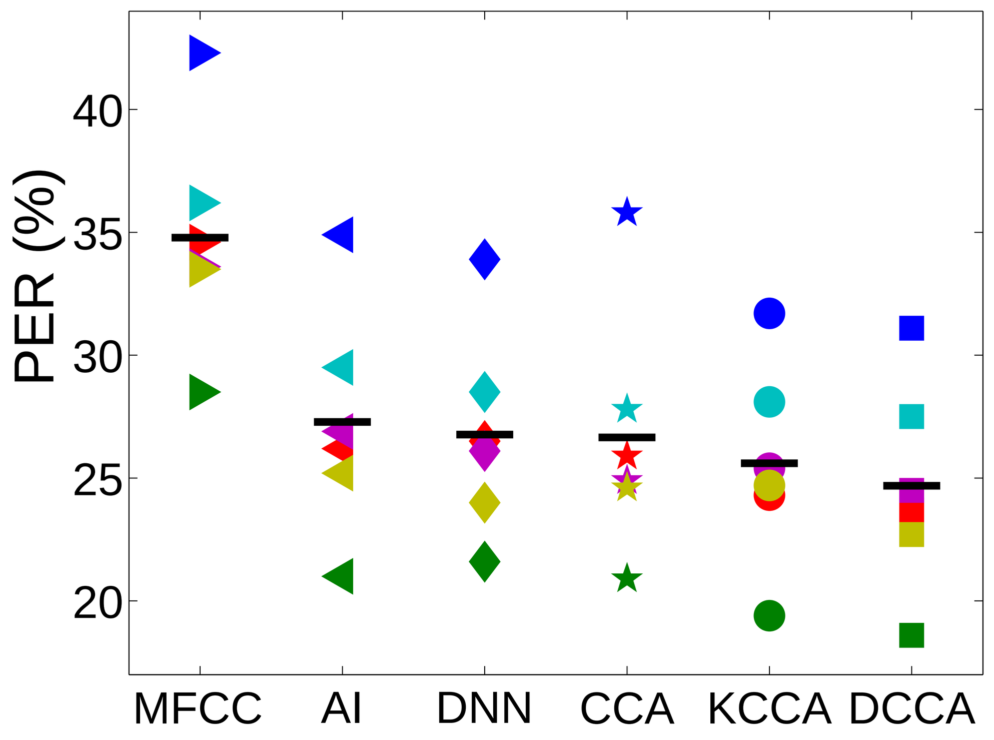


5 Experimental Results

Data: Wisconsin X-ray Microbeam Database, 47 speakers with ~ 50 utterances each, divided into 35/8/2/2 for feature learning/ASR train/tune/test. 6-fold cross-validation for ASR.

Acoustic input: 13 MFCCs + $\Delta + \Delta\Delta \times 7$ frames (273D).

Articulatory input: x, y displacements of 8 pellets $\times$ 7 frames (112D) + per-speaker mean/variance normalization.



Phone error rates using different feature transformations in a tandem recognizer.

Each color denotes a fold. Horizontal bars give the average PER over folds.

AI = articulatory inversion; DNN = supervised DNN features learned from ASR training data.

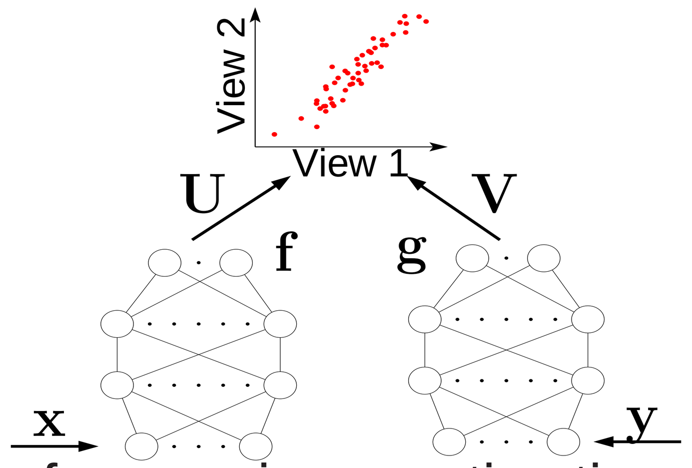
- DCCA is the best performer in all folds. All DCCA improvements over other feature types are significant at $p < 0.05$.
- Code is available from `http://ttic.edu/livescu`

3 Deep Canonical Correlation Analysis

- **Training data**: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^{D_x}$ and $\mathbf{y}_i \in \mathbb{R}^{D_y}$ are input features for i^{th} frame. Here x = acoustics (View 1) and y = articulatory measurements (View 2).
- **Feature mappings**: $\mathbf{f} : \mathbb{R}^{D_x} \rightarrow \mathbb{R}^{d_x}$ and $\mathbf{g} : \mathbb{R}^{D_y} \rightarrow \mathbb{R}^{d_y}$, optionally parameterized by $\mathbf{W}_1, \mathbf{W}_2$, for View 1 and View 2 respectively. Let $\mathbf{F} = \mathbf{f}(\mathbf{X}) = [\mathbf{f}(\mathbf{x}_1), \dots, \mathbf{f}(\mathbf{x}_N)]$, $\mathbf{G} = \mathbf{g}(\mathbf{Y}) = [\mathbf{g}(\mathbf{y}_1), \dots, \mathbf{g}(\mathbf{y}_N)]$.
- **CCA objective**: Find linear projections $\mathbf{U} \in \mathbb{R}^{d_x \times L}$ and $\mathbf{V} \in \mathbb{R}^{d_y \times L}$, and optionally parameters of $\mathbf{f}, \mathbf{g}$, with maximal canonical correlation:

$$\max_{\mathbf{U},\mathbf{V},\mathbf{W}_1,\mathbf{W}_2} \operatorname{tr}\left(\mathbf{U}^\top \mathbf{F} \mathbf{G}^\top \mathbf{V}\right)$$

$$\text{s.t.} \quad \mathbf{U}^\top \left(\frac{1}{N} \mathbf{F}\mathbf{F}^\top + r_x I\right) \mathbf{U} = I, \qquad \mathbf{V}^\top \left(\frac{1}{N} \mathbf{G}\mathbf{G}^\top + r_y I\right) \mathbf{V} = I.$$



where r_x, r_y are regularization coefficients for covariance estimation.

- **CCA variants**
  - Linear CCA (CCA): $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ and $\mathbf{g}(\mathbf{y}) = \mathbf{y}$.
  - Kernel CCA (KCCA): $\mathbf{f}/\mathbf{g}$ are feature maps induced by kernels k_x/k_y.
  - Deep CCA (DCCA): $\mathbf{f}$ and $\mathbf{g}$ are outputs of DNNs.
- For fixed $\mathbf{f}, \mathbf{g}$, optimal $(\mathbf{U}, \mathbf{V})$ given via singular value decomposition.
- **Final features**: $\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{U}^\top \mathbf{f}(\mathbf{x})$.

6 Conclusions

- DCCA significantly improves over previous multi-view methods, articulatory inversion, and supervised DNN features.
- Improvement over articulatory inversion suggests predicting details of articulation is not important or useful.
- Stochastic optimization allows DCCA to scale well to large data.
- Future work: Apply to hybrid ASR, new domains; incorporate supervision [4]; further analysis of stochastic training and network types

References

[1] Bharadwaj, Arora, Livescu, and Hasegawa-Johnson. *Multi-view acoustic feature learning using articulatory measurements.* IWSML 2012.

[2] Arora and Livescu. *Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains.* ICASSP 2013.

[3] Andrew, Arora, Livescu, and Bilmes. *Deep canonical correlation analysis.* ICML 2013.

[4] Arora and Livescu. *Multi-view learning with supervision for transformed bottleneck features.* ICASSP 2014.