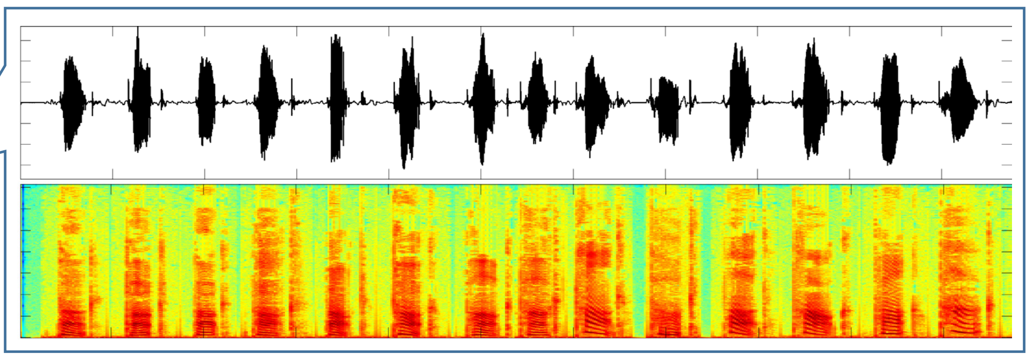
Discriminative Template Learning in Group-Convolutional Networks for Invariant Speech Representations

Chiyuan Zhang[1], Stephen Voinea[1], Georgios Evangelopoulos[1,2], Lorenzo Rosasco[1,2], Tomaso Poggio[1,2]
[1] Center for Brains, Minds and Machines, McGovern Institute for Brain Research at MIT
[2] LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology

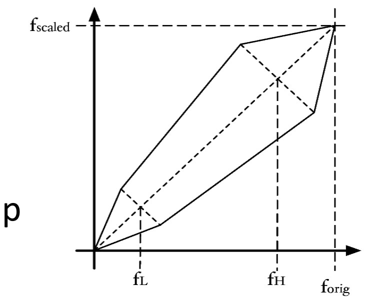
1. Overview

Goal: framework for learning (invariant) speech representations.
Motivation:
- Speech variations (speaker, tempo, accent, pronunciation, …): major ASR challenge for learning with few resources (labeled examples).
- Convolutional Neural Networks (CNNs): invariance to local frequency translations only.

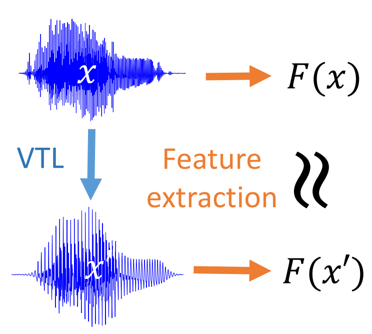
Contributions:
1. CNNs for generic transformations other than translations;
2. Theoretical justification of why CNNs work well via i-theory [1, 2];
3. Algorithm for discriminative learning of templates (exemplars) for group-transformation invariant speech representations [3];
4. Application to Vocal Tract Length (VTL) variation*:* improved frame-based phone classification errors on TIMIT and WSJ.



- VTL normalization: rectifies inter-speaker variability.
- Speaker adaptation through VTL modification (VTL warping).
- Transformation: piecewise linear map of the frequency axis.

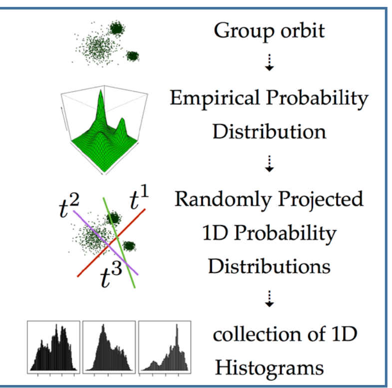


2. A Theory for Group Invariant Representations



Main Theorems [1, 2]
1. G is a group: the orbit $O_G(x) = \{\Phi_g(x) : g \in G\}$ is an invariant and selective feature map under transformations Φ_g.
2. G is locally compact: our network architecture is computing an approximation of $O_G(x)$.

Invariant: $\exists g:\ x' = \Phi_g(x) \Rightarrow F(x) = F(x')$; **Selective**: $F(x) = F(x') \Rightarrow \exists g: x' = \Phi_g(x)$



3. CNN as Approximately Invariant Orbit Map

Convolution: $\boldsymbol{y}[i] = \sum_j \boldsymbol{x}[j]\boldsymbol{t}[i-j] = \sum_j \boldsymbol{x}[j]\tilde{\boldsymbol{t}}[j-i] = \langle \boldsymbol{x}, \Phi_i(\tilde{\boldsymbol{t}})\rangle$,
Inner-product of the input and the transformed template (filter).

Pooling: $\boldsymbol{z}[i] = \max_{j \in G_i} \boldsymbol{y}[j]$, accumulating (MAX) statistics of inner-products.

4. CNN Over VTL Transformations

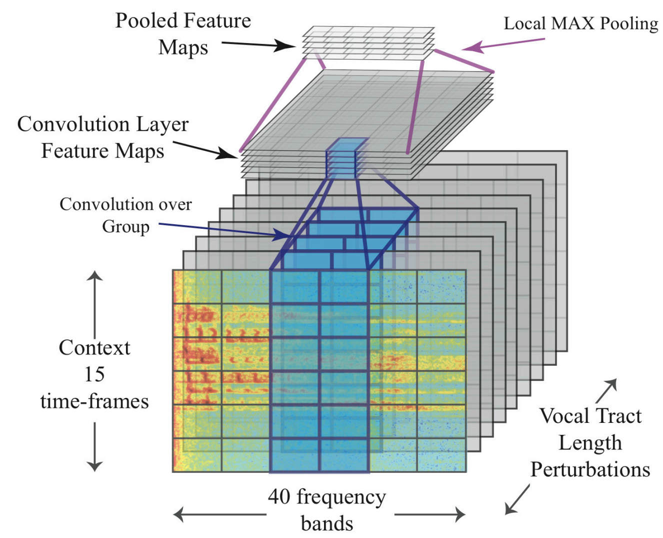
CNNs are approximately invariant to local (frequency) translation.
Generalized CNN: replace Φ_i with general transformations Φ_g.
VTL-CNN: (this paper) Use VTL transforms and propose networks with convolutions over VTL.

5. Learning VTL-CNN

Issue: Back-propagation with VTL-CNN: need to compute $\partial\left(\langle \boldsymbol{x}, \Phi_g(\boldsymbol{t})\rangle\right)/\partial t$ (*simple only for linear/parametric/known transformations*).

Observation: for unitary groups
$\langle \boldsymbol{x}, \Phi_g(\boldsymbol{t})\rangle = \langle \Phi_g^{-1}(\boldsymbol{x}), \boldsymbol{t}\rangle = \langle \Phi_{g^{-1}}(\boldsymbol{x}), \boldsymbol{t}\rangle$

Data Augmentation: work with transformed inputs; gradient becomes easy to compute.



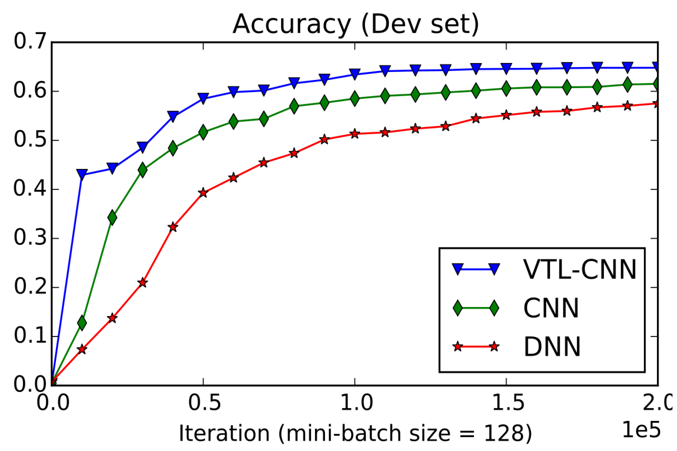
6. VTL-CNN Architecture

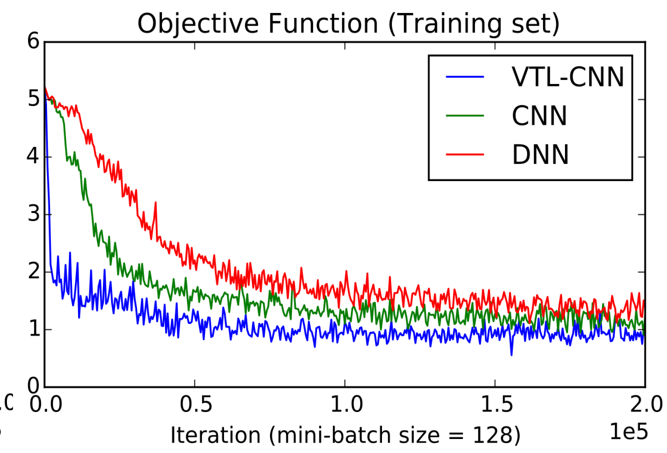
Input $\boldsymbol{x}$ *as 3D-tensor (15 x 40 x 9)*: 15-frame context × 40-dim filter-banks (no $\Delta, \Delta\Delta$) × 9-VTL perturbations (augmentation).
- Local filters $\boldsymbol{t}$ applied to inputs $\boldsymbol{x}$ transformed with different VTL perturbations Φ_g.
- MAX pooling over responses of $\Phi_g(\boldsymbol{x})$.
- Pooled feature maps fed into densely-connected layers.

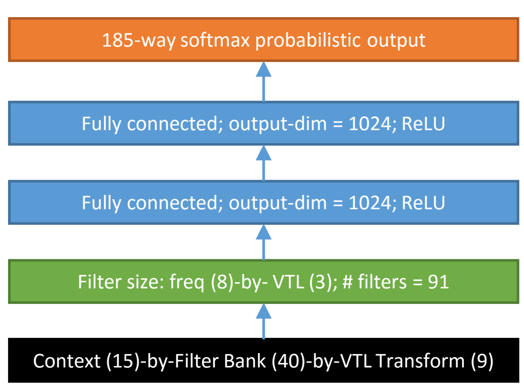
Filters are jointly learned with the DNN classifier

7. Experimental Evaluation (TIMIT, WSJ)

Training curves on TIMIT







← Proposed Architecture

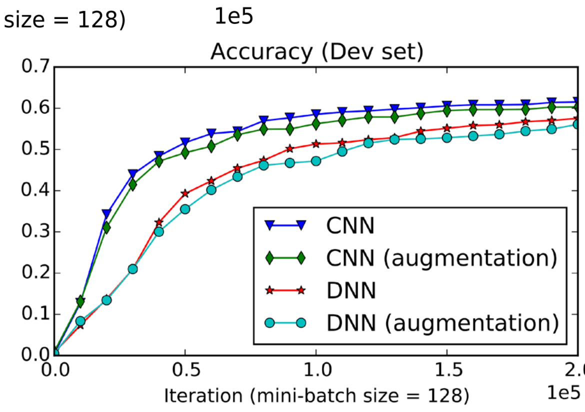
↓ Baseline Architectures
Baseline CNN: similar architecture, pools only over frequency translation.
Baseline DNN: replaces the conv-pool layer with densely connected layer.

All models: similar number of parameters.

Model	TIMIT	WSJ
DNN	58.25	63.65
CNN	62.41	66.78
VTL-CNN	64.62	70.08

Test Frame Accuracy (ALL)

Datasets: TIMIT and WSJ.
Task: Frame classification -- targets are forced-aligned mono-phone HMM states.
Baselines:
Densely-connected Deep Neural Network (DNN).
CNN that pools over frequency translation.
Training: SGD without pre-training, on standard train-dev-test splits.



Compare with Data Augmentation:

Baseline models were trained on same augmented data used by VTL-CNN. Performances are similar to models trained on the original datasets.

Conv-pool over VTL is crucial for taking full advantage of augmented data!

Acknowledgements: This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. Chiyuan Zhang and Stephen Voinea acknowledge the support of a Nuance Foundation Grant.

References: [1] F. Anselmi & T. Poggio. *Representation Learning in Sensory Cortex: a theory*. (2014).
[2] F. Anselmi et. al. *Unsupervised Learning of Invariant Representations in Hierarchical Architectures*, CBMM Memo 001, 2013.
[3] C. Zhang, S. Voinea, G. Evangelopoulos, L. Rosasco, T. Poggio. *Phone Classification by a Hierarchy of Invariant Representation Layers*. INTERSPEECH 2014.