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

Chiyuan Zhang[1], *Stephen Voinea*[1], *Georgios Evangelopoulos*[1,2], *Lorenzo Rosasco*[1,2,3], *Tomaso Poggio*[1,2]

[1] Center for Brains, Minds and Machines, Massachusetts Institute of Technology, USA
[2] LCSL, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
[3] DIBRIS, Universitá degli studi di Genova, Italy

{chiyuan,voinea,gevang,lrosasco}@mit.edu, tp@ai.mit.edu

Abstract

In the framework of a theory for invariant sensory signal representations, a signature which is invariant and selective for speech sounds can be obtained through projections in template signals and pooling over their transformations under a group. For locally compact groups, e.g., translations, the theory explains the resilience of convolutional neural networks with filter weight sharing and max pooling across their local translations in frequency or time. In this paper we propose a discriminative approach for learning an optimum set of templates, under a family of transformations, namely frequency transpositions and perturbations of the vocal tract length, which are among the primary sources of speech variability. Implicitly, we generalize convolutional networks to transformations other than translations, and derive data-specific templates by training a deep network with convolution-pooling layers and densely connected layers. We demonstrate that such a representation, combining group-generalized convolutions, theoretical invariance guarantees and discriminative template selection, improves frame classification performance over standard translation-CNNs and DNNs on TIMIT and Wall Street Journal datasets.

Index Terms: speech representations, neural networks, speech invariance, speech recognition

1. Introduction

Speech recognition systems built with statistical learning algorithms rely on the pre-assumption that the unknown probability distribution of speech data is fixed or similar for both train and test sets [1, 2]. Consequently, the mismatch caused by different speakers, speaking styles or pronunciations/accents is a major challenge for generic, real-world speech recognition. A number of normalization/adaptation techniques are applied to deal with such variability [3, 4]. On the other hand, human speech perception is remarkably robust to signal variations. Apart from the sophisticated contextual inference through complex language models, the lower-level neural representation of speech sounds might also be important for speech recognition [5, 6]. An invariant representation of the speech signal, in both biological and artificial systems, is crucial for improving the robustness of acoustic to phonetic mapping, decreasing the sample complexity (i.e., the required train set size) and enhancing the generalization performance of learning across distribution mismatch.

Representations invariant to transformations [7] that have a locally compact group structure can be obtained through a feature map that defines equivalence classes [8, 9]. The invariant map is obtained by the average over the group G of (possibly non-linear) measurements: the projections on a template $\boldsymbol{t}$ of the set of transformed signals $\{g\boldsymbol{x}|g \in G\}$. Additionally, multiple maps over a set $\{\boldsymbol{t}_k\}_{k=1}^K$ account for a selective representation with components given by $\mu_\eta^k(\boldsymbol{x}) = 1/|G| \sum_g \eta(\langle g\boldsymbol{x}, \boldsymbol{t}_k \rangle)$, where η is the non-linear measurement function. In extension, the approach can yield approximate invariance to unknown, non-group transformations, given access to a sufficient number of class-specific templates (and their transformed versions) [8]. The framework leads to biologically plausible neural representation models, forms predictions about representations in the ventral stream of the visual cortex [10] and provides mathematical justifications for a class of generalized convolutional network models [9]. Based on our previous work on audio representations [11, 12, 13], we propose a representation learning algorithm for invariance to more generic speech transformations.

Our contributions in this paper are: 1) a general framework for discriminative learning of invariant and selective representations, 2) a generalized convolutional network, defined for transformations other than local shifts, that can be discriminatively trained for optimal templates; our framework includes convolutional neural networks (CNNs) [14, 15] as a canonical special case, 3) an application to the transformations of vocal tract length (VTL) variations, and 4) improving, through VTL transformation invariance, the phone classification performance of standard CNNs and DNNs on TIMIT and World Street Journal (WSJ) datasets.

2. Related Work

Previous work on group-invariant speech representations focused on feed-forward, unsupervised networks using either random templates under speech-specific transformations [11, 12] or predefined wavelets under scaling and translations [16]. Such approaches are tractable to analyze and have nice theoretical guarantees regarding the type of invariances in the resulting representations. However, they are generally outperformed by explicit, large-scale, supervised training models [14, 15]. In this paper we extend a framework for group-invariant representations to the supervised regime, i.e. using data-specific, learned templates instead of random (or analytic) ones. CNNs can be shown to be a special case that approximates invariant representations for the translation group. Learning the filters jointly with the network parameters has been explored as an alternative to the base Mel-filter responses [17]. In contrast, we build on top of a first-layer, filterbank representation and learn the templates of a subsequent convolutional layer.

The idea to generalize CNNs beyond translations has been previously explored in tiled convolutional networks [18], that redefine weight-sharing for robustness to 2D rotations and scal-

ing, and scattering networks defined using wavelets over roto-translation groups [19]. Convolutional maxout networks [20] use an equivalent double pooling over translations and neuron outputs, without the group relationship in the pooled neuron functions and with a single (max) nonlinearity. Other generalizations of CNNs include kernel based interpolation for invariance to affine groups [21], layers of kernel maps [22] or the generalization of convolutions from spatial grids to graphs [23].

Most of these approaches however were developed for image data, where common transformations are relatively easy to model. On the contrary, the transformations relevant for speech are normally less intuitive or not tractable. This imposes difficulties in learning an invariant convolution/pooling module through back-propagation in our framework (Sec. 3.4). A simple workaround will be employed via augmenting the input samples. On its own, data augmentation is a widely-applied method for neural network-based speech recognition [24, 25].

3. Convolutional Neural Networks over Transformation Groups

3.1. Symmetries, groups and invariant representation

Symmetries of an object are transformations that keep a given property unchanged. A simple example is rotations of a circle centered at the origin of the 2D plane — although the points composing the object moved, the *shape* is *invariant* under rotations. Symmetries, which typically form a *group* [26], are being actively pursued as a theoretical tool for understanding *invariant feature representations* [7, 8, 9, 21, 27, 28], obtained from multilayer convolutional networks.

The symmetries of interest for signal representation are transformations that keep the "identity" unchanged, where by identity we refer to the classes of a learning problem. For example, let $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ be a speech segment, and $\Phi_g(\cdot) : \mathcal{X} \to \mathcal{X}$ a mapping between different speakers that preserves the sublexical identity, e.g., the phoneme label. Assume the mapping is parametrized by an element $g \in G$ of an abstract group G. Let $\rho(y|\boldsymbol{x})$ be the posterior distribution of the segment label $y \in \mathcal{Y}$ given the observation $\boldsymbol{x}$. A *symmetry* in this case would imply

$$\rho(y|\boldsymbol{x}) = \rho\left(y|\Phi_g(\boldsymbol{x})\right), \tag{1}$$

for all $\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$, $g \in G$, or equivalently that the transformation Φ_g is *irrelevant* for the purpose of inference on the phoneme labels. Then a feature map $F : \mathcal{X} \to \mathcal{F}$ that is *invariant* to the transformations Φ_g means

$$F(\boldsymbol{x}) = F\left(\Phi_g(\boldsymbol{x})\right) \tag{2}$$

for all inputs $\boldsymbol{x}$ and transformation $g \in G$. To avoid *degenerate* feature maps, we also require that F is *selective*, meaning that if $\boldsymbol{x}'$ is *not* a transformed version of $\boldsymbol{x}$, i.e. $\nexists g \in G : \boldsymbol{x}' = \Phi_g(\boldsymbol{x})$, then $F(\boldsymbol{x}') \neq F(\boldsymbol{x})$. It is easy to show that under the symmetry assumption, an invariant and selective feature map $F(\boldsymbol{x})$ is a *sufficient statistic* [29, 30] of the original signal $\boldsymbol{x}$. Therefore, learning on $F(\boldsymbol{x})$ is statistically equivalent to learning on $\boldsymbol{x}$. On the other hand, since F collapses equivalent points into one canonical representation, the size of the input sample space and the hypothesis space get reduced. Consequently, the *sample complexity* of the learning problem is reduced [8, 9].

3.2. Convolution and pooling over group orbits

The group orbit $O_{\boldsymbol{x}}$ is the set of all transformed versions of $\boldsymbol{x}$:

$$O_{\boldsymbol{x}} = \{\Phi_g(\boldsymbol{x}) | g \in G\} \subset \mathcal{X} \tag{3}$$

Our model is based on the fact that the feature map $F(\boldsymbol{x}) : \boldsymbol{x} \mapsto O_{\boldsymbol{x}}$ is naturally invariant and selective. We first show that when the transformations Φ_g are translations, convolutional neural networks (CNNs) are approximating this feature map. We then extend the architecture to more general transformations.

Recall that CNNs consist of interleaved spatial convolution and pooling layers. In the convolution layer, the convolution between an input $\boldsymbol{x}$ and a template (or filter) $\boldsymbol{t}$ is computed as

$$\boldsymbol{y}[i] = \sum_j \boldsymbol{x}[j]\boldsymbol{t}[i-j] = \sum_j \boldsymbol{x}[j]\tilde{\boldsymbol{t}}[j-i] = \langle \boldsymbol{x}, \Phi_i(\tilde{\boldsymbol{t}}) \rangle, \tag{4}$$

where $\tilde{\boldsymbol{t}}[i] = \boldsymbol{t}[-i]$ is the mirror reflected template, and Φ_i is an operation that shifts the signal by i samples. Succinctly, the convolution layer is computing the inner product between $\boldsymbol{x}$ and transformed templates $\Phi_i(\tilde{\boldsymbol{t}})$. In the pooling layer, a statistic (e.g., MEAN or MAX) of the output values $\boldsymbol{y}[i]$ is computed, typically within a local pooling range. We show that this is actually approximating the invariant feature map $F(\boldsymbol{x}) = O_{\boldsymbol{x}}$ under the translation group.

The orbit $O_{\boldsymbol{x}}$, being a set of signals, is uniquely associated with the probability distribution induced by the Haar measure on the transformation group G. From Cramér-Wold theorem [31], this high-dimensional distribution can be characterized in terms of projections onto unit vectors (templates). Specifically, for a unit-normed template $\boldsymbol{t}$, the set of inner product values

$$\{\langle \Phi_g(\boldsymbol{x}), \boldsymbol{t} \rangle | g \in G\} \tag{5}$$

specifies a one-dimensional distribution. Given enough templates, the set of one-dimensional distributions uniquely characterizes the original orbit $O_{\boldsymbol{x}}$. When the group is *unitary*, we could equivalently apply the (inverse) transformation to the template $\boldsymbol{t}$ and leave $\boldsymbol{x}$ unchanged:

$$\langle \Phi_g(\boldsymbol{x}), \boldsymbol{t} \rangle = \langle \boldsymbol{x}, \Phi_g^{-1}(\boldsymbol{t}) \rangle = \langle \boldsymbol{x}, \Phi_{g^{-1}}(\boldsymbol{t}) \rangle \tag{6}$$

Comparing this with Eq. (4), one can see that this is exactly what the convolution layer in CNNs is computing. The next step (the pooling layer) can be seen as finding a suitable description for those one-dimensional distributions. Natural choices include discrete histograms and moments, or the MAX statistic which is typically used in CNNs.

In summary, a CNN is approximating the invariant feature representation that maps $\boldsymbol{x}$ to the orbit $O_{\boldsymbol{x}}$, via interleaved convolution and pooling along the translation group. The above derivation applies to more general transformations than translations, under the main assumption of forming a unitary and locally compact group. Therefore, by replacing the translation group with other transformations, we get a natural generalization of CNNs. In this paper, we make the case for generic transformations for speech signals, specifically vocal tract length (VTL) perturbations, which we describe in Section 3.4.

3.3. Discriminative learning of optimal templates

The characterization of high dimensional distributions through one-dimensional projections is exact if *all* (infinitely many) templates on the unit sphere are used (Cramér-Wold theorem). For a finite set of input signals and finite number of *random templates* a *Johnson-Lindenstrauss lemma*-type argument can guarantee approximate characterization [9, 10]. The use of (random) samples of natural signals as templates [11] is also supported through the biologically plausible hypothesis that the collection and neural memorization of transformed signals is performed

through unsupervised observations (memory-based learning). The template selection can also satisfy *optimality* criteria from a theory perspective. For example, it can be shown that Gabor functions, which are also biological model functions of cortical simple cells, are optimal templates when maximal invariance to both scale and position is desired [10].

In this paper we explore learning the templates in a data-driven way, by jointly training for the templates and classifier that minimize a classification cost function on a labeled training set. This is the standard way of setting the filter weights (or choosing the templates) in a CNN through training data and supervision. Discriminatively learned templates have proven to be very effective in different speech tasks, including phone recognition and large vocabulary recognition [14, 15].

Given a loss function L (e.g., cross-entropy), the templates can be learned through (stochastic) gradient descent, by moving iteratively in the negative gradient direction

$$\boldsymbol{t} \leftarrow \boldsymbol{t} - \alpha \frac{\partial L}{\partial \boldsymbol{t}} \tag{7}$$

where α is the *learning rate* and the gradient vector $\partial L / \partial \boldsymbol{t}$ is computed via the chain-rule (back propagation) as

$$\frac{\partial L}{\partial \boldsymbol{t}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{t}} \frac{\partial L}{\partial \boldsymbol{y}} = \sum_i \frac{\partial \boldsymbol{y}[i]}{\partial \boldsymbol{t}} \frac{\partial L}{\partial \boldsymbol{y}[i]} = \sum_i \frac{\partial \Phi_i}{\partial \boldsymbol{t}} \boldsymbol{x} \frac{\partial L}{\partial \boldsymbol{y}[i]} \tag{8}$$

where $\boldsymbol{y}[i]$ (see Eq. (4)) is the output of the convolution layer, and $\partial L / \partial \boldsymbol{y}[i]$ is itself recursively computed via back-propagation from upper layers. In standard CNNs, Φ_i is a translation, so the derivative $\partial \Phi_i / \partial \boldsymbol{t}$ is straightforward to compute. However, for arbitrary transformations Φ_i, the derivative might be generally difficult to compute analytically. As a workaround, we modify the network architecture slightly: instead of applying the transformations to the templates $\boldsymbol{t}$, as in Eq. (6), we leave the transformations to the input $\boldsymbol{x}$. In this way, $\boldsymbol{y}[i]$ will depend only linearly on $\boldsymbol{t}$, making the derivative $\partial \boldsymbol{y}[i] / \partial \boldsymbol{t}$ easy to compute.

3.4. Vocal tract length perturbation and invariant module

Vocal tract length (VTL) normalization [3] is a widely applied technique in speech recognition for modeling and removing inter-speaker variability due to vocal tract length [4, 32, 33]. A warp factor, estimated for each speaker, is used to normalize the corresponding speech data to account for the difference of their vocal tract length from the *canonical mean*. The reverse, i.e., introducing perturbations of the VTL, has been recently explored as a means for augmenting the training set [24, 25]. In this paper, we will consider VTL as a transformation that preserves the posterior probability over phoneme labels, i.e., Eq. (1).

We propose an invariant signature extraction using group-convolution/group-pooling modules, one per template t, illustrated in Fig. 1. The input signal $\boldsymbol{x}$ is a speech frame, represented as 40-dimensional Mel-fiterbank, together with 7 context frames on both sides of the time index. VTL warpings, through a piecewise linear frequency warping function (implemented in Kaldi [34]), with 9 evenly distributed warp factors in $[0.9, 1.1]$ are applied to $\boldsymbol{x}$ to create transformed inputs. A template $\boldsymbol{t}$, randomly initialized and trained discriminatively, is a localized patch that shifts on both frequency and VTL axes. This is equivalent to a convolution operator in the joint transformation space. Similar to frequency-based CNNs for speech recognition [14, 15], the templates cover the whole time axis. In contrast, we introduce an additional dimension over the VTL transformations.

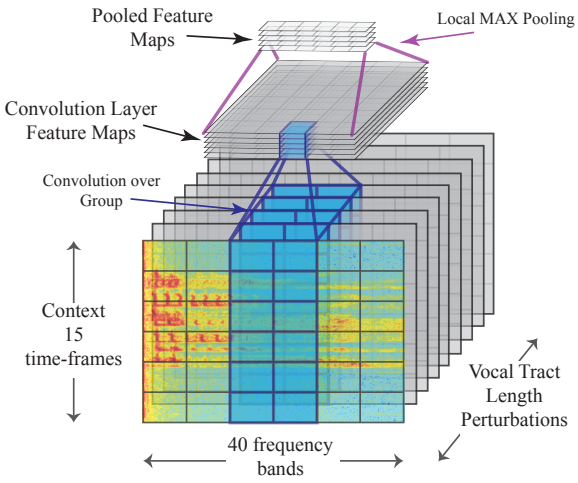

Figure 1: Invariant group-convolution and pooling module.

	TIMIT dev	TIMIT test	WSJ dev	WSJ test
DNN	58.70	58.25	58.33	63.65
CNN	63.07	62.41	61.07	66.78
VTL-CNN	65.26	64.62	64.27	70.08

Table 1: Frame classification accuracy.

The dot product values between templates and transformed input signals form the convolution layer feature maps; MAX pooling is performed, *locally across frequency regions*, similar to CNNs to avoid global frequency shift invariance, and *globally across all warps*.

4. Experimental Evaluation

We empirically evaluate the proposed model and compare it with canonical CNNs and fully connected Deep Neural Networks (DNNs). The input to each convolution/pooling module is a 15-by-40-by-9 dimensional tensor, described in Sec. 3.4. For the convolution layer, we use 91 filters with a kernel size of 8 and 3 on the frequency and VTL axis respectively, with a stride of 1. Pooling is over local regions of size 6 over frequency and global over VTL. The full network consists of the convolution layer, pooling layer, and two fully connected layers (1024 units) with Rectified Linear Units (ReLUs) and a final linear layer to predict the posterior probability of frame labels. Comparisons with the proposed model are performed under a similar setting (filter and pooling kernel sizes). To ensure the same number of parameters in the convolution layer, we use 3×91 filters for the conventional CNNs[1]. We also report baseline comparisons with a 3 hidden layer DNN, resulting by replacing convolution-pooling with a densely-connected layer.

4.1. Frame classification on TIMIT and WSJ

We evaluate the proposed model on two standard datasets: TIMIT [35] and Wall Street Journal (WSJ) [36]. We use the Kaldi speech recognition toolkit [34] for data pre-processing. Particularly, we compute the frame-level, 40-dimensional Mel-filterbank features as base data representation. The feature

[1] As the VTL dimension for our network disappears after pooling, the output of the pooling layer for CNNs is actually 3 times larger, leading to 3 times more parameters in the first fully connected layer.

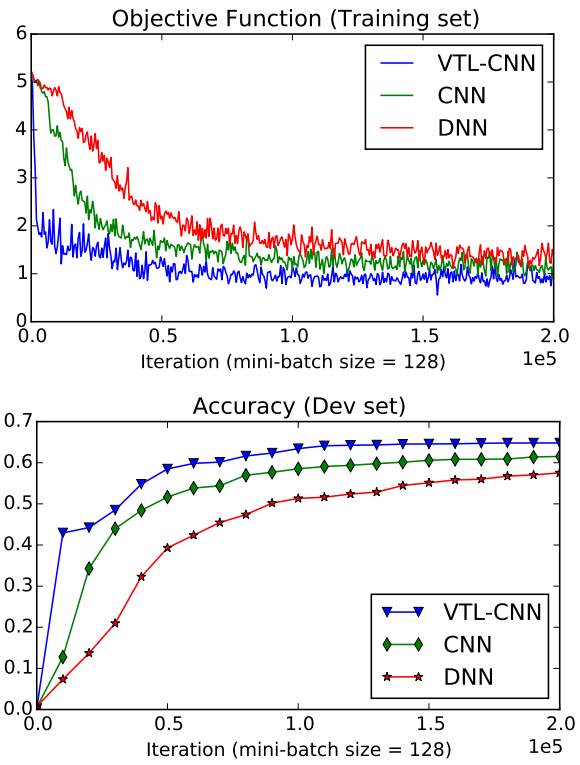

Figure 2: Training cost and dev set frame accuracy against the training iterations on the TIMIT dataset.

extraction module (convolution/pooling layer), together with the classifier (densely connected layers) are trained jointly, and all the parameters are initialized randomly without pre-training. The labels are monophone HMM states, 144 for the 48 TIMIT phone set, generated by running forced alignment on the ground-truth transcriptions with GMM-HMM decoders based on MFCC features. All networks were trained using mini-batch gradient descend and adaptive learning rate, with initial rate and momentum validated from the development sets.

For TIMIT, we use the standard *train*, development (*dev*) and *core-test* data partitioning. Figure 2 shows the training costs and frame accuracy on the *dev* set with respect to the training iterations. The DNN, despite having more parameters, performs worse than the CNN, consistently with recent work on CNNs for speech recognition [14, 15]. Our model, denoted as VTL-CNN, works better than both CNN and DNN. It also converges faster, which might be related to the fact that in principle an invariant representation reduces the learning sample complexity [8, 10], i.e., the network needs less data for a given level of performance. For WSJ, we use the *si84-half*, *dev93*, and *eval92* splits given by the *s5* recipe of Kaldi as the training, development, and test sets. Due to space limitations, we omit the performance curve for WSJ. The frame accuracy on the development and test sets for both datasets is summarized in Table 1.

4.2. Comparison with data augmentation

It is natural to ask whether the performance gain of our model comes from pooling over generalized transformations or simply because it sees "more" data, i.e., all the VTL warped speech frames. We explore this in depth by training the same models on the original and a VTL augmented training set, using the same warpings. The augmented dataset is explicitly shuffled to avoid

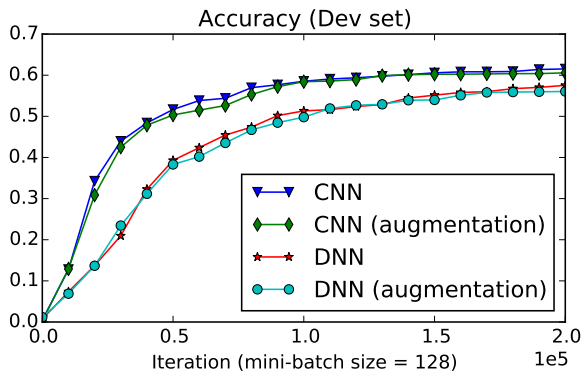

Figure 3: Performance of models trained on original and VTL-augmented TIMIT.

similar training samples in a mini-batch. The results for TIMIT are shown in Fig. 3. For both CNNs and DNNs, there is no significant difference between the networks trained on the original or the augmented sets; the number of parameters was kept the same and no learning parameter re-validation was performed. The VTL-CNN performance (Fig. 2) is thus not matched, indicating that the augmented data by itself does not necessarily boost performance; the architecture for explicit invariance of the proposed model is more important in this sense.

It is worth mentioning that alternative techniques can be applied for making better use of augmented training data. For example, in [24], *model averaging* is used to combine the predictions from VTL warped data. *Speaker-specific estimates* of the warping factor are used in [25] to introduce small VTL perturbations. In contrast, our feature is expected to be (approximately) invariant to VTL variation without requiring an estimate of the "true" warping factor in advance.

5. Conclusions

In this paper, we outlined a framework for learning a feature map that is approximately invariant under general transformation groups. We showed how CNNs are a special case of this framework for representations invariant to local frequency shifts. We proposed a new model that generalizes the convolutional networks to include specific invariance with respect to VTL perturbation and demonstrated its effectiveness on standard speech datasets. This work is an extension of our previous model [12] for invariant speech representations in the sense that we are discriminatively learning an optimal set of templates in a data-driven way by jointly training the feature extraction module and the classifier. The proposed model and extensions will be subsequently evaluated on phone recognition as well as large vocabulary continuous speech recognition tasks. To decrease the computational cost of explicitly transforming the input signals, while still avoiding computing derivatives of non-tractable transforms, we are actively exploring the use of more tractable, parametric transformations.

6. Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. CZ and SV acknowledge the support of a Nuance Foundation Grant. LR acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC.

7. References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning – From Theory to Algorithms*. Cambridge University Press, 2014.

[2] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 1996, pp. 346–348.

[4] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, Sep. 2005.

[5] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, Jan. 2014.

[6] T. O. Sharpee, C. A. Atencio, and C. E. Schreiner, "Hierarchical representations in the auditory cortex," *Curr. Opin. Neurobiol.*, vol. 21, no. 5, pp. 761–767, Jun. 2011.

[7] S. Soatto, "Steps Towards a Theory of Visual Information: Active Perception, Signal-to-Symbol Conversion and the Interplay Between Sensing and Control," *CoRR*, vol. abs/1110.2053, Oct. 2012. [Online]. Available: http://arxiv.org/abs/1110.2053

[8] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations with low sample complexity: the magic of sensory cortex or a new framework for machine learning?" CBMM Memo 001, 2014. [Online]. Available: http://arxiv.org/abs/1311.4158

[9] F. Anselmi, L. Rosasco, and T. Poggio, "On invariance and selectivity in representation learning," *CoRR*, vol. abs/1503.05938, 2015. [Online]. Available: http://arxiv.org/abs/1503.05938

[10] F. Anselmi and T. Poggio, "Representation learning in sensory cortex: a theory," CBMM Memo 026, Nov. 2014. [Online]. Available: http://arxiv.org/abs/1110.2053

[11] S. Voinea, C. Zhang, G. Evangelopoulos, L. Rosasco, and T. Poggio, "Word-level invariant representations from acoustic waveforms," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, Singapore, Sep. 14-18 2014.

[12] C. Zhang, S. Voinea, G. Evangelopoulos, L. Rosasco, and T. Poggio, "Phone classification by a hierarchy of invariant representation layers," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, Singapore, Sep. 14-18 2014.

[13] C. Zhang, G. Evangelopoulos, S. Voinea, L. Rosasco, and T. Poggio, "A deep representation for invariance and music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[14] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[15] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, Sep. 2014.

[16] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[17] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 297–302.

[18] Q. Le, J. Ngiam, Z. Chen, D. J. hao Chia, P. W. Koh, and A. Ng, "Tiled convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 1279–1287.

[19] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013, pp. 1233–1240.

[20] L. Tóth, "Convolutional deep maxout networks for phone recognition," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, Singapore, Sep. 14-18 2014.

[21] R. Gens and P. M. Domingos, "Deep symmetry networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2537–2545.

[22] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2627–2635.

[23] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6203

[24] N. Jaitly and G. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[25] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 4-9 2014, pp. 5582–5586.

[26] Y. Kosmann-Schwarzbach, *Groups and Symmetries, From Finite Groups to Lie Groups*. Springer, 2010.

[27] S. Mallat, "Group Invariant Scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[28] T. Cohen and M. Welling, "Learning the irreducible representations of commutative lie groups," in *International Conference on Machine Learning (ICML)*, Feb. 2014.

[29] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Wiley, 1994.

[30] S. Soatto, "Visual scene representations: Sufficiency, minimality, invariance and approximations," *CoRR*, vol. abs/1411.7676, 2014. [Online]. Available: http://arxiv.org/abs/1411.7676

[31] H. Cramér and H. Wold, "Some theorems on distribution functions," *Journal of the London Mathematical Society*, vol. s1-11, no. 4, pp. 290–294, 1936.

[32] Y. Qiao, M. Suzuki, and N. Minematsu, "Affine invariant features and their application to speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 4629–4632.

[33] R. E. Turner, T. C. Walters, J. J. M. Monaghan, and R. D. Patterson, "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2374–2386, Apr. 2009.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "DARPA, TIMIT acoustic-phonetic continuous speech corpus," *National Institute of Standards and Technology*, 1990.

[36] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. of the Workshop on Speech and Natural Language*, ser. HLT '91, 1992, pp. 357–362.