Phone Classification by a Hierarchy of Invariant Representation Layers

Chiyuan Zhang[*], *Stephen Voinea*[*], *Georgios Evangelopoulos*[*†],
Lorenzo Rosasco[*†], *Tomaso Poggio*[*†]

[*] Center for Brains, Minds and Machines | McGovern Institute for Brain Research at MIT
[†] LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
{chiyuan,voinea,gevang,lrosasco}@mit.edu, tp@ai.mit.edu

Abstract

We propose a multi-layer feature extraction framework for speech, capable of providing invariant representations. A set of templates is generated by sampling the result of applying smooth, identity-preserving transformations (such as vocal tract length and tempo variations) to arbitrarily-selected speech signals. Templates are then stored as the weights of "neurons". We use a cascade of such computational modules to factor out different types of transformation variability in a hierarchy, and show that it improves phone classification over baseline features. In addition, we describe empirical comparisons of a) different transformations which may be responsible for the variability in speech signals and of b) different ways of assembling template sets for training. The proposed layered system is an effort towards explaining the performance of recent deep learning networks and the principles by which the human auditory cortex might reduce the sample complexity of learning in speech recognition. Our theory and experiments suggest that invariant representations are crucial in learning from complex, real-world data like natural speech. Our model is built on basic computational primitives of cortical neurons, thus making an argument about how representations might be learned in the human auditory cortex.

Index Terms: Invariance, Auditory Cortex, Phonetic Classification, Convolutional Network

1. Introduction

Natural speech signals are complex sources of acoustic information due to the many factors of intrinsic and extrinsic variability such as vocalization variations (accent, pronunciation, articulation, pitch, volume, rate, etc.), mixing with background noise and reverberations or filtering by the sensory and transmission systems, just to name a few. Due to the effect of such *identity-preserving* variability, the set of perceivable speech units is infinitely large compared to the number of commonly used words in a language (e.g., the second edition of the *Oxford English Dictionary* lists 171,476 words in current use). Consequently, an effective acoustic feature representation, i.e., one that can discount variability while preserving identity, is a crucial component of automatic speech recognition (ASR) systems.

Representations that are robust to noise, invariant to task-irrelevant transformations, and rich in discriminative information are usually the result of clever feature engineering and speech domain expertise. Examples of such representations are the widely used the mel-frequency cepstral coefficients (MFCCs) [1] and perceptual linear predictive coefficients (PLPs) [2]. Despite feature sophistication however, machine-based speech recognition still under-performs compared to recognition by humans [3, 4]. This gap motivates a direction of studies towards understanding the processing and representation in the human auditory cortex [5, 6] and developing biologically inspired speech features [7].

Representation learning is a relatively mature field in machine learning, aimed at automatically extracting effective data representations from low-level inputs, such as sound waveforms or spectrograms. For example, dictionary learning and sparse coding have been used to learn representations for speech [8, 9, 10] and generic audio signals [11]. Similarly, bottleneck features [12, 13], are rediscovered and becoming increasingly popular with the recent advents of deep networks [14, 15].

Our contributions in this paper relate to both of the above directions, namely biologically-inspired and data-driven features, based on a theory and a hypothesis for processing in the visual ventral stream [16]: 1) We propose a biologically plausible computational module by wiring "neurons" in networks with predefined architectures, mimicking *simple-complex cells* [17] (Sec. 3). We show how these modules can compute representations that are provably invariant to compact group transformations and approximately invariant to more general classes of (smooth) transformations. 2) To factor-out different sources of variability, we build a layered architecture by re-using the basic modules (Sec. 4). 3) We present phonetic classification results on TIMIT dataset [18] (Sec. 4) and empirically evaluate different schemes for learning the module weights (Sec. 5).

2. Related Work

Acoustic modeling with deep neural networks has recently largely outperformed conventional pipelines based on MFCCs and Gaussian Mixture Models (GMMs) [19]. Different variants, successfully applied for speech, include fully connected architectures [20], convolutional networks with pooling over time [21] or frequency [22] and deep recurrent networks [23, 24]. Empirical studies have also shown that deep models can subsume many carefully designed speaker adaptation techniques [25]. Moreover, it was observed that the intermediate layer representations (bottleneck features) taken from a deep neural network can considerably improve the performance GMM-based acoustic modeling [14, 15]. However, there is still limited formal understanding of how or why deep architectures lead to effective representations.

Scattering networks [26] are a closely related framework to our model, towards a theory on how multilayer, hierarchical architectures can lead to stable and invariant representations. Scattering transforms compute feature maps via cascades of convolutions and modulus operators. A representation is then obtained by concatenating the output of successive layers. Our work shares the same of goal of building and explaining invariant representations through multilayer architectures. However, by using biologically plausible computational modules, we also reason about a possible mode of memory-based learning of invariances in the auditory cortex. In addition, we consider a model where different transformations are progressively factored out on different layers.

3. Invariant Representation Learning

In this section, we present the theory and implementation aspects of a biologically plausible module for computing features invariant to transformations and motivate a memory-based scheme for learning invariant representations from data.

3.1. Invariance and Learning Complexity

Real world signals are complicated in many different ways. One could argue though that the transformations that preserve semantic identity account for almost all the complexities in recognition tasks [16]. Such complexities have been the major difficulty in applying machine learning algorithms to real world problems. In recent years, due to advances in hardware, effective regularization techniques and improved optimization algorithms, sophisticated models are being trained on huge datasets containing multiple aspects of real world complexities. For example, $5,780$ hours of Google Voice input data was used in [27]. Augmenting the training set by applying label-preserving transformations to the data is also a technique known to boost system performance [28, 29].

Despite their impressive recognition performance, such systems are diverging from principles evident in human intelligence. A child, for example, can learn to recognize individual words by hearing a few examples of them, with limited supervision. Studies show that a two-year baby would probably be exposed to roughly 1000 hours of speech [30], which is mostly *unlabeled*. Both theoretical and empirical results [31, 16] support the argument that with a representation invariant to irrelevant transformations, fewer training samples (*sample complexity*) and simpler learning models (*model complexity*) could be used to achieve the same learning performance.

3.2. Group Theory and Invariant Representations

Mathematically, invariance is defined through transformations that leave some quantity unchanged. In our case, we are particularly interested in *identity-preserving* transformations. Such transformations (for example the rigid motion of objects) can often be modeled as *groups* [32].

Consider the space $\mathcal{X}$ of all possible input signals (e.g., speech sound waveforms), and an identity-preserving group G that acts on $\mathcal{X}$. That is, for the identity space, or label space $\mathcal{Y}$ (e.g., the set of all phonetic labels), and the ground-truth label map[1] $\mathfrak{g} : \mathcal{X} \to \mathcal{Y}$, we have $\mathfrak{g}(x) = \mathfrak{g}(g \circ x)$ for all $x \in \mathcal{X}, g \in G$. The goal is to construct a *representation map* $\mathsf{r} : \mathcal{X} \to \tilde{\mathcal{X}}$ to the feature space $\tilde{\mathcal{X}}$ in which the transformations of G keep the representation unchanged. In other words, $\mathsf{r}(x) = \mathsf{r}(g \circ x)$ for all $x \in \mathcal{X}, g \in G$.

The actions of G define an *equivalence relation* $\sim$ on $\mathcal{X}$ as: $x \sim x'$ if and only if $\exists g \in G$ such that $x = g \circ x'$. The equivalence classes $[x] = \{x' \in \mathcal{X} : x \sim x'\} = \{g \circ x : g \in G\}$ are called *orbit*s of G because $[x]$ is the "orbit" of applying all $g \in G$ to x. This equivalence relation induces a *quotient map* $\mathsf{q} : x \mapsto [x]$ from $\mathcal{X}$ to the *quotient space* $\mathcal{X}/G$ (a.k.a. the *orbit space*). It is easy to see that q is invariant to actions of G:

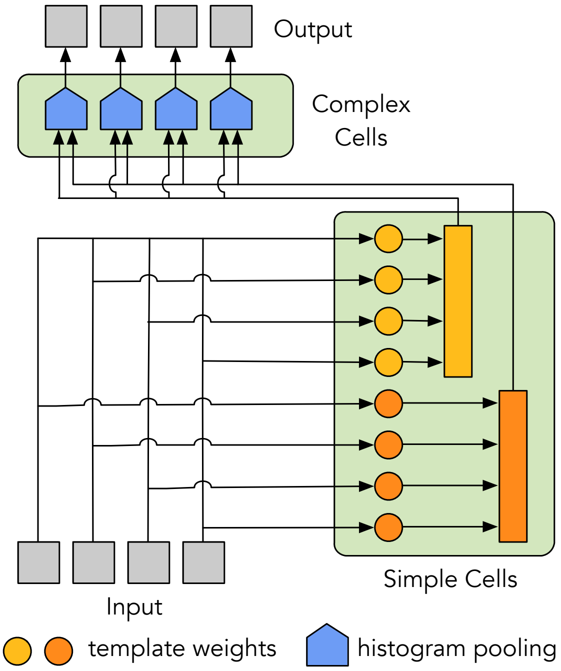

Figure 1: Illustration of an invariant representation module.

$$\mathsf{q}(x) = [x] = [g \circ x] = \mathsf{q}(g \circ x), \quad \forall x \in \mathcal{X}, g \in G$$

So letting $\tilde{\mathcal{X}} = \mathcal{X}/G$, $\mathsf{r} = \mathsf{q}$ will be an invariant representation map. Note this map is also discriminative, because $[x] \neq [x'] \Leftrightarrow \nexists g \in G, s.t.\ g \circ x = x'$, so x and x' will have different labels / identities.

3.3. Neuronal Modules for Computing Invariant Maps

For a compact group G and a fixed signal $x \in \mathcal{X}$, the normalized *Haar measure* on G induces a probability distribution p_x supported on all the signals in the orbit $[x]$. The high dimensional distribution p_x, having the same role as the orbit $[x]$, can be characterized by the collection of one dimensional distributions $\mathsf{p}_{\langle x,t \rangle}$ induced by projecting onto all the vectors t on the unit sphere [33, 16].

Under this reasoning, we propose a concrete model implementation of an invariant map, using only basic neuron primitives (high-dimensional dot product and nonlinearity) and several steps of approximation. First, a finite set of (randomly chosen) signals $t^1, \ldots, t^K$, called templates, are used. Then, the one dimensional distributions $\mathsf{p}_{\langle x,t^k \rangle}$ are approximated by discrete estimates (histograms of N bins). Specifically, N shifted *step functions* η_n are integrated to get the cumulative histogram counts for the projections on each t^k:

$$\mu_n^k(x) = \int_G \eta_n(\langle g \circ x, t^k \rangle)\, dg \qquad (1)$$

For neural modules, we use a smooth neuron nonlinearity (e.g., sigmoid) to approximate the step function. For a unitary G, it holds that $\langle g \circ x, t^k \rangle = \langle x, g^{-1} \circ t^k \rangle$. Thus, to compute (1), we do *not* need to obtain all the transformed versions of a signal x; instead, at training time, we *observe* and store a collection

[1] For simplicity, we assume this is a deterministic mapping.

of templates $\{t^k\}_{k=1}^K$ and all transformed versions $\{g \circ t^k\}_{g \in G}$ for each one, i.e. $\mu_n^k(x) = (1/|G|) \sum_{g \in G} \eta_n(\langle x, g \circ t^k \rangle)$.

To get the representation for new inputs, the *normalized* inner products with all transformed templates are computed and then histogram-pooled over each template orbit. The output is the $N \times K$-dimensional vector formed by K components of N histogram bins each. An illustration of such an invariant representation module is shown in Fig. 1. This module is reminiscent of the simple-complex cell modules [17], where inner products (filtering) are computed by the *simple cells* and pooling operations are carried out by the *complex cells*.

Note that in some cases it might be empirically favorable to use moment pooling with a nonlinearity function of the form $\eta_n(\cdot) = (\cdot)^n$ instead of histograms. This covers several special cases such as the *energy model* of complex cells [34] for $n = 2$ and *mean pooling* for $n = 1$. With proper normalization, *max pooling* [35] is also approximated for $n \to \infty$.

3.4. General Smooth Transformations

For groups that are only *locally compact*, we could pool over a subset $G_0 \subset G$ to generate a local signature as

$$\mu_n^k(x) = \frac{1}{V_0} \int_{G_0} \eta_n(\langle x, g \circ t^k \rangle)\, dg, \quad G_0 \subset G \qquad (2)$$

where $V_0 = \int_{G_0} dg$ is a local normalization constant in order for a valid probability distribution to be defined. It can be shown that this local signature is *partially* invariant to a restricted subset of transformations when the input and the templates have a *localization* property [16]. Dealing with general non-group transformations is more complicated. The basic idea is to rely on smoothness, and linearize locally at some *key transformations*. Approximate invariance could be achieved by combining these locally computed signatures. The more complicated the transformations are, the more *key transformations* are needed for a good approximation. Please refer to [16] for details. As a remark, not all transformations can be handled in our framework. For non-smooth transformations, it is difficult to control the behavior of local approximations in general. Furthermore, for non-invertible transformations (e.g., sound occlusions), one cannot recover the discriminative information that has been permanently lost.

3.5. Memory-based Learning of Invariance

Learning of invariant modules could be implemented through “memorizing” the transformed templates at training time and storing them in the “synapses” (i.e., the weights of module inputs). In this paper, we consider two modes of learning through stored transformations: *explicit* refers to generating the transformed templates by explicitly modeling the transforms and sampling by synthesis; *implicit* refers to forming a collection of grouped samples from the training data and assuming each group is generated via the same, *unknown* transformation. Specific examples of the two learning modes are given in Sec. 5.

One merit of our framework is that we do not need to explicitly know what the underlying transformation is in order to be invariant to it. The implicit mode, in the form of *unsupervised learning*, is very natural in vision, e.g., through the continuous observation of moving (transformed) objects. The temporal continuity accounts for collecting the set of a transformed template without knowing the identity of either the template (object) or the underlying transformation. Although the analogies for the auditory system are not apparent, in Sec. 5 we empirically study different schemes for implicit learning.

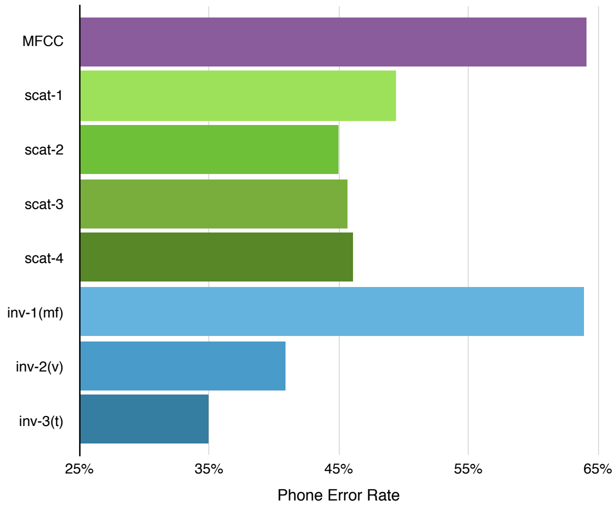

Figure 2: Phone classification error rates on TIMIT Core, using one-vs-rest regularized linear classifiers (scat-n: n-th order scattering [26], inv-n: n-th layer invariant representation).

4. Evaluation of a Layered Representation

In this section, we propose and empirically evaluate a multilayer architecture for invariance to multiple transformations, based on stacking multiple invariance modules. The proposed architecture is used to extract representations of segmented phones for phonetic classification on the TIMIT dataset [18].

Multilayer Invariant Representation: A three-layer representation is formulated by stacking modules for invariance to local frequency shifts, vocal tract length (VTL) and signal tempo. Note that VTL can be substantially the largest source of variability in the waveform of vowel sounds [36], and VTL normalization [37] is a commonly applied speaker adaptation technique in ASR systems.

The *first layer* computes standard Mel-frequency filter bank features from the raw waveform, which can be seen as a specific invariance module under our framework: the transforming templates are frequency-shifted Fourier basis; locally-weighted mean pooling (mel-filter energies) is applied on the projection of the input on the template set. Since pooling is only over a local subset of frequency-shifting transformations, the representation is only approximately invariant to small frequency shifts. The *second layer* attempts to handle speaker variability through a module of VTL-invariance. For obtaining the sets of transformed templates (explicit learning of the simple cells), we randomly sample frames from the training data. Each sample is a single template t^k, which we explicitly transform by VTL variations [37] with warping factors ranging in $[0.8, 1.2]$. Each template is thus associated with samples from a VTL orbit set. The *third layer* is aimed at compensating for the variations of the pace or speed of speech signals. Again, randomly sampled training frames are used as templates, transformed by varying the tempo within the range $[0.5, 1.9]$.

Experimental Setting: TIMIT consists of phonetically-balanced, prompted English speech recordings, including 6300 sentences spoken by 630 speakers from 8 dialect regions. We use the standard Train / Test partition and report phone clas-

Explicit					Implicit			Reference		
warp	pitch	tempo	time shift	VTL	kmeans	category	SA	MFCC	scat-1	scat-2
41.85%	42.39%	43.35%	43.86%	40.89%	46.26%	46.79%	42.99%	64.15%	49.37%	44.94%

Table 1: Phone classification error rate using different invariance modules.

sification error rates on the Core Test set. The dataset phonetic labels contain 61 classes. Following standard practices, we fold the 61 classes into 39 categories [38]. We formulate a multiclass classification problem, using one-vs-rest regularized linear regression classifiers, with the regularization coefficients decided by 5-fold cross validation. We explicitly avoided using kernel-based, non-linear methods in order to demonstrate the effectiveness of the representation for small model complexity.

Baselines and Results: We compare the multilayer model with the deep scattering spectrum (DSS) [26] representation and a baseline using MFCC (with Δ and $\Delta\Delta$) under the same setting. Fig. 2 shows phone error rates for all systems, including results from intermediate layers of the proposed multilayer architecture (denoted inv-1(mf), inv-2(v) and inv-3(t) for Mel, VTL and tempo layers respectively). First- to fourth-order DSS (denoted scat-n for each order n) all outperform the MFCC baseline significantly. However, adding layers beyond two does not improve performance, which was also observed in [26] (Note: Even though we are using the software package provided by the authors to compute scattering transforms, the reported errors are different from the ones in [26] as we are using a much weaker classifier). The performance of the first invariance layer inv-1(mf) is slightly better than the MFCC baseline. The second layer inv-2(v), by compensating for VTL variations, boosts the classification accuracy significantly, outperforming all orders of deep scattering transforms. The additional tempo layer inv-3(t) further reduces the overall error to 35%.

5. Evaluation of Explicit and Implicit Sets of Transformed Templates

In this section, we evaluate the relative significance of different transformations for speech representations, along with a preliminary study on data-driven methods for invariant layers without explicit transformation encoding. Phone error rates on TIMIT Core Set for different invariance modules are shown in Table 1.

A set of explicit transformations, for typical sources of speech variability, are considered for building transformation-specific invariant representations (see *Explicit* in Table 1):

- warp: transform the template signals $t[n]$ as $g_\varepsilon \circ t[n] = t_\varepsilon[n] = t[(1+\varepsilon)n]$ with warping factors ε within the range $[-0.4, 0.4]$ (step 0.1).
- pitch: scale the pitch of the template signals within the range $[0.7, 1.5]$ (step 0.02) while preserving the tempo.
- tempo: modify the tempo of the template signals within the range $[0.5, 1.9]$ (step 0.03) for fixed pitch.
- time shift: shift the time index of the template signals within the range of one third of the window size.
- VTL: apply vocal tract length perturbation to the template signals with warping factors within the range $[0.8, 1.2]$ (step 0.01).

Our results are consistent with studies on the role of VTL as a major source of transformation variability in speech signals [36]. A signature from pitch-based modules also works quite well, probably due to the similar effects of pitch to VTL variations. Surprisingly, the warping-based module also works very competitively.

As discussed in Sec. 3.5, our framework can also learn the invariances in an implicit, unsupervised mode, without explicitly or analytically modeling the underlying transformations, as long as the *orbits* of transformed templates can be identified and sampled. Three types for identifying (or observing) orbit samples are tested (see *Implicit* in Table 1):

- kmeans: cluster random training samples, and treat each cluster as an orbit.
- category: partition random training samples by their phonetic category, i.e., each phoneme class corresponds to one orbit.
- SA: like category but use the partition defined by the dialect sentences (SA) in TIMIT. This subset contains two sentences read by *all* the speakers from the eight dialect regions, defining a relatively clean partition that exposes speaker variability.

The dialect-based partition SA performs best out of the three, and category works slightly worse than kmeans. This is probably because orbits defined by phonetic categories encompass multiple and composite transformations, making it very difficult to approximate well in a single layer with limited number of templates and histogram bins (Sec. 3.4).

These preliminary results suggest that prior knowledge of the signal domain can help in aiming for invariance to specific, "typical" transformations. However, even without domain knowledge, invariant representations can be learned in a supervised or, most importantly, unsupervised way. The two can be potentially combined in multilayered, hierarchical architectures of low-level, part-based, analytic templates and higher-level, whole-based data-driven templates.

6. Conclusions

Based on a theory of invariant representations under compact group transformations and approximately invariant representations under general smooth transformations, we proposed invariant feature extraction modules that can be cascaded to factor out different transformations. Such modules can be constructed using templates and their transformed versions, either via explicit parametrization and sampling from the transformations or implicit (unsupervised) learning from data. The proposed modules and layers use basic computational primitives of "neurons", thus any results under this framework can form hypotheses on representation and recognition in the auditory cortex, and shed light on how humans learn to recognize speech.

7. Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. Lorenzo Rosasco acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC.

8. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, Jul. 1997.

[4] O. Scharenborg and M. P. Cooke, "Comparing human and machine recognition performance on a VCV corpus," in *ISCA Tutorial and Research Workshop (ITRW) on "Speech Analysis and Processing for Knowledge Discovery"*, Aalborg, Denmark, 2008.

[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] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, Nov. 2012.

[8] G. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4346–4349.

[9] O. Vinyals and L. Deng, "Are sparse representations rich enough for acoustic modeling?" in *Proc. INTERSPEECH 2012, 13th Annual Conf. of the ISCA*, Portland, Oregon, 2012.

[10] T. Sainath, D. Nahamoo, D. Kanevsky, B. Ramabhadran, and P. Shah, "A convex hull approach to sparse representations for exemplar-based speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2011.

[11] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, June 2010.

[12] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1635–1638.

[13] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Apr. 2007, pp. 757–760.

[14] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. INTERSPEECH 2011, 12th Annual Conference of the ISCA*, 2011, pp. 237–240.

[15] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 3377–3381.

[16] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," *CoRR*, vol. abs/1311.4158, 2013.

[17] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, no. 1, pp. 106–154, Jan. 1962.

[18] 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.

[19] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[20] A.-R. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[21] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems (NIPS) 22*, 2009, pp. 1096–1104.

[22] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4277–4280.

[23] A. Graves, A.-R. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.

[24] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2013, pp. 273–278.

[25] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 24–29.

[26] J. Andén and S. Mallat, "Deep scattering spectrum," 2013, IEEE Trans. Signal Processing (submitted). [Online]. Available: http://arxiv.org/abs/1304.6763

[27] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. INTERSPEECH 2012, 13th Annual Conf. of the ISCA*, Portland, Oregon, 2012.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS) 25*, 2012, pp. 1106–1114.

[29] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

[30] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Proc. EUROSPEECH, 8th European Conf. on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 2582–2584.

[31] T. Lee and S. Soatto, "Video-based descriptors for object recognition," *Image and Vision Computing*, vol. 29, no. 10, pp. 639–652, Sep. 2011.

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

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

[34] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 284–299, Feb. 1985.

[35] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition," *Nature Neurosience*, vol. 2, no. 11, pp. 1019–1025, Nov. 2000.

[36] 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.

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

[38] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.