INTERSPEECH 2014

Word-level Invariant Representations From Acoustic Waveforms

Stephen Voinea[1], *Chiyuan Zhang*[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
[voinea, chiyuan, gevang, lrosasco]@mit.edu, tp@ai.mit.edu

Abstract

Extracting discriminant, transformation-invariant features from raw audio signals remains a serious challenge for speech recognition. The issue of speaker variability is central to this problem, as changes in accent, dialect, gender, and age alter the sound waveform of speech units at multiple levels (phonemes, words, or phrases). Approaches for dealing with this variability have typically focused on analyzing the spectral properties of speech at the level of frames, on par with frame-level acoustic modeling usually applied to speech recognition systems. In this paper, we propose a framework for representing speech at the word level and extracting features from the acoustic, temporal domain, without the need for spectral encoding or preprocessing. Leveraging recent work on unsupervised learning of invariant sensory representations, we extract a signature for a word by first projecting its raw waveform onto a set of templates and their transformations, and then forming empirical estimates of the resulting one-dimensional distributions via histograms. The representation and relevant parameters are evaluated for word classification on a series of datasets with increasing speaker-mismatch difficulty, and the results are compared to those of an MFCC-based representation.

Index Terms: invariance, acoustic features, speech representation, word classification

1. Introduction

Humans have a truly remarkable ability to identify and recognize speech in adverse conditions and under a range of intrinsic variations. This is made all the more clear when one looks at human-machine comparisons in speech recognition [1, 2]. Human listeners not only demonstrate higher resilience, but they do so for more complex tasks (e.g., larger vocabularies) and with less supervision (e.g., exposure to mostly *unlabeled* speech during development [3]). A possible explanation for this gap is the effect of intraclass variability on the complexity of learning for the recognition task and the role of the speech representations. It can be shown that representations which are invariant to identity-preserving transformations can significantly reduce sample complexity [4], i.e., the number of training examples, for learning. This observation has made an indelible mark on the vision community, where invariance to transformations such as translation, rotation, scaling, and illumination has been leveraged to design state-of-the-art features [5, 6, 7, 8, 9].

Speech, as a sensory signal, is also characterized by natural sources of variability, ranging from extrinsic factors such as reverberation and noise, to instrinsic, speaker-specific variations like speaker identity, age, gender, accent/dialect, and physiology; perhaps more generally, there are variations which exist at an even finer, intraspeaker level, like pronunciation, enunciation, emotional state, and speaking rate [10, 11, 12]. Such differences are reflected in mismatches between expected and actual speech variations, e.g., captured in a train/test recognition set, and thus in the ability of a speech recognizer to generalize. While there are methods for directly compensating for this mismatch, such as training separate models for individual speakers or groups (males/females, adults/children), they usually rely on additional samples, i.e., increasing the set of expected variations, to adequately train with model adaptation [11].

In this paper, we propose a memory-based learning framework for extracting representations from the sound waveform, at a scale larger than conventional speech analysis windows, which are invariant to identity-preserving transformations. Our method and model is an extension of a recent theoretical framework for unsupervised learning of invariant sensory representations [4]. Speech segments are *projected* onto sets of randomly selected signal templates and their transformations; the projection outputs over each template set are *pooled* using nonlinear functions, approximating a one-dimensional empirical distribution which is invariant or quasi-invariant, while still remaining discriminative, for a sufficient number of templates, for different signal classes (Sec. 2). The resulting signature allows the training of simple classifiers that attain the same generalization error with less labeled examples.

Interestingly, the sequence of filtering (projection) and pooling operations is similar to simple-complex cell models, proposed by Hubel and Wiesel and used as models of early-processing in computational vision [13], that account for local feature response invariance. The implication of invariant maps using similar computational modules for auditory tasks, may relate to the process of sound and speech representation in the cortex, and further strengthen the analogies between the cortical visual and auditory processing [14].

The proposed method can be used to extract an invariant representation from essentially any low-level feature domain (e.g., waveforms, spectrograms, cepstral or predictive coefficients, etc.), and at any scale of the speech signal (frame, phone, word). We represent entire-word speech segments by extracting a word-level representation from the raw waveform domain. Working directly on the waveform, we circumvent the need for carefully designed, mid-level features adapted to work well for speech [15]. This can be important for performing acoustic modeling or building hierarchies on top of unprocessed low-level inputs [16, 17, 18]. Furthermore, by modeling entire words, we can better encapsulate long-term dependencies and obtain more discriminative features that can potentially be applied to acoustic modeling for word recognition [19, 20, 21].

2. Theory

This section provides a brief overview of the theory that forms the basis for our approach. The reader should refer to the rigorous exposition and supplementary information in [4] for more detail. Even though the theory takes a somewhat strong "vision" perspective to motivate its core principles, all the ideas can be applied to more general sensory signals, such as speech.

Our ultimate goal is to map a speech signal, in our case on the raw waveform domain, to a new representation which will be unaffected by identity-preserving transformations, even if these transformations greatly alter the waveform [10]. More formally, we consider signals in a Hilbert Space $\mathcal{X}$ equipped with a norm and inner product, and as a concrete example, we can take $\mathcal{X} = \mathbb{R}^d$ where d is the signal dimension. If $x \in \mathcal{X}$ is the waveform of a word, and g is a transformation which preserves the identity of that word ($gx \equiv x$), we require a map M with the property $M(gx) = M(x)$. It is also expected that if $x, y \in \mathcal{R}^d$ with $x \not\equiv y$, then $M(x) \neq M(y)$, i.e., the map will not mix word identity. In effect, M should be unique and invariant to g.

More generally, M should be invariant to a family of transformations (e.g., all time-shifts of a word). In the case that the set of transformations forms a group G (and in particular, a finite compact group), we can find such a map M by considering the *actions* of G on $\mathcal{X}$ [22]. Each g essentially permutes the elements of $\mathcal{X}$ in a way that adheres to the group operations (e.g., composition of shifts are additive). This simple abstraction provides us with a very convenient structure, the *orbit* of a signal x under G,

$$O_x = \{gx : \forall g \in G\}. \tag{1}$$

which is the set set of all realizations of x under the transformations by G. For example, the word "one" shifted at all time locations under the action of the translation group.

The orbit provides a *unique* and *invariant* (to G) representation of a signal, as the set of all orbits partitions the space $\mathcal{X}$ under the equivalence relation $x \equiv y$ if and only if $gx = y$ for some $g \in G$. Effectively, if x and y differ only by a transformation under g, which we assume preserves identity, then $O_y = O_{gx} = O_x$ (invariance), and yet if $x \not\equiv y$, then $O_x \neq O_y$ (uniqueness). In terms of our example, the set of all time-shifts of "one" is exactly the same, regardless of the initial time; however, there exists no time-shift of "one" that results in "two", and so the orbits of these words remain separate.

A representation defined by the group orbit, i.e., setting $M(x) = O_x$, discounts identity-preserving variability while maintaining interclass separability. In place of directly comparing orbits, we rely on two observations for deriving a more concrete representation. First, we note that the action of G on $\mathcal{X}$, along with the Haar measure on G, can be used to define a set of random variables $Z_x : G \to \mathbb{R}^d$, indexed by the signals $x \in \mathcal{X}$, such that $Z_x(g) = gx$, i.e., the realizations of random variable Z_x are the elements of O_x. It can be shown (see [4] for a proof) that there is an equivalence between orbits and the induced probability distributions, so that for signals x and y,

$$O_x = O_y \iff P_x = P_y, \tag{2}$$

The corresponding probability distribution to be used for the map $M(x) = P_x$, is invariant and unique, i.e., $x \equiv y \iff P_x = P_y$.

The distance in the new representation space will depend on some metric between the d-dimensional distributions P_x. Such a metric can be approximated by

$$d_K(x, y) = \frac{1}{K} \sum_{k=1}^{K} \|\mu^k(x) - \mu^k(y)\|. \tag{3}$$

through the one-dimensional distributions μ^k, of the projections of x on a finite number K of randomly chosen, unit-norm template signals. This follows from the Cramér-Wold Theorem, and applying concentration of measures [23, 4], on the approximation of an infinite sum of projections on directions of the unit sphere. Estimations of the one-dimensional empirical distribution can be computed through a set of N smooth nonlinear functions, η_n, each computing one bin of the empirical CDF:

$$\mu_n^k(x) = \frac{1}{|G|} \sum_{g \in G} \eta_n \left(\langle gx, t^k \rangle \right), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is the normalized dot-product in $\mathcal{X}$, and G is a finite transformation group. Note that we require all actions of G on x; however, we can rewrite $\langle gx, t^k \rangle$ as $\langle x, g^{-1} t^k \rangle$, which allows us to obtain the distribution (per template) by accessing the transformed templates instead. The final invariant representation is formed by repeating this computation for all K template signals and concatenating the results, giving us

$$M(x) = (\{\mu_n^1(x)\}, \ldots, \{\mu_n^K(x)\}) \in \mathbb{R}^{NK}. \tag{5}$$

3. Implementation

To compute representations which are invariant to a particular set of transformations, the theory requires access to a number of templates and their realizations under these transformations. In this paper, the samples we wish to classify originate from a different dataset than the templates. We obtain the main train/test partitions from TI-DIGITS (see Table 1) and the templates from TIMIT. This is to empirically support the claim that the templates can be chosen irrespective of the data, scale, and conditions of the learning task.

Our templates are waveform segments of support smaller than the average word length in our task. A natural choice, coinciding with the structure in speech signals, is choosing word sub-units, such as phones. This implies that the final representation is a collection of "word parts", invariant to transformations of the part-based templates.

For the templates, we chose K random phoneme instances from the phonetically segmented TIMIT dataset, downsampled to 8kHz. To each template, we applied three types of transformations: I **time shifts** (with a stride of 50 samples), J **pitch-shifts** (range of -400 to 400 cents, with a step size of 50), and M **tempo** changes (adjusted twice, at half speed and double speed) [24]. For template t^k, this yields a set of $I \times J \times M$ vectors $\{t^k_{ijm}\}$, where each t^k_{ijm} is an instance of the original template at a particular time shift, pitch, and speaking rate. This is repeated for every template, yielding K sets of vectors.

Each word sample x in the train and test set is mapped onto the representation space defined by the K random templates. First, for template t^k, we compute the normalized dot product between x and each vector in $\{t^k_{ijm}\}$. We then pool over this set of projections with a histogram, whose size N (bin count) is a free parameter. This process is repeated for each template, resulting in a set of K histograms which are concatenated to produce the final feature vector of length $N \times K$.

4. Experiments

To analyze the efficacy of our approach for transformation-invariant features, we use the extracted features in a set of restricted-vocabulary, word classification tasks, with an increasing level of of variability and difficulty. Note that since we are

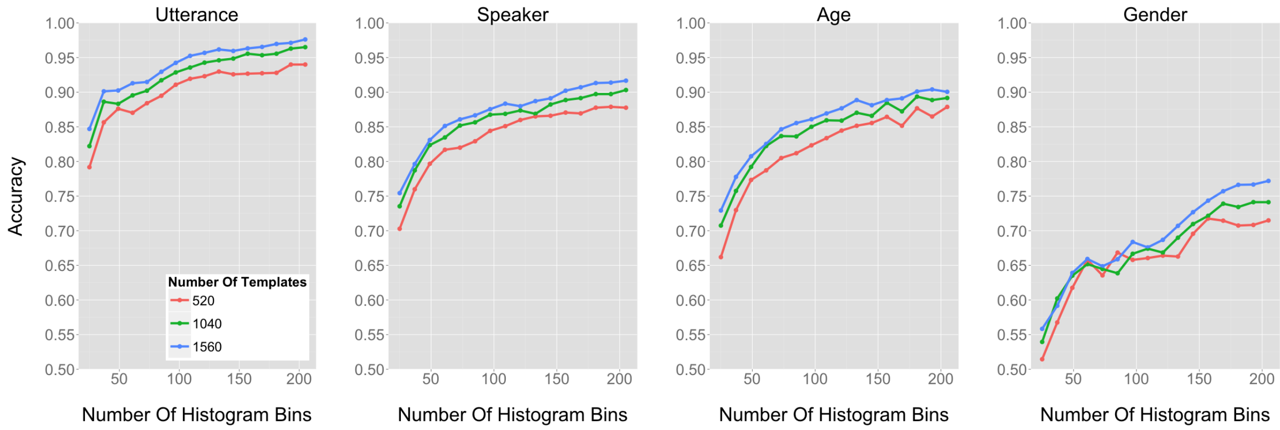

Figure 1: *Digit classification accuracy of invariant features using various numbers of templates and pooling with different numbers of histogram bins.*

extracting a signature for the entire word, we start from varying-length inputs and result in a fixed-length feature vector used to assign a label on the entire word [25]. To better understand the discriminative and invariant properties of the representation, we use a simple regularized linear classifier with leave-one-out cross-validation [26].

Aside from testing the resilience of the representation, the formed datasets provided a way to explore how the system performance depends on the representation parameters, such as the number of transformations (e.g., *simple cells*) or the number of histogram bins and templates (e.g., *complex cells*). It has been shown that the selection of even simple parameters, like the number of hidden nodes in a neural network, or the step-size between extracted features, can have as much an effect on performance as the choice of a weight-learning algorithm or the number of layers for deep networks [27].

Starting from all isolated utterances of the TI-DIGITS set (downsampled to 8kHz), we form four different partitions of train and test sets, shown in Table 1, in order to emphasize different types of train/test speaker mismatch conditions. The dataset names are meant to convey the type of mismatch.

In the first set, termed “Utterance”, the train and test sets contain different utterances of exactly the same speakers. This was used to see how our features handled intraspeaker variability, and is expected to be the easiest test of our system. The set “Speaker” has completely different speakers in the train and test sets, and anticipated to be more difficult due to the larger differences from one speaker to another. However, we note that both train and test sets have members from all main speaker groups (women, men, girls, and boys).

Next we consider age differences (“Age”) by training on adults and testing on children. Children have posed many problems for ASR, largely due to shorter vocal tract and vocal folds as compared to adults [12]. They also tend to have poorer pronunciation than the typical adult, though this may not have been an important factor here due to the small vocabulary size. Researchers have largely relied on domain-specific knowledge to account for this mismatch, as well as methods for adapting the acoustic features of children to fit models that were trained on adults [28, 29].

Finally, we consider gender variability (“Gender”) by training on men and testing on women. Again, a large part of the speech variation here can be attributed to differences in vocal tract length, so vocal tract length normalization is usually applied to compensate for this, along with gender-dependent models that are regularly used in ASR systems [12]. A breakdown of the four datasets, including sample size and group mismatch, is provided in Table 1. Classification accuracy results using the proposed features on each set, for different numbers of templates and a varying number of histogram bins, is shown in Fig. 2.

As a baseline comparison, we computed MFCC features (with Δ and $\Delta\Delta$ features) on 25ms windows (10ms offset) on each word. To obtain a fixed length representation, we averaged over thirds of the word and concatenated the results, yielding a 117 dimensional code which we refer to as MFCC3 [30]. Averaging over more parts significantly boosted performance in some cases, so we additionally include results from a 390 dimensional coding obtained from averaging over ten parts of the word (MFCC10). The MFCC-based features are used with the same linear classifiers as the proposed features (InvR), and compared on all sets to the best performing template/bin combination (InvR with 1560 templates, 201 histogram bins).

To address the large discrepancy in the dimensionality of the invariance-based features (313560) and MFCC-based coding, we also report results using a reduced dimensionality version (InvRPCA). This is designed specifically to match the dimension of the MFCC10 vector by projecting the data onto the first 390 principal components of the training set. It is worth remarking that the full feature vector (InvR) is extremely sparse (due to the large number of fixed histogram bins), with an average 1/3 of the dimensions being zeros on all samples from train and test.

5. Results

Figure 1 shows the performance on all four datasets measured by the classification accuracy on isolated digits (averaged across the 10 digit categories, chance is 0.1). Each feature set differs with respect to the number of templates (520, 1040, or 1560) and the number of histogram bins (25-205 with a step-size of 12). Each curve corresponds to the performance of a fixed number of templates, with a varying number of bins displayed on the x-axis of the plots.

As is immediately apparent in all sets, increasing either the number of templates (more simple-complex cells) or the num-

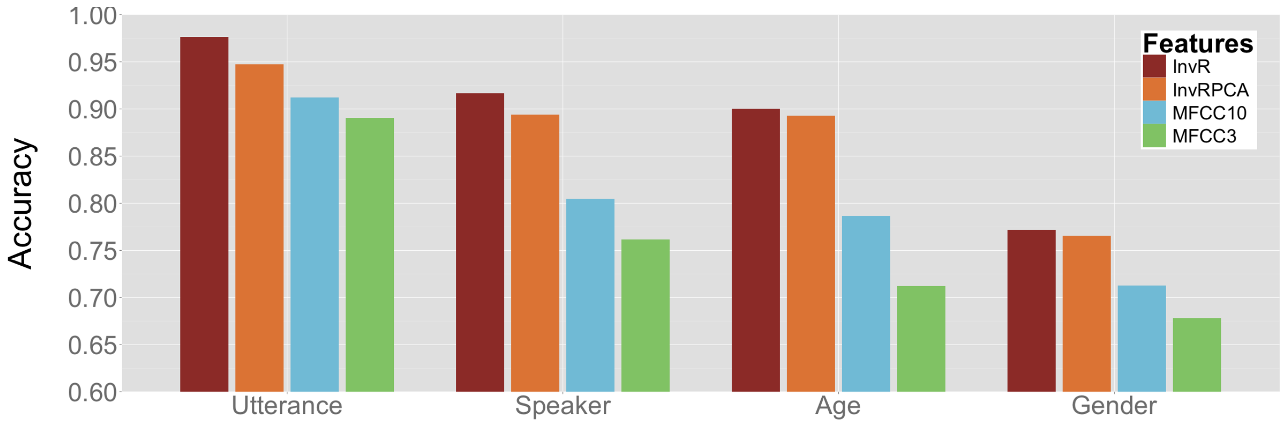

Figure 2: *Digit classification accuracy for four sets of features over all four datasets.*

Dataset		Training Set	Test Set
Utterance	#Samples	3260	3260
	Groups	All Speakers (a)	All Speakers (b)
Speaker	#Samples	3260	3260
	Groups	1/2 Speakers (a,b)	1/2 Speakers (a,b)
Age	#Samples	4500	2020
	Groups	All Adults (a,b)	All Children (a,b)
Gender	#Samples	2220	2280
	Groups	All Men (a,b)	All Women (a,b)

Table 1: *Dataset partitions from the subset of TI-DIGITS containing isolated samples of the digits 0-9, spoken by 326 different speakers, each providing two utterances of each (labelled a and b). For example: 1/2 Speakers (a,b) means the collection of both utterances (of each digit) for half of the speakers.*

ber of histogram bins via finer pooling (more complex-cells) results in improved performance. This would be the equivalent of increasing the number of nodes on hidden and output layers of multilayer neural networks. The two parameters can be increased independently, or in tandem, to yield stronger classification accuracy. In fact, it was always observed to be the case that features corresponding to both the largest number of templates and histogram bins gave the best results.

While these trends are observed across all four datasets, there is a clear difference in the overall performance among sets. As expected, the best results were achieved on the low-mismatch "Utterance" dataset, where the mismatch was limited to intraspeaker variability. A drop-off is noted in the "Speaker" dataset, most certainly due to the introduction of the interspeaker variations; however, it's interesting to note that there is virtually no drop in performance between the "Speaker" and "Age" datasets, suggesting that the extracted features are accounting for the different sources of mismatch in the two sets equally well; however, it should be noted that the ratio of train-to-test sample size is twice as large for the "Age" than the "Speaker" dataset. Finally, we observe the poorest performance on the "Gender" set, which is not surprising, considering the amount of gender-dependent modeling used in modern ASR systems.

Figure 2 shows comparison results with the MFCC-based word-features. We note that, similar to the observed trends for the proposed features, MFCCs do quite well on the "Utterance" dataset, where intraspeaker variability is the biggest source of mismatch. Most importantly though, the invariance-based features outperform MFCCs on both the "Speaker" and "Age" datasets. This is also true when the comparison is made with the InvRPCA features, which have a dimension equal to that of the MFCC10 features. Evidently, there is small difference in performance between InvR and InvRPCA representations, despite the enormous gap in the dimensionality of the feature vectors. This supports our theoretical observations that the derived features are more discriminative, even when embedded in much lower dimensional spaces, and more invariant than standard features.

6. Conclusions

We presented the design and a potential implementation of a feature map for word-level speech representations built on the level of the acoustic waveform. Based on a theory for forming invariant representations using stored templates and their transformations under compact or locally-compact groups, we proposed an implementation using smaller-support, speech part-like templates, corresponding to phonetic units from an external database. The method can transform variable-length, raw waveforms of whole-word segments into a representation which was empirically shown to be very discriminative for the purpose of word classification, and yet invariant to common sources of speaker variability such as identity and age. The invariance-based feature vectors consistently outperformed an MFCC-based representation, especially when interspeaker variability such as age is a strong source of train/test mismatch.

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] R. P. Lippman, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[2] B. Meyer, T. Wesker, T. Br, A. Mertins, and B. Kollmeier, "A human-machine comparison in speech recognition based on a

logatome corpus," in *Proc. Speech Recog. and Intrinsic Variation*, Toulouse, France, May 2006, pp. 95–100.

[3] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners." in *Proc. of EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 2582–2584.

[4] F. Anselmi, J. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations in hierarchical architectures," *CoRR*, Jan. 2014. [Online]. Available: http://arxiv.org/abs/1311.4158v5

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of International Conference on Computer Vision*, Kerkyra, Greece, Sept. 1999, pp. 1150–1157.

[6] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, Jan. 2007.

[7] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, San Diego, CA, USA, June 2005, pp. 886–893.

[9] S. Mallat, "Group invariant scattering," *Communications in Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

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

[11] D. D. O'Shaughnessy, "Acoustic analysis for automatic speech recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1038–1053, May 2013.

[12] M. Benzeghiba, R. De Mori, O. Deroo, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. J. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, Oct.-Nov. 2007.

[13] 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.

[14] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, Aug. 2001.

[15] N. Jaitly and G. E. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 5884–5887.

[16] O. Abdel-Hamid, A. rahman Mohamed, H. J. 0001, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Proc. of ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4277–4280.

[17] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[18] J. Andén and S. Mallat, "Deep scattering spectrum," *CoRR*, vol. abs/1304.6763, 2013. [Online]. Available: http://arxiv.org/abs/1304.6763

[19] C.-H. Lee, B.-H. Hwang, F. Soong, and L. Rabiner, "Word recognition using whole word and subword models," in *Proc. of ICASSP*, Glasgow, Scotland, May 1989, pp. 683–686.

[20] K. Kintzley, A. Jansen, and H. Hermansky, "Map estimation of whole-word acoustic models with dictionary priors." in *Proc. of INTERSPEECH*, Portland, Oregon, Sept. 2012.

[21] H. Hermansky and S. Sharma, "Traps – classifiers of temporal patterns," in *Procs. of ICSLP*, Sydney, Australia, Nov. 1998, pp. 1003–1006.

[22] D. S. Dummit and R. M. Foote, *Abstract algebra*. Hoboken, NJ: John Wiley & sons, 2004.

[23] H. Cramer and H. Wold, "Some theorems on distribution functions," *Journal of the London Mathematical Society*, vol. s1-11, no. 4, p. 290294, Oct. 1936.

[24] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Proc. of ICASSP*, Minneapolis, MN, USA, Apr. 1993, pp. 554–557.

[25] A. L. Maas, S. D. Miller, T. M. O'Neil, A. Y. Ng, and P. Nguyen, "Word-level acoustic modeling with convolutional vector regression," in *ICML Workshop on Representation Learning*, Edinburgh, Scotland, July 2012.

[26] A. Tacchetti, P. K. Mallapragada, M. Santoro, and L. Rosasco, "GURLS: a least squares library for supervised learning," *CoRR*, vol. abs/1303.0934, 2013. [Online]. Available: http://arxiv.org/abs/1303.0934

[27] A. Coates, H. Lee, and A. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. of the Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, USA, Apr. 2011, pp. 215–223.

[28] S. Panchapagesan and A. Alwan, "Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc," *Computer Speech & Language*, vol. 23, no. 1, pp. 42–64, Jan. 2009.

[29] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, Mar. 1999.

[30] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Proc. of EUROSPEECH*, Rhodes, Greece, Sept. 1997.