Acoustic Spatiotemporal Modeling using Deep Machine Learning for Robust Phoneme Recognition

Itamar Arel
EECS Department
University of Tennessee
Email: itamar@ieee.org

Shay Berant and Tsvi Slonim
Binatix, Inc.
Palo Alto, CA
Email: shay,tsvi@binatix.com

Ami Moyal
Afeka Academic College of Engineering
Tel-Aviv, Israel
Email:amim@afeka.ac.il

Bo Li, Khe Chai Sim
Computional Lingusitics Laboratory
National University of Singapore
Email: li-bo,simkc@comp.nus.edu.sg

***Abstract*— Robust phoneme recognition continues to play a key role in automatic speech recognition systems. A prerequisite task to phoneme recognition involves the accurate capturing of low-level acoustic features. Most preprocessing schemes consider a fixed window of time from which acoustic features are extracted. In this paper, we introduce the use of a deep machine learning architecture as a hierarchical feature extraction mechanism capable of capturing temporal dependencies that span different time scales, thereby offering improved acoustic context to the phoneme classification system. We show that using this framework a state-of-the-art frame error rate (FER) of 24.16% is achieved on the TIMIT speech corpus phoneme recognition benchmark. The principles introduced can be applied to a broad range of speech processing tasks, such as speaker identification and diarization.**

I. Introduction

Phoneme recognition continues to play a very important role in many speech processing applications. Phoneme strings form a fundamental representation for spoken language analysis and are particularly critical in key-word spotting applications. A core component of any phoneme recognition systems is the acoustic modeling employed. The latter maps speech signal information, typically obtained by applying Mel power spectral estimates in short analysis windows, to a feature space that is provided as input to inference and decoding stages. The output of the decoding stage, which is typically implemented using the Viterbi algorithm, is a sequence of inferred phonemes. Due to the desire to keep the dimensionality of the feature vector adequately small, a common approach for feature extraction involves compressing an acoustic data window of 100 to 300 msec to a vector typically 100 to 200 elements in size.

Gaussian Mixture Models (GMMs) combined with Hidden Markov Models (HMM) are a popular choice for achieving speech modeling and recognition. However, GMMs inherently treat utterances as a collection of static pieces of data thereby neglecting broad temporal information. Moreover, these models suffer from other limitations such as conditional independence of observations given state sequences, feature extraction imposed by frame-based observations, weak duration modeling and the likelihood terms being dominated by output probabilities rather than by the transition probabilities. These assertions were recently solidified through an extensive study in which it was determined that without the introduction of a novel modeling technique that will effectively represent temporal information, the performance of existing speech recognition systems is bound to remain limited. Modeling temporal information within utterances across different time scales is perceived as computationally intractable using current techniques and is thus avoided in practice.

Deep machine learning (DML) architectures have recently emerged as promising biologically-inspired frameworks for effective modeling of complex signals[1][5]. In DML, a hierarchical architecture for information representation is employed whereby higher layers of the hierarchy represent broader, more abstract characteristics pertaining to the signal modeled. However, limited application of DML to speech analytics has been reported in the literature. Published work on the topic primarily transforms the speech signal into a two-dimensional signal, resembling an image, by unfolding the signal in time, and processes the observations as if it were static images [4]. This is primarily due to the lack of a natural approach for modeling multi-scale temporal dependencies in existing DML systems.

In this paper, we report on a speaker-independent phoneme recognition system based on spatiotemporal acoustic features obtained using the Binatix DML architecture, Hierarchical Deep Recurrent Network (HDRN). HDRN comprise of a hierarchy of homogenous cortical circuits each of which is trained to capture spatiotemporal regularities in its observations, thereby serving as a robust data-driven feature extraction engine. The rich feature space produced by HDRN is applied to standard inference and decoding stages, yielding state-of-the-art performance on the standard TIMIT speech corpus for phoneme recognition.

The rest of the paper is organized as follows. In section II we briefly describe the deep machine learning system employed. Section III describes the phoneme recognition system considered, while section IV presents the experimental results. Conclusions are drawn in Section V.

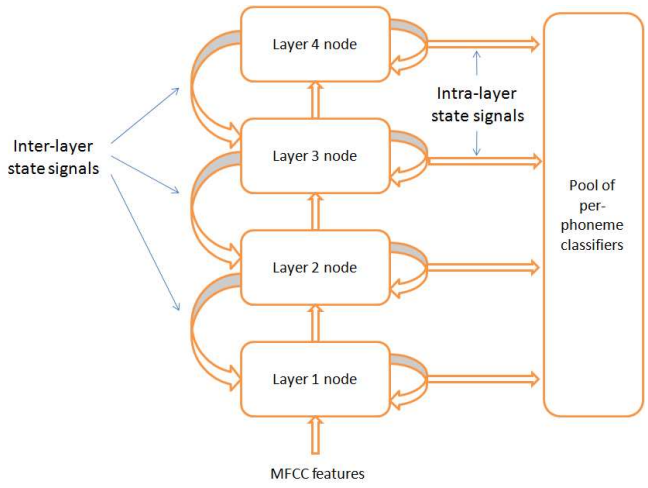

Fig. 1. Four-layer HDRN model for capturing multi-scale spatiotemporal dependencies in the speech signal. The nodes maintain state signals which serve as a feature vectors to a collection of per-phoneme classifiers.

II. Hierarchical Deep Recurrent Network

The DML architecture used here is the Binatix Hierarchical Deep Recurrent Network (HDRN) which comprises of a hierarchy of multiple layers each hosting an identical cortical circuits (or node), which is homogeneous to the entire system, as illustrated in Figure 1. Each node is implemented using a neural network and models the inputs it receives in an identical manner to all other nodes. The nodes have internal feedback signals (inter-node state signals) and signals that higher-layer nodes offer to lower layer nodes (intra-node state signals). These state signals serve as memory units and thus capture temporal dependencies that characterize the signal across multiple time scales.

A goal of the node modeling, which can be viewed as a form of lossy compression, is to represent the inputs in a compact form that captures only the dominant spatiotemporal regularities in the observations. The system is trained in an unsupervised manner by exposing the hierarchy to a large set of observations and letting the salient attributes of these observations be formed across the layers. In a typical setting, signals are then extracted from this deep-layered inference engine to a supervised classifier for the purpose of robust pattern recognition. Robustness here refers to the ability to exhibit classification invariance to a diverse range of transformations and distortions, including various noise and channel modulations.

The internal signals of the cortical circuits comprising the hierarchy may be viewed as forming a feature space, thus capturing salient characteristics of the observations. As one ascends the hierarchy broader, more abstract, features of the input data are formed, which are often most relevant for the purpose of pattern recognition. DML combined with a classifier may be viewed as a general semi-supervised phoneme recognition system, in which training is generalized as follows:

- During the first step, a set of unlabeled samples (i.e. inputs/observations that do not have a known class label associated with them) are provided as input to the HDRN. These samples arrive in the form of a sequence of Mel-frequency Cepstral coefficients (MFCC). The HDRN will learn from such samples about the general underlying structure of the speech signal it is presented with, particularly the temporal dependencies it exhibits.
- Next, each utterance is processed through the HDRN model and a feature vector is consequently produce by it. A classifier is then trained on labeled phoneme examples (i.e. inputs that have a distinct class labels associated with them) in a supervised manner.
- Testing is achieved by presenting unseen observations (e.g. novel utterances) and evaluating the output of the classifier relative to the actual phoneme class.

The above describes the steps needed in order train a single classifier and is easily enhanced to address multiple phonemes by utilizing a classifier for each phoneme..

III. System Configuration and Methodology

A. Speech Corpus and Acoustic Features

We used the TIMIT speech corpus in our experiments for evaluating the proposed system. The acoustic featured used were the standard MFCC 39-dimensional feature vectors consisting of 13 static coefficients with cepstral mean subtraction and their approximate first order and second order derivatives, extracted using the HTK package. We considered the 462 speaker training set and removed all identical sentences for all speakers in the database (SA records), as they can bias the model and classifiers. We used the HTK decoder in tandem mode whereby posteriors from the classifiers formed the inputs (i.e. feature space) for a standard GMM/HMM decoder. A set of 50 speakers was used for tuning the decoder parameters. In accordance with standard practice, results are reported using the 24-speaker core test set. We produced the training labels with a forced alignment of an HMM baseline. Since there are three HMM states per phone and 61 phones, the classifiers produced a 183-element softmax vector. We focused our attention on frame error rate, which reflects the raw recognition robustness of the system, as it is independent of any post-processing, such as that of a decoder. However, we do apply the decoder in order to also provide results that are comparable to the ones published in the literature.

Once the training labels have been created, the HMM baseline is omitted. Following the decoding phase, starting and ending silences were removed and the 61 phone classes were mapped to a set of 39 classes as in [2] for scoring. Removing starting and ending silences was performed in order to be consistent with published results. Our decoder comprised a simple bigram language model over phones. Minimal optimization of the Viterbi decoder parameters was conducted. We are confident that further tuning of the decoder would result in improved phoneme error rate results.

B. Viterbi Decoder

With the phoneme state posteriors generated from by deep models, a set of HMMs were estimated using the log of the posteriors for decoding. Due to the higher dimensionality of the posterior feature vectors, heteroscedastic linear discriminant analysis (HLDA) projections were considered in order to compress the posterior information into a lower dimension. However, the gains from the HLDA projections proved to be quite marginal and thus the full posterior vectors were adopted in our system. Both context-independent monophone HMM and context-dependent triphone HMM were experimented with. The standard Viterbi decoding was finally adopted to generate the phoneme sequence. The HTK tools were used for the HMM estimation and Viterbi decoding.

IV. Experimental Results

The features extracted from the HDRN were provided as input to a set of 183 single-layer, feedforward multi-layer perceptron neural networks. These networks were trained as binary classifiers with a mean-square error cost function, as commonly practiced. Several variants have been investigated in training the set of classifiers on the HDRN extracted features as described in the previous section. Gradient collection with standard back-propagation rule was used to batch update the weights. Moreover, a decaying step size (i.e. learning rate) over time was applied in order to speed convergence. The initial step size value was 1/32 while the terminal value was 1/256. For the purpose of regularization, additive white Gaussian noise has been applied to the inputs with an SNR of 20dB and the number of hidden neurons has been limited. In some variants, PCA was applied to the inputs as a preliminary stage to whiten the features as well as reduce dimensionality. The validation set was used to tune these parameters.

As each classifier in the set was trained to produce positive output for a specific phoneme tri-state of all 183 possible tri-states, the positive and negative groups were inherently highly unbalanced. To avoid undesirable effects, such as a low true-positive rate, a target-based emphasis training scheme has been devised, in which the gradients collected from positive targets were boosted by a fixed factor over those collected from negative targets. Further significant FER improvement has been achieved by combing scores from several sets of classifiers to produce a "committee of experts" system. The latter formed a basic boosting technique, in which the weights of the mixture components have been adjusted heuristically using the validation set. The best result achieved was an FER of 24.16% at a PER of 23.60%, composed of 15.04% replacements, 6.10% deletions and 2.46% insertions. To the best of our knowledge, this is the best reported FER result on the TIMIT benchmark [3]. With regards to the PER result, experimental outcomes have indicated that triphone HMM yielded the best performance with a relative improvement of 0.24% PER over the monophone HMM system.

V. Conclusions

This paper introduced a novel methodology for combining deep machine learning based features with a balanced classification technique to yield a state-of-the-art phoneme recognition system. The core contribution of the deep learning features involves the capturing of long-term temporal dependencies that stem from regularities in the observations. The proposed scheme is resource-efficient, in both computational and storage requirements, and naturally scales. Moreover, it can be broadened to other important speech processing tasks such as speaker identification and diarization.

References

[1] I. Arel, D. Rose, and T. Karnowski, "Deep machine learning - a new frontier in artificial intelligence research," *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13–18, 2010.

[2] K. fu Lee and H. wuen Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989.

[3] J. Keshset, D. McAllester, and T. Hazan, "Pac-bayesian approach for minimization of phoneme error rate," in *The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[4] A. rahman Mohamed, G. E. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[5] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Processing Magazine*, vol. 28, pp. 145 – 154, Jan 2011.