A Phoneme Recognition Framework based on Auditory Spectro-Temporal Receptive Fields

Samuel Thomas¹, Kailash Patil¹, Sriram Ganapathy¹,², Nima Mesgarani², Hynek Hermansky¹,²

¹Department of Electrical and Computer Engineering, ²Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, USA.
{samuel,kailash,ganapathy,nmesgar1,hynek}@jhu.edu

Abstract

We propose to incorporate features derived using spectro-temporal receptive fields (STRFs) of neurons in the auditory cortex for phoneme recognition. Each of these STRFs is tuned to different auditory frequencies, scales and modulation rates. We select different sets of STRFs which are specific for phonemes in different broad phonetic classes (BPC) of sounds. These STRFs are then used as spectro-temporal filters on spectrograms of speech to extract features for phoneme recognition. For the phoneme recognition task on the TIMIT database, the proposed features show a relative improvement of about 5% over conventional feature extraction techniques.

Index Terms: Spectro-temporal receptive fields, Phoneme recognition

1. Introduction

Important information for classification of speech sounds into subword or phonetic classes is present in the spectro-temporal dynamics of the speech signal. Conventional features for phoneme recognition capture this information separately either in short spectral analysis windows across frequency bands (e.g. PLP features [1]) or in longer temporal analysis windows in each frequency band (e.g. TRAP features [2]). In order to jointly capture these spectro-temporal dynamics, two dimensional Gabor filters [3] have also been used for extracting features. All these features are then used to train classifiers to distinguish between speech sounds.

Analysis of phoneme recognition experiments using a conventional classifier shows that about 80% of misclassifications appear within phonemes of the same broad phonetic class (BPC) [4]. This is primarily because these phonemes have similar spectro-temporal characteristics. To reduce the confusions within each phonetic class, several classifier architectures have been proposed. For example in [5], a hierarchical architecture where separate classifiers are trained for each BPC is proposed. This architecture of training separate classifiers for each BPC is useful when used along with an optimized set of features which discriminate between phoneme classes of the same BPC, rather than those between all phoneme classes.

Several approaches have been proposed to address the issue of selecting different kinds of features instead of using a single feature set to train each of the BPC classifiers. In [4], empirical results are used to choose a heterogeneous set of features including zero-crossing rate, pitch and energy, suitable for each BPC. A mutual information (MI) criteria is used in [5] to select time frequency patterns that are relevant to discriminating between phonemes within the same BPC.

In this paper we investigate the usefulness of a new set of features along with the hierarchical architecture proposed in [5]. Our features are based on recent work on the representation of speech in the primary auditory cortex [6]. In these studies, the behavior of each neuron is characterized by its spectro-temporal receptive field (STRF), which is a linear approximation of the transfer function between sound spectrogram input and neural output. These STRFs are estimated from the neural responses recorded in the primary auditory cortex. Since these STRFs exhibit distinct patterns based on stimulus frequency, scale and rate, we use them to discriminate between speech sounds. We select sets of STRFs which respond differently to phonemes within a BPCs. Features derived using these STRFs are then used to train individual classifiers to differentiate between phoneme classes of the same BPC. The outputs from each of these BPC classifiers are then combined into a single posterior probability vector similar to those derived from a single classifier trained jointly on all classes. Our experiments show that the proposed features are able to perform well compared to conventional features in phoneme recognition experiments.

The rest of the paper is organized as follows. In Section 2 we describe the STRFs that we use for our experiments and how we use them for feature extraction. Before we train classifiers for each BPC, we select STRFs that are relevant for each class. We describe a mutual information based criteria that is used for STRF selection in Section 3. Section 4 discusses how we build a hierarchical classifier for the phoneme recognition. Section 5 describes our experiments and results using these features. We conclude with a discussion of the results in Section 6.

2. Spectro-temporal representations of speech

To learn how the brain analyzes speech, representation of continuous speech in the primary auditory cortex neurons have been measured in [6]. These representations are recorded as spiking activity in isolated single neurons in the primary auditory cortex of awake, passive ferrets to phonetically transcribed continuous speech from the TIMIT database. The activity of the auditory cortical neurons are then modeled as linear STRFs, $h(t,f)$
which map the spectrogram of speech $s(t, f)$ to neural output $r(t)$:

$$r(t) = \sum_{f=1}^{F} \int h(\tau, f) s(t - \tau, f) d\tau$$

In this model, the neural output response is represented as the sum of LTI filters each operating on one center frequency. In [6] the STRF for each neuron is estimated by normalized reverse correlation of the neuron’s response to the auditory spectrogram of the stimulus.

An important observation from these studies is that neurons with different temporal modulation tuning (rate) and spectral shape tuning (scale) characteristics generate multidimensional representations in which different phoneme categories form unique patterns. These unique patterns are exhibited for different classes of vowels - open and closed and for different categories of consonants - fricatives, nasals and plosives. Further it is observed that there is always a distinct sub-population of neurons that responds well to the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space.

Some examples of population neural responses to different phoneme classes are shown in Figure 1a. Each row in Figure 1a is the average neural responses to different phonemes, sorted by best frequency, best scale and best rate of the neurons accordingly. Red regions in Figure 1a indicates the neural population that responded strongly to each phoneme. For example, phoneme /s/ activates the high-frequency neurons that match its spectral pattern. From the neural responses to vowels, it is observed that open vowels (such as /aa/ in Figure 1a) tend to evoke maximal responses in cells tuned to broad spectra (low scales at $< 1$ Cyc/Oct). An example of such broad neuron is shown in Figure 1b, neuron 2. Closed vowels on the other hand (such as /i/ in Figure 1a) evoke maximal responses in narrowly tuned cells (scales $> 1$ cyc/oct). An example of such neurons is shown in Figure 1b, neuron 4. From the neural responses to consonants sorted using best frequency, rate and scale it is observed that members of three groups of consonants: plosives, fricatives and nasals fall in distinct clusters. For example, a neuron that is tuned to rapid temporal modulations (e.g. Neuron 1 in Figure 1b) responds better to plosives (e.g. plosive /t/ in Figure 1a), while slower neurons (e.g. Neuron 3 in Figure 1b) is more responsive to fricatives (e.g., fricative /s/ in Figure 1a). Nasals on the other hand evoke strong response in narrowly tuned cells (e.g. nasal /n/ in Figure 1a).

In this work, we use 480 distinct neural representations collected in [6] for phoneme recognition. Each STRF is considered as a transformation from a two dimensional auditory spectrogram (sub-band energies stacked as a function of time) to a one dimensional neural firing pattern (energy as a function of time). We convolve the auditory spectrogram of every sentence in TIMIT with each STRFs to derive corresponding neural firing patterns (sub-band energies and rates). Once these neural firing patterns have been derived we select the responses that can discriminate between different classes of phonemes. We describe a multilayer perceptron (MLP) based feature selection technique to pick features for phoneme classes in the next section.

3. Feature selection for phoneme classes

We use a mutual information (MI) based criteria to select input features that are useful for recognizing a particular output class. We define the MI based criteria between an input feature $x$ and an output class $y$ of a classifier as

$$MI(x, y) = \sum_{t} \frac{\sum_{i} x_i \tilde{y}_i}{\sum_i x_i}$$

where

$$\tilde{y}_i = \begin{cases} +1 & \text{if class } y \text{ is correctly predicted at time } t \\ -1 & \text{if the prediction is incorrect} \end{cases}$$

The more a particular feature $x$ occurs when the label $y$ is correctly predicted, higher is the value of $MI(x, y)$. This measure hence serves as a useful indicator of which features contribute the most to the recognition of a particular label.

To select STRFs corresponding to each BPC using this framework, we first train an MLP using all the 480 STRF outputs. When trained on sufficient amounts of data, it has been
shown that MLPs estimate Bayesian a-posteriori probabilities of phoneme classes $\hat{p}(y/x)$ [7]. For each neural firing response, we find also its corresponding $\tilde{x}_t$ sequence i.e. $\tilde{x}_t$ is set to 1 if the actual response is greater than the mean of the neural firing pattern by at least twice its variance (a threshold selected empirically) and 0 otherwise. We now use the MI criteria described above with

$$\tilde{y}_t = \begin{cases} +\hat{p}(y_t/x_t) & \text{if prediction is correct} \\ -\hat{p}(y_t/x_t) & \text{if prediction is incorrect} \end{cases}$$

This modified definition of $\tilde{y}_t$ allows us to accumulate soft counts corresponding to the posteriors from the trained MLP. For each sound class, the neural firing patterns are sorted based on the MI criteria. In our experiments, we select 65 STRFs with highest MI to train classifiers for each sound class.

### 4. Hierarchical classifier for phoneme recognition

Once we select features for each BPC, we use these features to build a hierarchical classifier. A schematic representation of the classifier is shown in Figure 2. These hierarchical classifiers are built by dividing the phonemes into different categories. The first categorization is at the broad phonetic level where we build a classifier to recognize between vowels, nasals, fricatives, stops and silence. Table 1 shows the broad phonetic categorization for 61 TIMIT phonemes. This *inter-class classifier* generates probabilities of a particular frame of speech being either one of these broad phonetic classes - shown as $\{V,N,F,S,\text{Sil}\}$ in Figure 2. Using the feature selection strategy described in the previous section, we select 65 STRFs that best respond to the broad classes. After decorrelating the neural firing patterns using a DCT transform, we use a 9-frame context to derive 585 features to train the classifier.

For each broad phonetic class, we now train *intra-class classifiers*. These classifiers focus on reducing the confusions between phonemes within the same class. For each of these classifiers we select 65 STRFs that respond the best to phonemes within each class. The neural firing patterns are decorrelated and used along with a 9-frame context to train each of the classifiers. Each classifier outputs posterior probabilities of an input frame belonging to phonemes within each class. These are shown as $\{v,n,f,s\}$ in Figure 2. To derive the complete posterior vector corresponding to an input frame of...
speech, we now combine outputs from each intra-class classifier with the output from the inter-class classifier similar to [5]. Posterior vectors from each intra-class classifier are weighted by the corresponding inter-class posterior which gives a Bayesian estimate of class conditional probabilities as shown in Figure 2. This final posterior vector is now decoded to obtain the underlying phoneme sequence.

5. Experiments

For our experiments we use a phoneme recognition system based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [8]. Experiments are performed on the TIMIT database, excluding ‘sa’ dialect sentences. All speech files are sampled at 16 kHz. The training data consists of 3400 utterances from 425 speakers, cross-validation data set consists of 296 utterances from 34 speakers and the test data set consists of 1344 utterances from 168 speakers. A three layered MLP is used to estimate the phoneme posterior probabilities. The network is trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the error in the frame-based phoneme classification on the cross validation data. In our hierarchical phoneme recognition system, each MLP classifier consists of 585 input nodes and 1000 hidden neurons. The number of output nodes (with soft max nonlinearity) varies for each classifier depending on the number of output classes. For the broad class classifier there are 5 output nodes corresponding to the classes shown in Table 1. The number of output nodes for the other classifiers are as follows - 25 for the intra-vowel classifier, 8 for the intra-nasal classifier, 12 for the intra-fricative classifier and 6 for the intra-stop classifier. While each of these classifier is trained only on frames corresponding to their respective classes, the broad class classifier is trained on all frames in the training data. The final combined posterior vector corresponding to 61 labels is mapped to the standard set of 39 phonemes before decoding. In the decoding step, all phonemes are considered equally probable (no language model). The optimal phoneme insertion penalty that gives maximum phoneme accuracy on the cross-validation data is used for the test data.

Table 2 shows the results of our experiments. We first build a single classifier using conventional PLP features with a 9 frame context to discriminate between all phoneme classes. The MLP classifier consists of 1000 hidden neurons, and 61 output neurons representing the all phoneme classes. Before decoding these posterior are mapped to the the standard set of 39 phonemes. In our second experiment, we build a hierarchical classifier as described in the previous section using PLP features. Although the recognition accuracies are similar, the hierarchical classifier performs slightly worse. This probably indicates that the hierarchical classifier performs well when class specific features are used to train each classifier. In order to evaluate the performance of the STRF based features, we use two classifier architectures. In the first architecture, we use the 65 STRFs that best respond to the broad classes to train a single classifier with the same MLP architecture used to train the baseline features. The STRF outputs are decorrelated and used along with a 9 frame context. In our final set of experiments, we use the STRF based neural firing patterns to train a hierarchical classifier. The phoneme recognition accuracy using a single classifier shows improvements over conventional features. The improvements are more when these features are used along with a hierarchical classifier. The results show a significant relative improvement of about 5% using this architecture and features over the baseline systems.

6. Conclusions

In this paper, we have investigated the usefulness of neural firing patterns corresponding to actual STRFs recorded from the primary auditory cortex for the task of phoneme recognition. The primary motivation for using these features is to understand the usefulness of these unique patterns which are exhibited for different sound classes. In order to capture the effectiveness of these features, we train a hierarchical classifier built up from individual classifiers for each BPC. STRFs that are used at each level of the hierarchy are chosen using a MI based feature selection scheme. These features perform well when trained in this fashion and demonstrate how different spectro-temporal properties of speech sounds can be captured. In future, we plan to investigate the usefulness of these features in the presence of different kinds of noise.

7. References


Table 2: Phoneme Recognition Accuracies (%) for different features and classifier architectures

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Classifier Type</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP features</td>
<td>Single Classifier</td>
<td>70.2</td>
</tr>
<tr>
<td>PLP features</td>
<td>Hierarchical Classifier</td>
<td>69.7</td>
</tr>
<tr>
<td>STRF based features</td>
<td>Single Classifier</td>
<td>71.1</td>
</tr>
<tr>
<td>STRF based features</td>
<td>Hierarchical Classifier</td>
<td>71.8</td>
</tr>
</tbody>
</table>