NEURAL NETWORKS FOR VOICED/UNVOICED SPEECH CLASSIFICATION

Aage Bendiksen and Kenneth Steiglitz
Department of Computer Science
Princeton University
Princeton, NJ 08544

ABSTRACT

This paper describes the results of designing, training, and testing a neural network for the voiced/unvoiced speech classification problem. A feed-forward multilayer backpropagation network was used with 6 input, 10 internal, and 2 output nodes — for a binary decision. The six features are common and easily computed. Training was done with 72 frames from two speakers; testing was done with 479 frames from four speakers; a total of 2 errors (0.4%) occurred. Thus a small neural network performs well on the V/U problem.

1. Review of Previous Research

Voiced/Unvoiced (V/U) classifiers can be grouped into two general categories [SL77]: 1) classifiers which determine the V/U content of a segment of speech as a byproduct of an attempt to determine the primary pitch of the segment, and 2) classifiers which determine the V/U content of a segment of speech by examining one or more speech signal features which are known to be correlated with the V/U distinction. This paper describes a neural network to do V/U classification using the latter approach. Our motivation was to produce an accurate V/U classifier for high-quality analysis-synthesis using linear predictive coding.

Multi-feature techniques have some drawbacks, and it is important to be aware of the practical limitations imposed by these methods. One problem with the techniques presented below is their lack of immunity to non-stationary noise effects [AR76, CB80]. This problem is largely due to the fact that these methods do not incorporate some form of continuous adaptation to the environment being sampled—they involve training a V/U classifier in a fixed noise environment, to operate in that same environment. For a discussion of V/U classification techniques designed to operate in the presence of varying noise conditions, see [CB80, KH87, BG87, KS79]. In addition, the tradeoff between time and accuracy is important in many applications. In particular, it may not be practical to use a large number of features in real-time applications, especially if the features used are computationally complex. The ongoing development of digital signal processing hardware should cause a continuous reconsideration of such tradeoffs. In the analysis below, the techniques considered are applicable when high accuracy is desired, computation time is not a severe constraint, and noise effects can be determined in advance.

Two issues are involved in building a V/U classifier of this general class. First, it is necessary to determine a set of speech features adequate to the V/U decision task, and second, it is necessary to find an explicit rule for making the V/U decision based upon the values of the features. Although these two aspects cannot be completely isolated from one another, it is useful to view them as separate parts of the problem. (Sometimes, as in [SB80], the set of features used is tailored to the classification method.) Separation allows a more systematic approach to each part.

Among the features that have been used or proposed for V/U classification are:

1) rms or log(rms) energy of the signal [AR76, SL79a]
2) rms energy of the preemphasized signal [SL79a, SL77]
3) normalized autocorrelation coefficient at unit sample delay [AR76, SL79a]
4) normalized autocorrelation coefficient at unit sample delay, of the preemphasized signal [SL79a, SL77]
5) LPC normalized minimum error [AR76]
6) LPC normalized minimum error, of the preemphasized signal [SL79a, SL77]
7) first LPC predictor coefficient [AR76]
8) first LPC predictor coefficient, for the preemphasized signal [SL79a, SL77]
9) number of zero crossings in the signal [AR76, SL79a]
10) ratio of high frequency (above 4kHz) signal energy to low frequency (below 2kHz) signal energy [SL79a, SL77]
11) bispectrum of the signal [WB85]
12) low frequency energy [NE78, CT86]
13) an LPC distance measure [RL77]
14) an energy distance measure [RL77]
15) bit alteration rate of linear delta-modulated signal [UC80]
16) peak amplitude of the signal [SL79a, SL77]
17) ratio of preemphasized signal energy to normal signal energy [CT86]
18) causal pitch prediction gain [CT86]
19) non-causal pitch prediction gain [CT86]

Many of these features are highly correlated with one another (see [SL77, AR76]). [SL79a] described additional speech features useful for the V/U/Mixed problem.

Although a variety of methods have been used to perform V/U/M analysis based on a set of speech features, the discussion here will focus on pattern recognition techniques that have been successfully applied to the problem. These techniques have yielded classifiers that are among the most accurate available [AR76, SL77, SB80a]. Other decision methods have been less general, and therefore applicable to only the specific speech features used (as in [CT86] or [NE78]), or have been unable to achieve accuracy as high as pattern recognition techniques have achieved (as in [RV80]). Pattern recognition techniques have a distinct advantage in that they are sufficiently general to be adapted to any set of available speech features.

2. Neural Network Classifiers

The feed-forward multilayer back-propagation network is particularly well suited to the V/U/M classification problem. We use the logistic activation function

\[ o_j(x) = \frac{1}{1 + e^{-\left(w_j^T x + w_{0j}\right)}} \]

where \( o_j(x) \) is unit \( j \)'s activation value when presented the input vector \( x \), \( w_j \) is a column vector containing unit \( j \)'s input weights, and \( w_{0j} \) is unit \( j \)'s scalar bias weight (see [RM86] for more details).

To train the network, we use the following back-propagation formula to modify a unit's weights:

\[ \Delta w_{lk} = \eta \delta_j o_i, \]

where \( \Delta w_{lk} \) is the change to be made to the weight of unit \( l \)'s input from unit \( i \), \( \eta \) is the learning rate, \( \delta_j \) is an error signal available at unit \( j \), and \( o_i \) is the activation value of unit \( i \) after the input training vector has been presented to the network as input and the network has settled. The way of computing the error signal \( \delta_j \) for each unit depends on the type of unit involved. If unit \( j \) is an output unit, the error signal is computed as

\[ \delta_j = [t_j - o_j](1 - o_j), \]

where \( t_j \) is the target vector element corresponding to output unit \( j \). If unit \( j \) is a hidden unit, the error signal is computed as

\[ \delta_j = o_j(1 - o_j) \sum_k \delta_k w_{kj}, \]

where \( k \) indexes all units which have unit \( j \) as an input. \( \delta_k \) is the error signal available at unit \( k \), and \( w_{kj} \) is the weight of unit \( k \)'s input from unit \( j \). We use a feed-forward network, so all the units indexed by \( k \) are in layers succeeding the layer of unit \( j \), and error signals propagate back from the output layer.

3. Feature Selection, Network Structure and Training

Six speech features were used with ten hidden units, and two output units. Each input unit output was connected to an input of every hidden unit and to an input of every output unit. Each hidden unit output was connected to an input of every output unit. Connection weights were initially set to small real pseudo-random numbers. The number of hidden units used in the network was near the high end of what seemed reasonable for the problem; a relatively large number of hidden units was used to safeguard against training difficulties.

The output units were meant to be binary indicators of voicing, the first output indicating unvoiced excitation, and the second output indicating voiced excitation. With this convention, an output vector of (1,0) would indicate a purely unvoiced frame, and a vector of (0,1) would indicate a purely voiced frame. Thresholds of 0.1 and 0.9 were used to make binary decisions.

Speakers were recorded on standard grade audio cassette tape, using a low fidelity portable tape recorder. Speech was digitized to 16 bits per sample, at a sampling rate of 28000 Hz. Features were computed for input frames of 256 samples, corresponding to a frame duration of 9.1 milliseconds, and a frame rate of 109 frames per second. The six input features used were:

1) the rms energy of the signal:

\[ f_1 = \left( \frac{1}{N} \sum_{i=1}^{N} x_i^2 \right)^{1/2} \]

2) the rms energy of the preemphasized signal:

\[ f_2 = \left( \frac{1}{N} \sum_{i=1}^{N} p_i^2 \right)^{1/2} \]

3) the signal's normalized autocorrelation coefficient at unit sample delay:

\[ f_3 = \frac{\sum_{i=1}^{N} s_i s_{i-1}}{\left( \sum_{i=1}^{N} s_i^2 \right)^{1/2} \left( \sum_{i=1}^{N} s_i^2 \right)^{1/2}} \]

4) the preemphasized signal's normalized autocorrelation coefficient at unit sample delay:

\[ f_4 = \frac{\sum_{i=1}^{N} p_i p_{i-1}}{\left( \sum_{i=1}^{N} p_i^2 \right)^{1/2} \left( \sum_{i=1}^{N} p_i^2 \right)^{1/2}} \]

5) the ratio of signal energy above 4000Hz to signal energy below 2000Hz:

\[ f_5 = \frac{e_s(4000)}{e_s(2000)} \]

6) the product of signal energy above 4000Hz and signal
were all correctly classified to within about five percent. About a thousand training presentations, with a learning rate of 1.0 had output values of at least 0.94, and output units whose target value was 0.0 had output values of no more than 0.06. Once all of the frames were included, the network was allowed to train for an additional 25,000 iterations, in order to push the outputs even closer to the targets. Output errors were reduced to no more than 0.013 for any given output unit. The mean output error was 0.00068.

4. Testing the Classifier

A performance test was made of the classifier obtained. Test data was taken from four speakers, using the same sentence as was used in training. The frames used for testing were manually classified, in the same manner as the frames used for training. Two of the speakers, AB and LM, were the ones used to train the network. Two others, JE and VB, were speakers who were not included in the training exercise. AB and JE were male, VB and LM were female. The results of the test are summarized in table 1.

<table>
<thead>
<tr>
<th></th>
<th>JE</th>
<th>VB</th>
<th>AB</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced Frames</td>
<td>101</td>
<td>76</td>
<td>61</td>
<td>52</td>
</tr>
<tr>
<td>Unvoiced Frames</td>
<td>68</td>
<td>48</td>
<td>38</td>
<td>35</td>
</tr>
<tr>
<td>Total Frames</td>
<td>169</td>
<td>124</td>
<td>99</td>
<td>87</td>
</tr>
<tr>
<td>Errors</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Experimental Test Results

There were a total of 479 frames included in the test set. Only two classification errors were observed; in further training these two misclassified frames would be added to the training set. Both errors occurred for speaker VB, and both occurred in unvoiced frames with low energy. This represents an error rate of 0.4% over the entire set of test data, and 1.6% over the data for VB.

5. Conclusions and Future Work

A small multilayer neural network appears to work well for the V/UV problem. A network with 6 input, 10 hidden, and 2 output nodes did not encounter difficulties, such as slow learning or local minima of the error function, that could have prevented successful training. The network trained on two speakers correctly classified 186 additional frames from those same test speakers, plus all but two frames out of 293 additional frames from two other speak-
ers. The observed error rate is at least as good as other reported classifiers based on pattern recognition, and the network is simple to implement and easy to train.

In future work we will study the relative power of additional internal nodes, sensitivity to choice of feature set, and extensions to the V/UV/Mixed problem.

References


