Some vowel recognition experiments using multilayer perceptrons

Hult, G.

journal: STL-QPSR
volume: 30
number: 1
year: 1989
pages: 125-130

http://www.speech.kth.se/qpsr
SOME VOWEL RECOGNITION EXPERIMENTS USING MULTILAYER PERCEPTRONS

Gunnar Hult
Technical Dept., Swedish Telecom

Abstract

A multilayer perceptron was used to classify eleven long, stressed Swedish vowels, spoken by nine male and five female speakers in an h-vowel-t context. The perceptron was trained on speech material from eight male and four female speakers and the remaining two speakers were used as test material. Two different perceptron topologies were evaluated, one having F0 and the formants F1-F3 as input, and one using 128 LPC spectral samples as input. The vowel recognition rate for the two new speakers was 82% and 73% (F0 and formants), and 73% and 73% (LPC samples), with most errors being confusions among front vowels. Inspection of the weights of much simpler networks produced some evidence for the importance of the F2-F1 dimension.

BASIC THEORY FOR THE MULTILAYER PERCEPTRON

The multilayer perceptron (Rumelhart et al., 1986), cf. Fig. 1, is a feed-forward network consisting of nodes and weighted inter-node connections. It can be shown that four layers of nodes is sufficient for almost any classification problem. The output of each node is a non-linear function of the weighted outputs of nodes in the preceding layer, viz.

\[ o_j^P = f \left( \sum_{i} w_{ij} o_i^P + \theta_j \right) \]

Fig. 1. A multilayer perceptron with 3 layers, 10 inputs and 3 outputs.
where

\[ o_j^p = \text{the output from node } u_j \text{ when pattern } #p \text{ is applied} \]
\[ w_{ij} = \text{the weight from node } u_i \text{ to node } u_j \]
\[ \theta_j = \text{the internal threshold level in node } u_j \]

A common choice of nonlinearity is

\[ f(x) = \frac{1}{1+e^{-x}} \] (2)

The net is trained as a classifier from a list of prescribed input-output pairs. For each input-output pair, \( \{w_{ij}\} \) and \( \{\theta_j\} \) are adjusted such that the output part of the pair is produced when the input part of the pair is applied at the input nodes. The process is repeated iteratively until the net produces the desired input-output relationships for the training set.

Training typically uses a gradient descent algorithm to minimize differences between actual and desired outputs. This leads to the following equations for updating weights and thresholds:

\[ D^p w_{ij} = \alpha D^{p-1} w_{ij} + (1-\alpha) \gamma \delta_j^p o_i^p \] (3)

\[ D^p \theta_j = \beta D^{p-1} \theta_j + (1-\beta) \mu \delta_j^p \] (4)

where

\[ D^p w_{ij} = \text{the change made to the weight } w_{ij} \text{ after input/output pair } #p \text{ is applied} \]
\[ D^p \theta_j = \text{the change made to the threshold value } \theta_j \text{ after input/output pair } #p \text{ is applied} \]
\[ \alpha, \beta = \text{smoothing factors (typically in the range } [0.5, 0.9]) \]
\[ \gamma, \mu = \text{gradient step sizes (typically in the range } [0.5, 1.0]) \]
\[ t_i^p = \text{the desired output when input/output pair } #p \text{ is applied} \]

\[ \delta_j^p = \begin{cases} 
  f_j^p \left( \sum_i w_{ij} o_i^p + \theta_j \right) \cdot (t_j^p - o_j^p), & \text{if node } u_j \text{ is an output node} \\
  f_j^p \left( \sum_i w_{ij} o_i^p + \theta_j \right) \cdot \sum_k \delta_k^p w_{jk}, & \text{otherwise} 
\end{cases} \] (5)

with the summation over \( k \) in (5) being over all nodes in the layer to which \( u_j \) feeds its output, hence the name back propagation for this training algorithm.
CLASSIFICATION TASK AND SPEECH MATERIAL
The purpose of these experiments was to train a multilayer perceptron to make speaker-independent classification of the eleven long, stressed Swedish vowels

\[ \text{a}: \quad \text{o}: \quad \text{u}: \quad \text{e}: \quad \text{i}: \quad \text{y}: \quad \text{æ}: \quad \text{ø}: \quad \text{ø}: \quad \text{e}: \]

spoken in isolated words in an h-<vowel>-t context (hat, hot, ...). Nine male and five female speakers were asked to read the list of eleven words. The speech was bandlimited to 6.4 kHz and sampled at 16 kHz. The location of each vowel was determined through manual segmentation.

VOWEL RECOGNITION BASED ON F0, F1, F2 AND F3
Much has been written about the importance of formants and F0 for vowel perception, and the purpose of the first experiment was to use these parameters and compare the strategy of the trained perceptron to data and perceptual theories in the literature, e.g. (Fant, 1973) and (Traunmüller, 1984).

Formants and F0 were automatically extracted at a single location 1/3 into the vowel. Formants were extracted using peak picking and median filtering on LPC19 spectra and F0 was calculated using a time-domain, peak-picking algorithm (Gold & Rabiner, 1969). All extracted values were manually inspected and errors were corrected by simultaneous inspection of synchronized wideband spectrograms.

The chosen perceptron topology had 4 input nodes and 11 output nodes with each output node corresponding to a vowel. The number of nodes in each intermediate layer, somewhat arbitrarily set to 30, was chosen large enough to avoid convergence problems during training. The training data used 8 male and 4 female speakers and the test used the remaining 1 male and 1 female speaker.

An example of the performance of the trained network is shown in Fig. 2 where the network correctly classifies a vowel from a speaker not included in the training set.

Fig. 2. Classification of vowel /a:/ from a new male speaker. Only the input and the output layers are shown.
Training and testing results for this perceptron topology are given in Table I below.

**Vowel Recognition Based on LPC Spectral Samples**

A perceptron classifier based on formants and F0 has several disadvantages. It requires considerable hand-editing of input data since pitch and formant trackers make occasional errors. It may also exhibit convergence problems during training since we require that input patterns which are closely spaced in the four-dimensional input space (e.g., front vowels) produce widely spaced output patterns (Rumelhart et al., 1986).

A new perceptron structure was thus tested, this time using 128 spectral samples from an LPC19 analysis as input. Training time was reduced by zeroing all spectral samples below the average spectral level, thus providing a sparse input. The spectral samples were again measured 1/3 into the vowel, but this time no manual corrections were made.

The chosen perceptron topology had 128 input nodes and 11 output nodes with each output node again corresponding to a vowel. The number of nodes in the intermediate layers was, again somewhat arbitrarily, set to 40 and 20, respectively. The training data used 8 male and 4 female speakers and the test used the remaining 1 male and 1 female speaker.

Training and testing results for this perceptron topology are given in Table I below.

<table>
<thead>
<tr>
<th>Input</th>
<th>Training time</th>
<th>Errors in test type</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0-F3</td>
<td>10h 39m 57s</td>
<td>i→y: m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y→i: m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ο→ό: f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ω→e: f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ο→e: f</td>
</tr>
<tr>
<td>LPC samples</td>
<td>5m 2s</td>
<td>ω→y: m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>i→y: m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ο→ω: m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ο→ό: f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>i→ο: f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y→e: f</td>
</tr>
</tbody>
</table>

Table I. *Training and test results of two multilayer perceptrons.*

**DISCUSSION**

The errors in Table I are mainly confusions among front vowels and are due to the fact that the training material is too small to capture important speaker variability. Other errors may also be seen in Table I, e.g., the first error for the female talker in the formant-based perceptron, an ο: to ο: confusion, caused by the fact that the network almost completely ignores F0. The fact that F0 is ignored is clearly seen in Fig. 3,
which shows the weights from the input layer to the next layer. The weights applied to F0 are generally much smaller in magnitude.

![Fig. 3. The weights from the input layer to the next layer for a trained, 4 layer, formant- and F0-based perceptron classifier. Black is positive, white is negative, and the size of the square is proportional to the magnitude. F0 weights are on the top.](image)

In spite of several simulations, we have not seen any evidence for proposed perceptual theories such as Z1-Z0 (using Bark scale frequencies) as an indication of degree of openness (Traunmüller, 1981) or Z2-Z1 as a cue to degree of roundedness among front vowels. This does not in any way disprove these theories. The back propagation training procedure will provide one possible solution, most likely there is a large number of other possible solutions as well, some of which may be consistent with proposed theories. In fact, even if the perceptron does use such a proposed strategy, we may not be able to see it directly in the highly distributed weight patterns. The fact that we don't use the Bark scale here is probably irrelevant: the nonlinearities in the net can approximate logarithmic characteristics if such are desirable.

Using a 2-layer network with 4 input nodes (formants, F0), 11 output nodes and trained on a single speaker, we can force the training procedure to pay attention to F0, cf. Fig. 4, although it is hard to see a systematic behaviour. There is a tendency, however for front vowels to have opposite signs on the F1- and F2-weights which indicates that F2-F1 (or F1-F2) may be important. The 2-layer perceptron is, however, not very useful as a speaker-independent classifier since the only possible decision boundaries are hyperplanes in the input space (Rumelhart et al., 1986).

![Fig. 4. The weights from the input layer to the output layer for a trained, 2 layer, formant- and F0-based perceptron classifier. Black is positive, white is negative, and the size of the square is proportional to the magnitude. From the top F0, F1, F2, F3.](image)
References


