AUTOMATIC CLASSIFICATION AND FORMANT ANALYSIS OF FINNISH VOWELS USING NEURAL NETWORKS

Toomas Altosaar and Matti Karjalainen

Helsinki University of Technology
Acoustics and Speech Processing Laboratory
Otakaari 5A
02150 Espoo, Finland

ABSTRACT

In this paper we report results from a study of using feedforward neural networks with error back-propagation in order to see their inherent ability to learn speaker independent classification and formant analysis of Finnish vowels.

1. INTRODUCTION

The recognition and analysis of vowels is an important problem in the field of speech recognition and phonetics. Neural networks [5] are shown to give excellent performance in many speech recognition subtasks [1],[2]. They can be described as "black-boxes" that when given an input and desired output can actually learn to associate the input with the output. The performance levels achieved with neural nets can be very high and their use is an attractive method when performing vowel recognition or analysis [5].

In our study we used feedforward nets with error back-propagation. Figure 1 shows a possible net topology where data flows from the input layer to the output layer via a hidden layer. Each layer is fully connected with the next one. The dimensionality of the net can be stated as the number of nodes in each layer (10-6-2 in figure 1).

This paper describes the application of neural networks to vowel recognition and analysis. Experimental results of vowel recognition and formant analysis are presented along with a summary regarding the usefulness of neural nets in this problem domain.

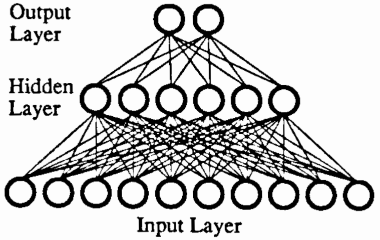

Figure 1. A possible network topology (general structure of a feed-forward net).

2. VOWEL RECOGNITION

For our vowel recognition experiments we used speech taken from 12 female and 24 male speakers. Static auditory spectra (288 in total) each consisting of a 48 point real-valued vector were used as the input representation [2]. The topmost curve in figure 2 shows the auditory spectrum of the vowel /ä/. The 0-24 Bark critical-band scale corresponds to approximately 0-15 kHz.

We defined a criterion for when a neural net had learned all of the input material: a) all of the inputs had to be correctly classified, and b) a 0.75 minimum level had to be measured for the correct output layer node. The target values during training were 0.0 or 1.0.

In the first experiment we determined how many nodes were required in the hidden layer as well as which spectral representation performed best to correctly learn 8 vowels from a single male speaker. What is meant by spectral representation is the scale or resolution of the input data. We applied a Gaussian band-pass filter to the original auditory spectra to obtain a fine-scale representation that

would emphasize formant-like local structures in the spectrum. A higher level of smoothing was also applied to yield a coarse-scale representation that emphasized more global spectral trends. The fine and coarse representations for the vowel /ä/ can also be seen in figure 2.

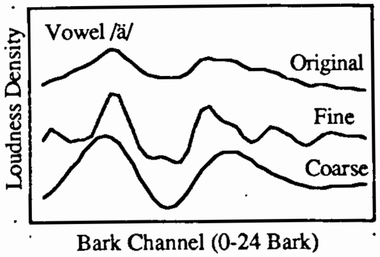

Figure 2. Original, fine, and coarse auditory spectrum representations of /ä/.

We then trained 100 separate nets with similar initial parameters of dimension 48-3-8 (48 input nodes, 3 hidden nodes, and 8 output nodes each corresponding to one of the eight Finnish vowels). We repeated this test for 4 to 9 hidden nodes, and for all three representations. The results which can be seen in figure 3 indicate that the fine spectral representation learned the 8 vowels most frequently, followed by the original and coarse representations. This result is explainable since emphasized formants help to distinguish each of the eight vowels of a single speaker.

For a larger input set (24 male speakers, 192 vowel spectra) these results changed somewhat and are shown in figure 4. Here the number of nodes was varied between 3 and 14 and only the original and fine spectral representations were compared. The ability of learning the input set perfectly when using the fine resolution was always lower than for the original representation. A possible explanation for this is that in general the fine representation will emphasize formants, and since several examples of each vowel exist in the training set with different formant frequencies, the variability of the input representation increases making it more difficult for the net to learn the differences. For this reason we decided to use only the original spectral representation in the remaining tests.

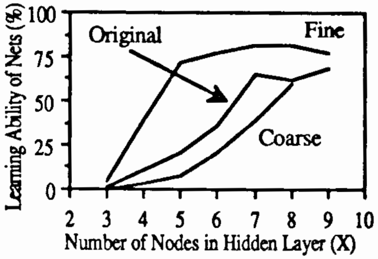

Figure 3. 48-X-8 Net's Ability to Learn 8 Vowel Spectra

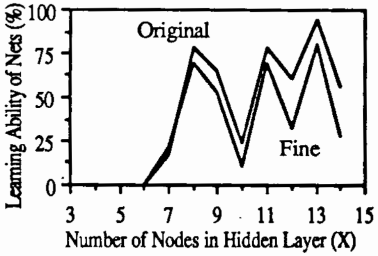

Figure 4. 48-X-8 net's ability to learn 192 male spectra.

2.1 Effect of F0 on Classification

A central part of the study was to see if the pitch frequency as additional information to the auditory spectrum could improve the classification performance of the nets by providing extra information for spectral normalization. For this test we created three training sets: male (24 speakers), female (12 speakers), and a male+female set (36 speakers), in order to see the degree of speaker independence and difficulty of the learning problem in each set.

For all three sets the number of hidden nodes was varied from 3 to 48. Figure 5 shows the learning ability for the 24 male set. Each test was repeated 100 times to gain statistical confidence. With eight hidden nodes approximately 80% of the nets were able to learn the male training set entirely. No significant difference in performance level was observed if F0 was included or not. This result is somewhat surprising because it is often assumed that human listeners do spectral normalization based on the pitch of the speaker.

For the female and male+female training sets the results were similar to the male

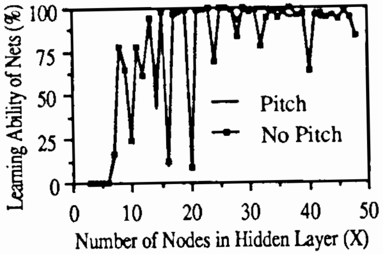

Figure 5. 24 male speakers with and without pitch information.

training set test, i.e. no significant improvement or degradation of learning frequency was found by including pitch information.

3. FORMANT ANALYSIS

The second main topic of this study was to investigate the usefulness of neural networks in analyzing continuous parameters or features of vowels. Specifically we wished to teach nets to be able to identify the location of the first two formant frequencies of vowels in the auditory spectrum. A traditional method to perform this task automatically is to calculate the envelope of the spectrum and peak-pick the formants. Another method utilizes solving for the poles by LPC.

We trained networks of dimensions 48-X-2, $X \in [2,15]$ to estimate the two first formant frequencies F1 and F2 of vowels. These estimates were based on the auditory spectrum input and we hypothesized that the network could be more robust than traditional methods to find and label the formant frequencies. The output level nodes of the net were modified by removing the sigmoid non-linearity thus allowing continuous valued output values to be realized. As a training set we selected 64 vowels and diphthongs from a single male speaker. The formant frequencies were located by hand by an experienced speech scientist.

Figure 6 shows the average F1 and F2 absolute errors as a function of the number of hidden nodes. F2 exhibits a larger error since a larger input variation exists for it but drops down to ≈0.15 Bark when the number of hidden nodes is seven or higher. This error corresponds to approximately 35 Hz at 1.5 kHz. The F1 error being considerably smaller was found to be 0.08 Bark which corresponds to 10 Hz at 400 Hz.

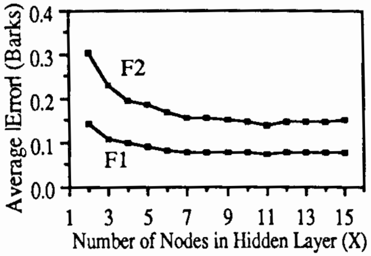

Figure 6. Average Formant Analysis Error of 64 Male Spectra as a Function of Net Size.

We evaluated the performance of the 48-12-2 net on three independent (with respect to the training set) evaluation sets: male (3 speakers), female (3 speakers), and male+female (3 male and 3 female speakers). As can be seen in figure 7 the average absolute error for F1 (labelled "F1 error") when evaluated on the male set of spectra (3M) was ≈0.5 Bark, and for F2 (labelled "F2 error") 0.8 Bark. The F2 error was very large when evaluated on the female set (3F) - 2.2 Barks which corresponds to ≈600 Hz at 1.5 kHz. Notice that the net was trained by data from a single male speaker.

To see if we could reduce the average absolute F2 error for females we trained a similar net with the original 64 vowels and diphthongs but also included eight static vowels from one female speaker. When re-evaluated on the independent sets the F2 error (labelled "F2 error 1F"), as seen in figure 7, was substantially smaller dropping to ≈1.3 Barks which corresponds to ≈330 Hz at 1.5 kHz for the female (3F) evaluation set.

The overall accuracy for the formant analysis tests was not always good but the nets showed a robust behaviour avoiding gross errors such as incorrect formant ordering, which is very difficult to achieve by traditional methods. We also observed that networks based the formant estimates on the general shape of the auditory spectrum but didn't generalize to search for exact auditory peaks. Further studies are needed to see how accurate and robust the method could be if a more complex net is used with more training material.

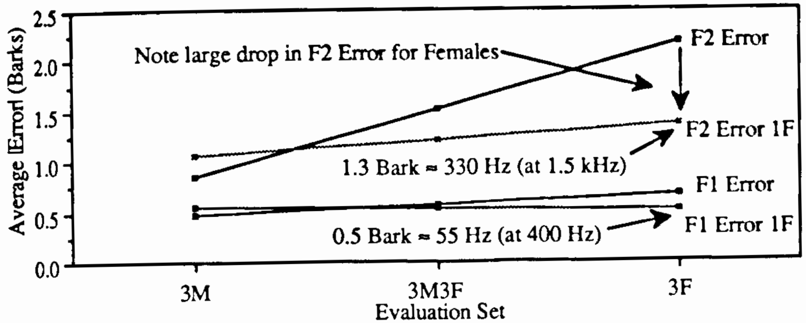

Figure 7. Evaluations of Trained Net on Independent Spectra.

4. COMPUTATIONAL ENVIRONMENT

These experiments were carried out on an object-oriented signal processing environment called QuickSig [3], developed in our laboratory. QuickSig, which is an extension to the Symbolics Common Lisp and Flavors environment runs on Symbolics Lisp Machines. To speed up the tests by a factor of 150 over the Symbolics Lisp Machines a Texas Instruments TMS320C30 digital signal processor was used.

5. SUMMARY

This study has shown that neural networks are very useful tools in the classification and analysis of vowels. The ability of a neural network to generalize is an attractive feature since this means that a trained net, even if it has never seen a certain input before, can make an intelligent decision.

Specifically we found that F0 does not help in achieving better performance levels for vowel recognition. This confirms earlier work [4]. The number of nodes in the hidden layer was found to affect the learning potential. With too many nodes the net will learn but will not generalize (it will learn each training element individually). On the otherhand, given too few nodes all the inputs will not be classified correctly. We also found that the preferred spectral representation when having to choose from a set of representations derived from the auditory spectrum was the unmodified auditory spectrum itself.

In the formant frequency analysis experiments more spectra need to be used to verify the accuracy and potential of the approach. Eventhough performance may not reach the levels of other well established methods such as LPC, neural networks may provide a useful general indication of formant locations for later, more detailed analysis, or rule-based combination of multiple methods.

6. REFERENCES

[1] LIPPMANN, R. (1987) "An Introduction to Computing with Neural Nets", In *IEEE ASSP Magazine*, 4.

[2] KARJALAINEN, M. (1987) "Auditory Models for Speech Processing", *11th ICPhS*, Tallinn.

[3] KARJALAINEN, M. et al. (1988) "QuickSig - An object-oriented signal processing environment", Proc. IEEE *Int. Conf. on Acoustics, Speech, and Signal Processing*.

[4] MUTHUSAMY, Y. et al. (1990) "Speaker-Independent Vowel Recognition: Spectrograms versus Cochleagrams", Proc. IEEE *Int. Conf. on Acoustics, Speech, and Signal Processing*.

[5] WAIBEL, A. et al. (1989) "Phoneme Recognition Using Time-Delay Neural Networks", In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3).