Speaker Identification Based on a Modified Kohonen Network

Karina Vieira, Bogdan Wilamowski, and Robert Kubichek
Department of Electrical Engineering,
University of Wyoming
Laramie, WY 82070
vieira@uwyo.edu, wilam@uwyo.edu, kubichek@uwyo.edu

Abstract

A human information processing system is composed of neurons switching at speeds about a million times slower than computer gates. Yet humans are more efficient than computers at computing complex tasks such as speech and visual interpretation. A neural network (NN) method was developed to reproduce one of the abilities and power of the human brain: speaker recognition. To realize this method, the input patterns used for this network were the reflection coefficients (k_p) of the speech signal. The speech was analyzed using an auto regressive LPC technique and the k_ps were found by applying the Levinson Recursion algorithm. Next, a simple transformation of these input patterns onto a hypersphere in augmented space was made using a Multi Layer Perceptron (MLP) neural model. The Kohonen approach is commonly used for computing a distance in multidimensional input space so that the input patterns are projected on a hypersphere with unit radius. Although this technique is very efficient for clustering of patterns, it has one significant drawback. By normalizing the input pattern, the information about its magnitude is lost. The proposed modified network has a relatively simple architecture but is shown to be very effective in performing speaker recognition.

1. Introduction

Extensive studies have been done based upon conventional techniques of speech signal processing [2]. Much research in neural network computing techniques has been devoted towards improvement of classification performance [3]. Examples include the Minimum Distance Classifier (MDC), the Radial Basis Function Networks (RBFN), the Dynamic Time Warping (DTW), and Hidden Markov Models (HMM). The RBFN is criticized for not reassembling biological neuron. The DTW is unsatisfactory in speaker recognition due to the lack of the ability to generalize and extensive computation requirements. Even though HMM based word recognition system can achieve high recognition rates, the a-priori choice of the model topology hinders the flexibility of the model by including non-explicit knowledge of the speech recognition process [4].

The speaker identification task is to classify an unlabeled voice token as belonging to one set of N reference speakers. The idea is to identify the inherent differences in the articulatory organs and the manner of speaking. The major disadvantage of speaker recognition is to compress the speech signal in a binary form of 0s and 1s, so it can be used by the computer, when a speech signal is compressed into a character representation, a great deal of redundant information is present, such as noise. To improve the performance of recognition, a modified Kohonen algorithm was implemented.

Speaker recognition systems that perform well under no-noise environments usually experience serious performance degradation when noise is present. Therefore several research projects have been devoted to make a robust speaker recognition system which performs well in noisy environments [5] [6] [7]. Although their performances were good, their real time applications are inferior to NNs which possess parallel processing structure. The proposed algorithm will be compared with the performance of the RBFN for speaker recognition.

2. Speech Data and Background

In this study a Multi Layer Perceptron (MLP) [9] network architecture and a Modified Regression Algorithm (MRA) [8] for network training were used. The data files presented to the network were made from voice recording in a robust environment (i.e., a moderate level of background noise was present) with a 16-bit digitizer. The data input was sampled at 11,025 Hz. These files were created for 15 people including males and females reciting the word "hello" 20 times. Each word, consisting of approximately 5000 to 8000 amplitude samples was processed to generate 20 reflection coefficients as already described. The data set was divided into 2 parts: a training set consisting of 15 instances of "hello" per person and a

0-7803-4122-8/97 $10.00©1997 IEEE

test set with 5 instances of "hello" per person. The training set was used to adjust the connection weights, whereas the test set was used to assess the performance. To obtain as reliable a performance as possible, the set of 15 reflection coefficient vectors were averaged prior to training.

3. Problems with the Radial Basis Function Network

The Radial Basis Function (RBF) Network consists of two layers, usually one hidden layer with special neurons. Each of these neurons responds only to the input signals close to the stored pattern. The output signal is computed using

$$o = \exp\left(-\frac{\|x_s - \bar{x}\|^2}{2\sigma^2}\right) \quad (1)$$

where x_s is the input vector and $\bar{x}$ is the stored pattern representing the center of the cluster, and σ is the radius of this cluster. The behavior of this neuron significantly differs from the biological neuron due to the fact that the excitation is not a function of the weighted sum of the input signal. If the distance between the input and stored pattern is zero the neuron responds with an output magnitude equal to one. The neurons in the RBF network have localized receptive fields because they only respond to inputs that are close to their centers.

The graphs shown above are the results from the classification of the speakers. The y-axis represents the speaker, where each letter corresponds to one speaker. The x-axis on the other hand represents how well the speakers were recognized. This means that the letters closer to x=1 are the winner for that particular speaker.
In this case a 94.67% recognition rate was obtained from the set of training patterns and an 88.06% recognition rate using the testing patterns.

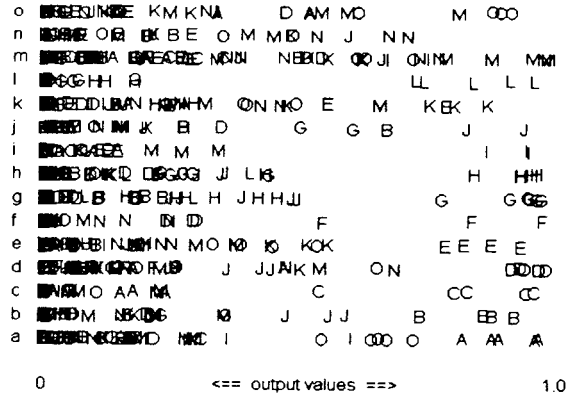
4. Auto Regressive and Levinson Recursion Techniques

The goal of determining speaker identity from recorded speech motivates using features that are sensitive to individual vocal tract (VT) characteristics. The simplest Vt approximation model consists of p rigid tubes connected in series and excited at one end by a glottal signal, E(z), to produce the output speech, S(z). The VT transfer function can be shown to be

$$H(z) = \frac{S(z)}{E(z)} = \frac{\sigma}{1 + \sum_{i=1}^{p} a_p(i) z^{-i}} \quad (2)$$

where σ is the gain, and $a_p(i)$ are the linear predictive coefficients. Related to the $a_p(i)$ coefficients are the reflection coefficients, $k_p(i)$, which indicate VT reflectivity at the i-th tube boundary. Both sets of coefficients can be estimated from the speech data using the Levinson-recursion algorithm, and both have been used successfully for speaker recognition tasks [reference = Parson's]. The relation between $a_p(i)$ and $k_p(i)$ is: $k_p(p) = a_p(p)$, and $a_p(i) = a_{p-1}(i) + k_p(i)\, a_{p-1}(p-i)$. Feature vectors in our speaker-recognition approach are based on $k_p(i)$ rather than $a_p(i)$ since averages of $a_p(i)$ vectors can lead to unstable transfer functions that lack meaningful physical interpretation.

0 to 1 scale



0.8 to 1 scale

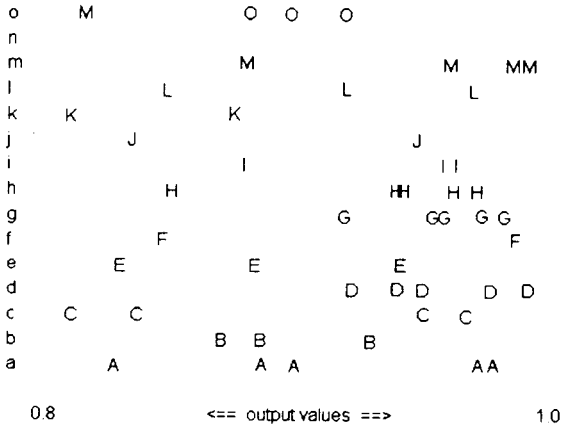


Figure 1. Results for RBF network (a) verification patterns – output scale from 0 to 1, and (b) verification patterns – expanded output scale from 0.8 to 1.

5. Proposed Algorithm

The block diagram of the network architecture is shown in Figure (2). The speech data are reduced to 20-reflection coefficient per record as described above. In the first block, the 20 inputs are transformed onto a hypersphere by increasing the problem's dimensionality. A single hyperplane is now capable of separating clusters lying on the hypersphere. Therefore a single-layer neural network can perform the classification. The output from the neural network is then applied to the Winner Take-All (WTA) architecture where the neuron with the highest output is chosen.

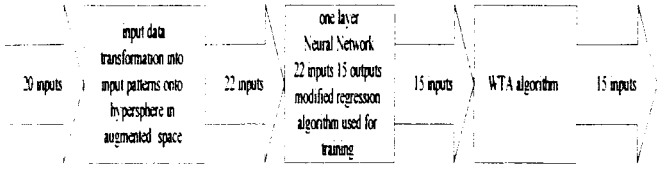


Figure 2. Neural Network Architecture.

The architecture of the NN consists of an input layer of 20 neurons (20 k_p), a hidden layer containing 15 neurons of 22 inputs including a threshold and $\sqrt{r^2 - x^t x}$ where $r^2 > x^t x$, and an output layer of 15 neurons. Each of the weights connecting the neurons were trained using the MRA (1) so that a mapping from the speech data to the output of a particular person was established.

$$w^{k+1} = w^k + \Delta w \tag{3}$$

where

$$\Delta w = \left(x^t x\right)^{-1} x^t del \quad \text{and} \quad del = \frac{D_p - O_p}{f_p'} \tag{4}$$

All neurons from the MLP network compute the weighted sum of the incoming signal $net = w^t x$, then the weighted input data was applied to sigmoidal type activation function to compute the output.

$$o = \frac{1}{1 + \exp(-anet)} \tag{5}$$

The output units were then encoded by the Winner Take All Network which allows the network to choose the neuron with highest output. The one layer neural network is easy to train. The fastest way to do this is to use the Regression algorithm which gives the same results as the Widrow-Hoff LMS algorithm [11]. However, the Regression and the LMS algorithms lead to non-optimum results. Therefore the MRA as described in [8] should be used.

6. Results

These graphs follow the same format used in Figure 1. It should be noted that the experiments were performed with very limited preprocessing of the data. Furthermore, the word "hello" is not the most appropriated to use for recognition due to the fact that identification is done by checking the inherent differences in the articulatory organs and the manner of speaking.

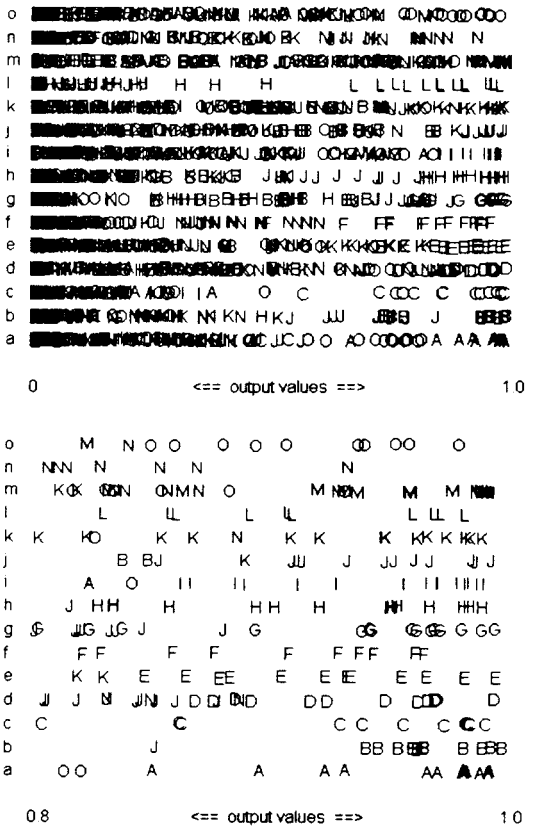


Figure 3. Results for the training patterns (a) output scale from 0 to 1, (b) expanded output scale from 0.8 to 1.

7. Conclusion

The new neural network method with a noise-robust speaker recognition system has been described. Experimental results from graphs above show that the proposed algorithm provides recognition rate of 97.78% for training data and 92.54% for testing data. On the other

hand, the RBF network provided a 94.67% of recognition for training data and 88.06% for testing data. It is very clear that the proposed algorithm generates more accurate recognition rates than the RBF network. This is primarily due to the fact that a training procedure has been used, while in the case of RBF network the mean pattern values are used as "weights". The training procedure automatically gives various scaling factors to each input and this leads to better results. Training the RBF network would also be possible, but it is more difficult than in the case of simple one layer NN, as in the proposed approach. However, improvements can be achieved by training the speech signal with White Gaussian Noise. Another approach would be to use the some kind of spectral information uniquely identifying the person independent of the text token. This kind of "voice print" would be more user-friendly than systems using code-words. Results show that this proposed algorithm is a promising architecture for speaker identification systems.

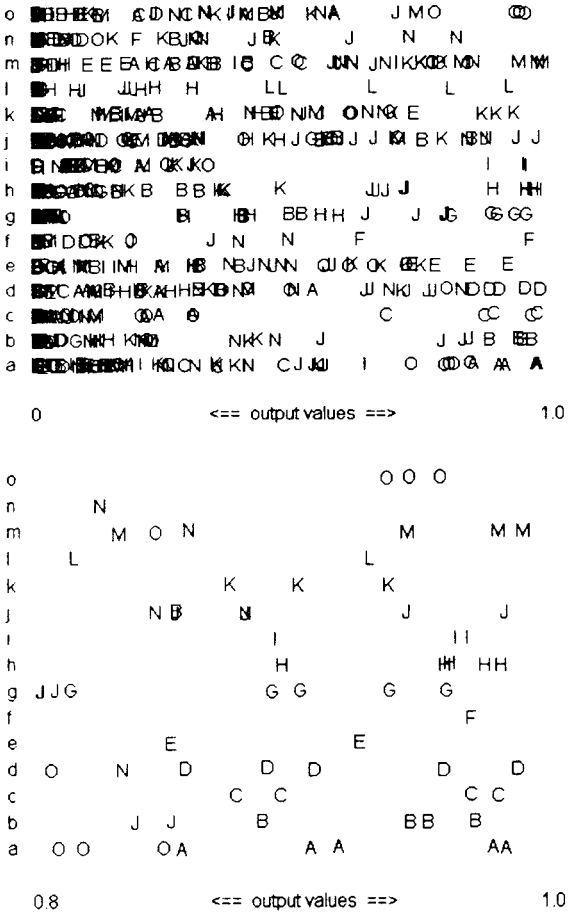


Figure 4. Results for the verification patterns (a) output scale from 0 to 1, (b) expanded output scale from 0.8 to 1.

References

[1] Wolds, A. and Wroldsen, J., "Person Identification By Neural Networks And Speech Processing," *World Congress on Neural Networks*, vol. **4**, pp. 562-567, June, 1994.

[2] Matsui, T. and Furui, S., "Speaker Recognition Technology," NTT Review, vol. **7**, No. 2, pp. 40-48, 1995.

[3] Bennani, Y. and Galliwari, P., "Connection Approaches For Automatic Speaker Recognition," *Proc. ESCA Workshop on Automatic Speaker Recognition*, pp. 95-102, 1994.

[4] Rabiner, L., Levinson, S., and Sondhi, M., "On The Use Of Hidden Markov Models For Speaker Independent Recognition Of Isolated Words From A Medium Size Vocabulary," *AT&T Bell Laboratories Technical Journal*, vol. **63**, No. 4, April, 1984.

[5] Mansour, D. and Juang, B., "The Short-Time Modified Coherence Representation In Noisy Speech Recognition," *IEEE-ASSP*, December, 1989.

[6] Wang, M. and Young, S., "Speech Recognition Using Hidden Markov Model Decomposition And A General Background Speech Model," *IEEE-ASSP*, 1992.

[7] "Signal Processing," IEEE-ISSN, vol. **13**, No. 5, pp. 58-70, September, 1996.

[8] Andersen, T. J. and Wilamowski, B. M., "A Modified Regression Algorithm For Fast One Layer Neural Network Training," World Congress on Neural Networks, vol. **1**, pp. 687-690, July, 1995.

[9] Wilamowski, B. M. and Jaeger, R. C., "Implementation Of RBF Type Networks By MLP Networks," ICNN'96 vol. **3**, pp. 1670 -1675, 1996.

[10] Chen, K., Xie, D., and Chi, H., "Speaker Identification Based On Input/Output HMMs," World Congress on Neural Networks, pp. 37-40, September, 1996.

[11] Zurada, J. M., "Introduction To Artificial Neural Systems," West Publishing Company. 1992.

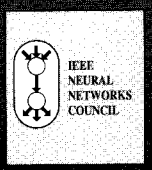
[12] Parsons, T., "Voice and Speech Processing," McGraw Hill, 1986.

1997 International Conference on Neural Networks (ICNN '97)

Westin Galleria Hotel, Houston, Texas, USA

TUTORIALS: June 8, 1997

CONFERENCE: June 9-12, 1997



Proceedings

Volume 4 of 4