Speaker Verification for Security Systems Using Artificial Neural Networks

Karina Vieira, Bogdan Wilamowski, and Robert Kubichek
Department of Electrical Engineering, University of Wyoming
Laramie, WY 82071

Abstract -- **This paper investigates automatic speaker recognition systems, which can be used for security purposes. The speech signal is compressed using linear prediction analysis and recognized by neural networks. This neural network technique is presented for the task of speech recognition and speaker verification. This technique first uses pattern recognition to identify the speech, then it is used to distinguish each user from all other speakers (impostors). With this method, unknown speech can be accurately classified as user or impostor speech. The approach used is based on the following steps: extraction of spectral features; training of an initial neural network to identify the speech; extraction of LPC-reflection coefficients for each user, training of a secondary neural-network to identify the user; and classification of unknown speech as either user or impostor.**

I. INTRODUCTION

Neural Networks NN have been used for speech recognition for many years. Early experiments were devoted to isolated word recognition on small vocabulary and extensive comparisons with classical systems have been performed [1], [2]. More recently hybrid systems combining NN and dynamic speech alignment techniques have been developed either for isolated word recognition [3], [4], [5], [6], [7] or for continuous speech recognition [8], [9], [10].

While the area of speech recognition is concerned with extracting the linguistic message underlying a spoken utterance, speaker recognition is concerned with extracting the identity of the person speaking the utterance. Applications of speaker recognition are wide ranging, including facility or computer access control [11], telephone voice authentication for long-distance calling or banking access [12], intelligent answering machines with personalized caller greetings [13], and automatic speaker labeling of recorded meetings for speaker-dependent audio indexing (speech-skimming) [14]. The proposed system can make tree types of decisions: speech recognition (identifying the message), speaker identification (identifying 1 out of N talkers), speaker verification (accept or reject a claimed identity) for continuous speech recognition.

In a previous paper [15], we presented a robust speaker identification system based on a modified Kohonen algorithm. This algorithm has several advantages for classification tasks, including better performance and reduced training time. The accuracy of the " training network for a single word "hello was 98-99%,. This is not sufficient for security lock which typically require error rates of 10^{-3} or less. In order to solve this problem, we have used a continuous-speech text-dependent system. In a text-dependent system, the speech used to train and test the system is constrained to be the same word or phrase.

In this paper the following approach is used. First, the system recognizes if the correct password was spoken. This is accomplished using RMS instantaneous power and zero-crossing frequencies. Once the correct password is identified, the system recognizes the speaker using reflection coefficients.

II. SPEECH DATA

The data files presented to the network were made by recording the voice in a robust environment (i.e., a moderate level of background noise was present) with a 16-bit digitizer sampling at 11,025 Hz.

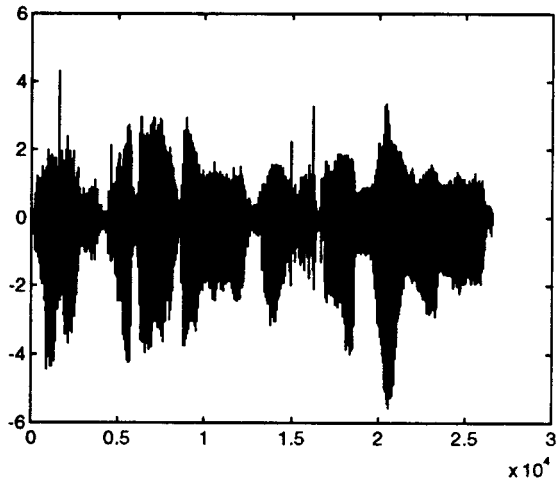


Fig. 1. Raw data of the sentence "I am the student of the University of Wyoming"

Figure 1 shows the raw recording data for the sentence: "I am the student of the University of Wyoming". The power spectrum as a function of time and frequency is shown in Fig. 2. In both cases the size of the speech data is enormously large. The data size can be reduced by using only the RMS values of the speech, as shown in Fig. 3. Another way of speech data reduction is to display only average zero-crossing rate as function of time shown in Fig. 4. The waveform of Fig. 4 can be further processed using the FFT. Magnitudes of Fourier coefficients instead of the waveform are shown in Fig. 5.

0-7803-3932-0

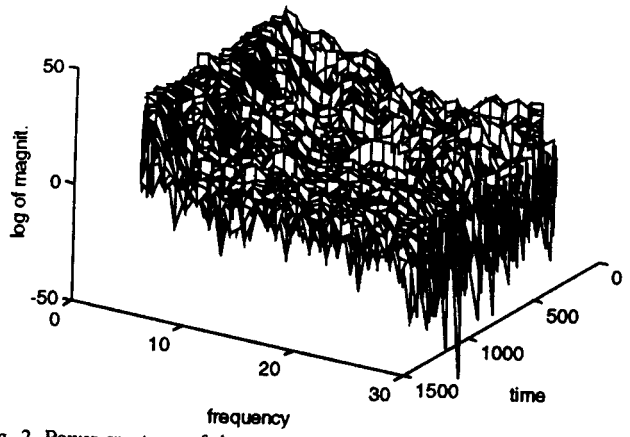


Fig. 2. Power spectrum of the sentence "I am the student of the University of Wyoming"

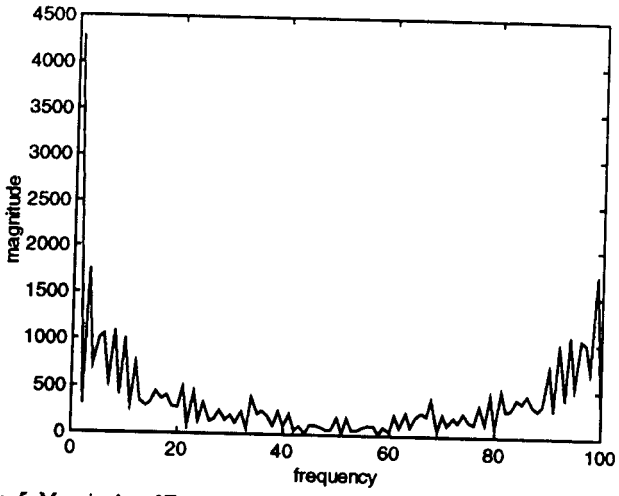


Fig. 5. Magnitudes of Fourier coefficients of data in of Fig. 4.

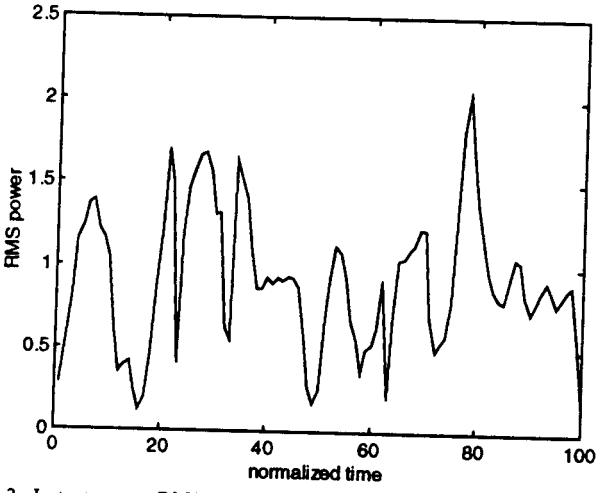


Fig. 3. Instantaneous RMS power of the sentence "I am the student of the University of Wyoming"

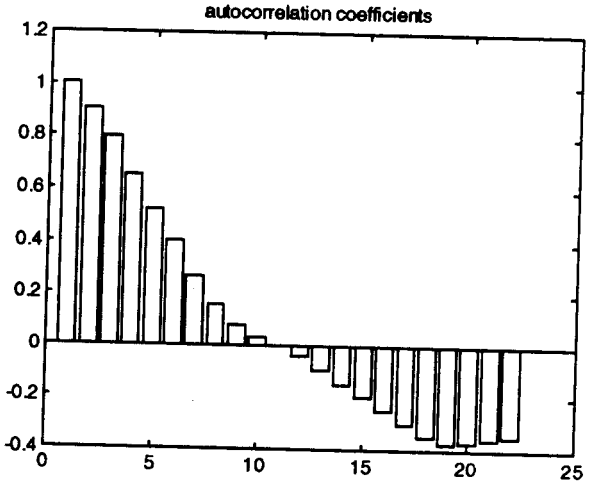


Fig. 6. Autocorrelation coefficients of the sentence "I am the student of the University of Wyoming"

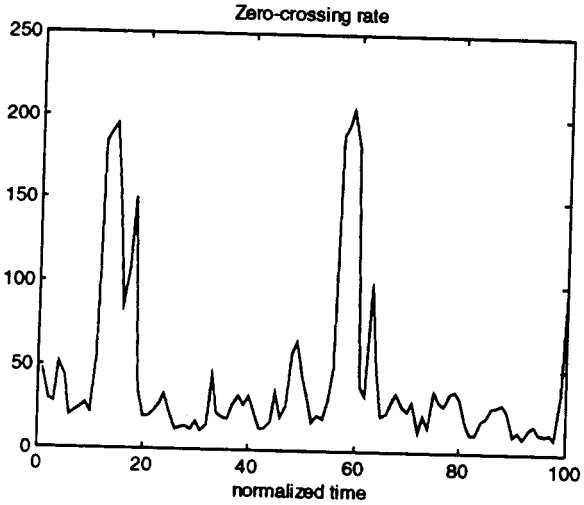


Fig. 4. Zero-crossing frequency of the sentence "I am the student of the University of Wyoming"

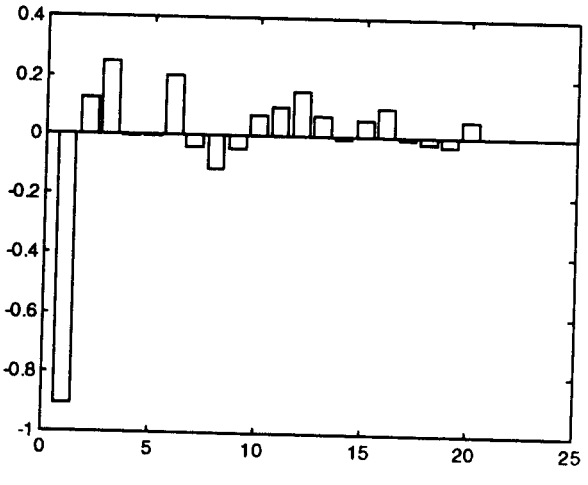


Fig. 7. Reflection coefficients of the sentence "I am the student of the University of Wyoming"

Other possible approaches are to use the autocorrelation coefficients (Fig. 6) or reflection coefficients (Fig. 7). Various data representations have their advantages and disadvantages. For example, instantaneous RMS power (Fig. 3) and zero-crossing frequencies (Fig. 4) depend more on the content of the speech, while the reflection coefficients in Fig. 7 are more sensitive to the speaker. When speech is used for identification of an authorized person, both the speech content and the speaker must be recognized.

III. RECOGNITION OF THE SPEECH CONTENT

Input data to any recognizer consists of a mix of relevant and irrelevant information. Feature selection is the process of jettisoning as much irrelevant information as possible and representing relevant data in compact and meaningful form. We have observed that any utterance carries at least two types of information: the message itself and the identity of the talker. In a speech-recognition system we wish to select the first type of feature and ignore the second; in a speaker-recognition system we wish to do just the opposite.

Figures 8 through 11 show various representation of the 20 samples of the word "February". Figures 12 through 15 show the similar representations for the word "March". Note that are instantaneous RMS power and zero-crossing frequencies appear to be most sensitive to the content.

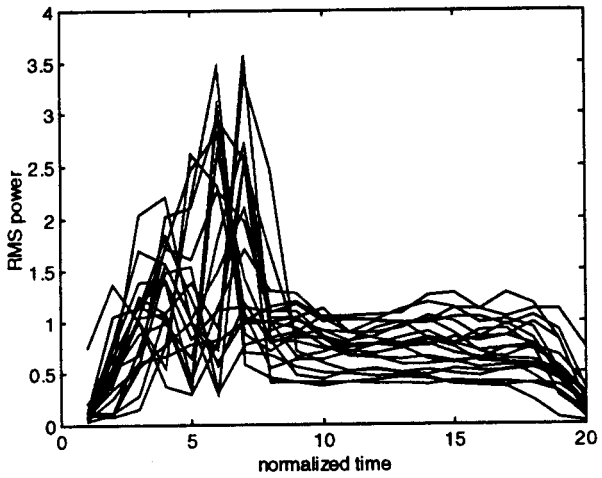


Fig. 8. Instantaneous RMS power of the 20 samples of the word "February" from the same talker.

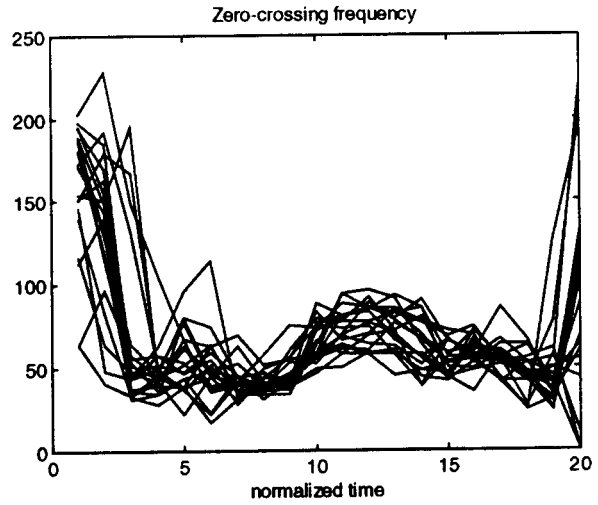


Fig. 9. Zero-crossing frequencies of the 20 samples of the word "February"

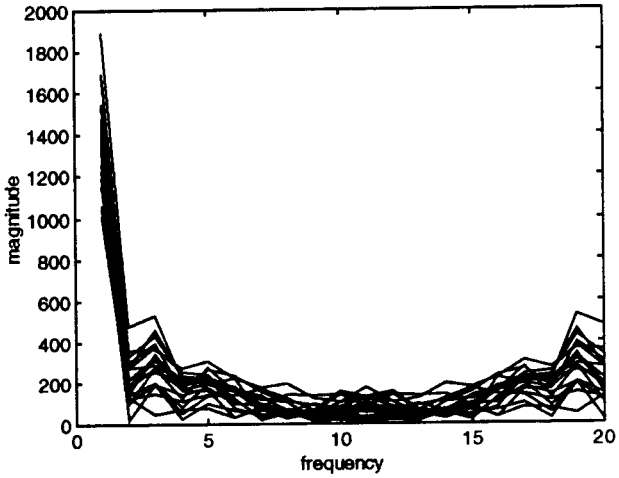


Fig. 10. Fourier coefficients for data shown in Fig. 9

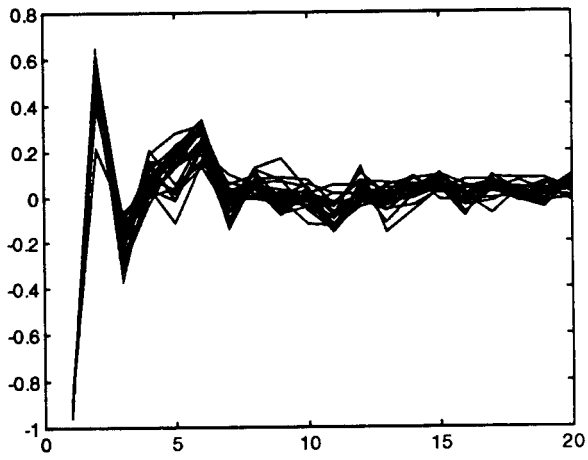


Fig. 11. Reflection coefficients of the 20 samples of the word "February"

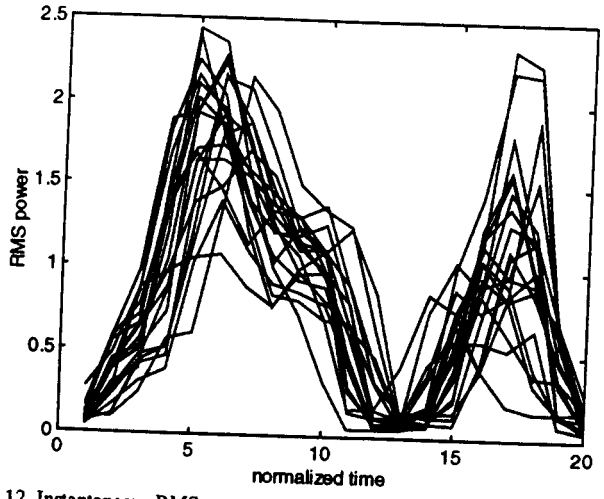


Fig. 12. Instantaneous RMS power of the 20 samples of the word "March"

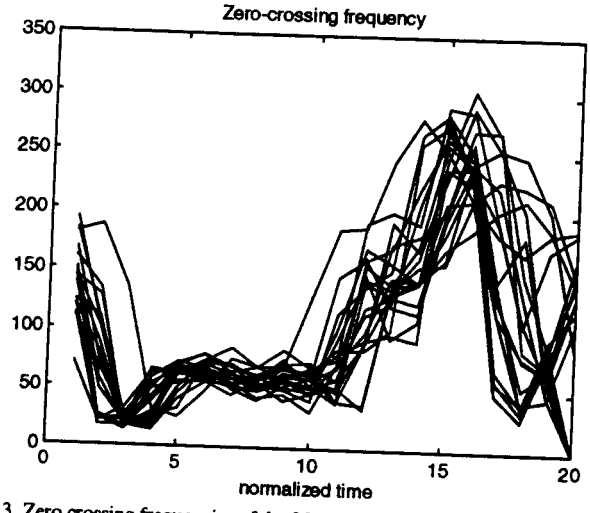


Fig. 13. Zero crossing frequencies of the 20 samples of the word "March"

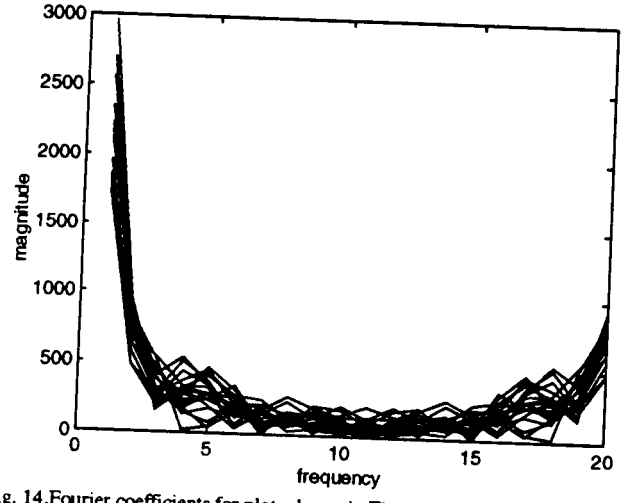


Fig. 14.Fourier coefficients for plots shown in Fig. 13

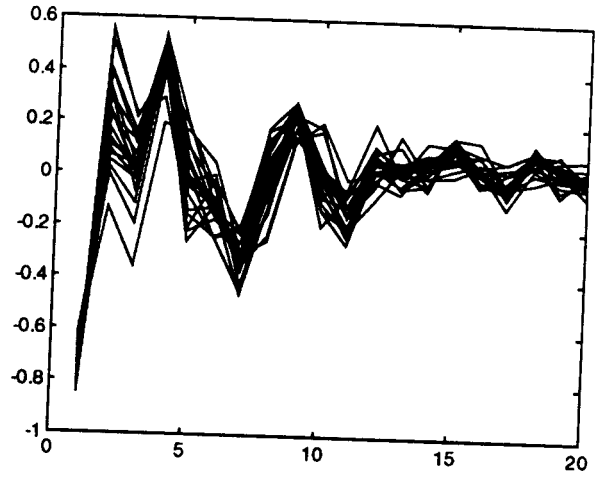


Fig. 15. Reflection coefficients of the 20 samples of the word "March"

IV. LINEAR PREDICTION AND AUTOCORRELATION METHOD

Fant developed a linear model of speech production in the late 1950's [20] where the glottal pulse, vocal tract, and radiation are individually modeled as linear filters. The goal of determining speaker identity from recorded speech motivates using features that are sensitive to individual vocal tract (VT) characteristics. The vocal tract is usually modeled as a concatenation of non-uniform lossless tubes of varying cross-sectional area that begins at the vocal cords and ends at the lips [19]. 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 source is either a quasi-periodic impulse sequence for the voiced sounds or a random noise sequence for unvoiced sounds with a gain factor G set to control the intensity of the excitation.

In order to perform the LP analysis of a speech segment consisting of N samples, the following p order all-pole filter H(z) is assumed:

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

The gain G is usually ignored to allow the parameterization to be independent of the signal intensity. With this transfer function, the difference equation for synthesizing the speech samples s(n) is obtained as

$$s(n)=-\sum_{i=1}^{p}a_i s(n-i)+Gu(n)$$

It can be noted that s(n) is predicted as a linear combination of the previous p samples. Therefore, the speech production model is often called the linear prediction (LP) model, or the autoregressive model 25]. Autoregressive analysis (AR), is a set of techniques that assume that the signal

spectrum can be represented by the all-pole transfer function.

The linear prediction (LP) analysis technique provides a procedure for separating the excitation-source and the linear-system components of the speech production model. However, to achieve this, it assumes an all-pole model for the linear system. The all-pole filter coefficients also called linear prediction coefficients are computed from the speech signal on the basis of a least-squares fit between the observed-signal values and the values linearly predicted from the preceding samples. The autocorrelation method and the covariance method are two standard methods of solving for the predictor coefficients [18][22]. When the speech signal is applied to this filter as its input, its outputs the LP error signal whose n-th order sample is given by:

$$e(n) = s(n) + \sum_{i=1}^{p} a_p(i) s(n-i)$$

In the LP analysis, the LP coefficients are determined by minimizing the total-squared value of the estimation error,

$$E = \sum_{n=n_1}^{n_2} e_n^e$$

Where the summation range depends on which of the two methods, the autocorrelation or the covariance methods, is used for LP analysis. In the autocorrelation method, the summation ranges from $-\infty$ to $+\infty$ which means that the speech signal is available for all time. This can be achieved by windowing the speech signal and assuming the samples outside this window to be zero. The methods differ with respect to the details of implementation. The autocorrelation method is computationally simpler than the covariance approach and, unlike its covariance counterpart, assures that all the poles of H(z) lie within the unit circle. Thus, the autocorrelation method guarantees the stability of the estimated all-pole filter.

Since speech is time-varying in that the vocal tract configuration changes over time, an accurate set of predictor coefficients is adaptively determined over short intervals (typically 10 ms to 30 ms) called frames, during which time-invariance is assumed. 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 calculated from the speech data using the Levinson-recursion algorithm [24], and both have been used successfully for speaker recognition tasks [23]. The relation between $a_p(i)$ and $k_p(i)$ is:

$k_p(p) = a_p(p)$, and $a_p(i) = a_{p-1}(i) + k_p(i)\, a_{p-1}(p-i)$

V. NEURAL NETWORK ARCHITECTURE

In this study a Multi Layer Perceptron (MLP) [16] network architecture and a Modified Regression Algorithm (MRA) [17] for network training were used for both the content identification and for speaker recognition. Three neural networks having the same architecture are used. The first neural network identifies correctness of the content using instantaneous RMS power. The second neural network identifies correctness of the content using zero-crossing frequencies. Once both neural networks recognize the correct password, the third neural network is used to identify a speaker. The neural network architecture, as shown in Fig. 16, has 15 inputs, and number of hidden neurons is equal to number of authorized speech patterns. The output neuron performs the OR operation. The input layer the augmented by two inputs. One additional input equal to +1 is used for biasing, and the second additional input is computed using formula:

$$x_{17} = \sqrt{r^2 - x_1^2 - x_2^2 - x_3^2 \cdots - x_{15}^2}$$

This way all input patterns are transformed onto hypersphere with the radius r so they can be clustered much easier.

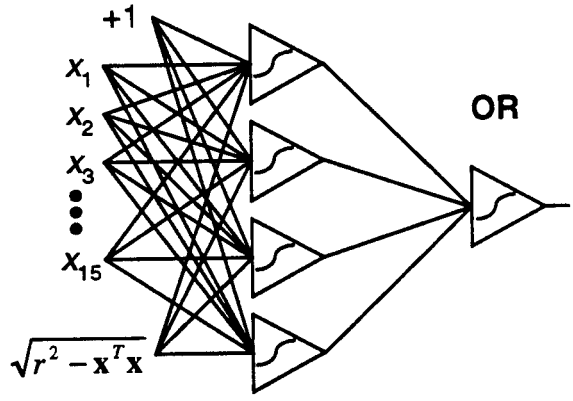


Fig. 16. Neural network architecture

The speech sample is first normalized so the average magnitude is zero and the standard deviation is one, then the sample is normalized with time. In the case of instantaneous RMS power, the normalized speech sample is divided into divided into 20 equal intervals and average power of each interval is used an input to the neural network.

In the case of zero crossing frequencies the speech sample is also divided into 20 equal intervals and zero crossings are counted for each interval. In the case of the reflection coefficients 20 reflection coefficients are calculated using Levinson algorithm [24]. For each speech pattern 20 recording are made.

The basic problem of speaker verification is to decide whether or not the individual whose identity was claimed spoke an unknown speech sample. The problem is similar to that of speech recognition in which the problem is to normalize out, in some sense, the individual speaker and extract the message content of the speech. Here, the problem is to normalize out, in some sense, the message content and extract information about the individual speaker. Because of

the similarities of these two problems, the processing for speaker verification is similar to that of speech recognition.

A speaker verification system can make two types of errors: it can reject a true customer (Type I error) or it can accept an imposter (Type II error). The goal of most verification systems is to try to bound Type I errors while minimizing Type II errors.

VI. CONCLUSION

The basic techniques used in [15] for training the were implemented in this system, however, the following improvements were introduced:

1. Several words was used instead of a single word so chances of misclassification of an intruder as the authorized person was significantly reduced.
2. Synchronization of the speech signal was introduced. True comparisons of the speech samples can be done, only when the starting times are synchronized. This way segments of the speech signal will always fall in the same portions of the sentences when reflection coefficients are calculated.
3. Password system was used. For each person a different sentence was used to improve the security level.
4. During the training procedure the distant samples were rejected so the radiuses of clusters can be kept smaller.

The neural network method with a noise-robust speaker verification system has been described. It is very clear from the graphs above that the algorithm generates more accurate results since more information is provided. In previous studies [15], we were able to obtain 98-99.9% accuracy by only using speaker information since the word spoken was the same for all the speakers in the set. In this case, information about the sentence spoken is also provided, thus providing higher level of security for the system. It is important to note that even though the proposed algorithm was able to generate 100% accuracy for the sample data used, the results were just an estimation since it is difficult to verify using actual number of impostors.

REFERENCES

[1] Waibel, Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., "Phoneme recognition using time-delay neural networks," *Technical Report TR-I-0006*, ATR Institute, Japan, 1987.

[2] Iso, and Watanabe, T., "Large vocabulary speech recognition using neural prediction model," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 57-60, Toronto, Canada, 1991.

[3] Driancourt, Bottou, L., and Gallinari, P., "Multi layer perceptron, learning vector quantization and dynamic programming: Comparison and cooperation Training," *IEEE-IJCNN*, 1991.

[4] Devillers, and Dugast, C., "Incorporating acoustic phonetic knowledge in hybrid TDNN/Viterbi framework," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 421-424, San Francisco, USA, 1992.

[5] Driancourt, and Gallinari, P., "A speech recognizer optimaly combining learning vector quantization, dynamic programming and multi-layer perceptron," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, San Francisco, USA, 1992.

[6] Haffner, and Waibel, A., "Multi-state time delay neural networks for continuous speech recognition," NIPS, pp.135-142.

[7] Bengio, Mori, R., Flammia, G., and Kompe, R., "Global optimization of a neural network hidden Markov model hybrid," *IEEE Trans. Neural Networks*, Vol. **3**, No. 2, pp. 252-259.

[8] Bengio, Mori, R., Gori, M., "Learning the dynamic nature of speech with back propagation for sequences," *Pattern Recognition Letters*, Vol. **13**, pp. 375-385.

[9] Bourlard, Morgan, N., Wooters, C., and Renals, S., "CDNN a context dependent neural network for continuous speech recognition," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, San Francisco, USA, pp. 349-352, 1992.

[10] Mellouk, and Gallinari, P. "A discriminative neural prediction system for speech recognition," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, Minneapolis, USA, 1993.

[11] Naik, and Doddington, G., "Evaluation of a high performance speaker verification system for access control," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 2392-2395, 1987.

[12] Naik, Netsch, L., and Doddington, G., "Speaker verification over long distance telephone lines," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 524-527, 1989.

[13] Schmandt, and Arons, B., "A conversational telephone messaging system," *IEEE Trans. Consumer Electronics*, vol. **30**, pp. Xxi-xxiv, 1984.

[14] Wilcox, Chen, F., Kimber D., and Balasubramanian V., "Segmentation of speech using speaker identification," *Proc. Internat. Conf. Acoust. Speech Signal Process*, pp. I-161-I-164, 1994.

[15] Vieira, K., Wilamowski, B., and Kubichek, R., "Speaker Identification Based on a Modified Kohonen Network," Accepted for *ICNN'97*, Houston, June 9, 1997.

[16] 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.

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

[18] Markel, J. D. and Gray Jr., A. H., "Linear Prediction of Speech," Springer-Verlag, Berlin Heidelberg, New York, 1976.

[19] Rabiner, L. R. and Schafer, R. W., "Digital Processing of Speech Signal," Prentice Hall, Englewood Cliffs, NJ, 1978.

[20] Fant, G., "Acoustic Theory of Speech Production," Mouton and CO., Gravenhage, The Netherland, 1960.

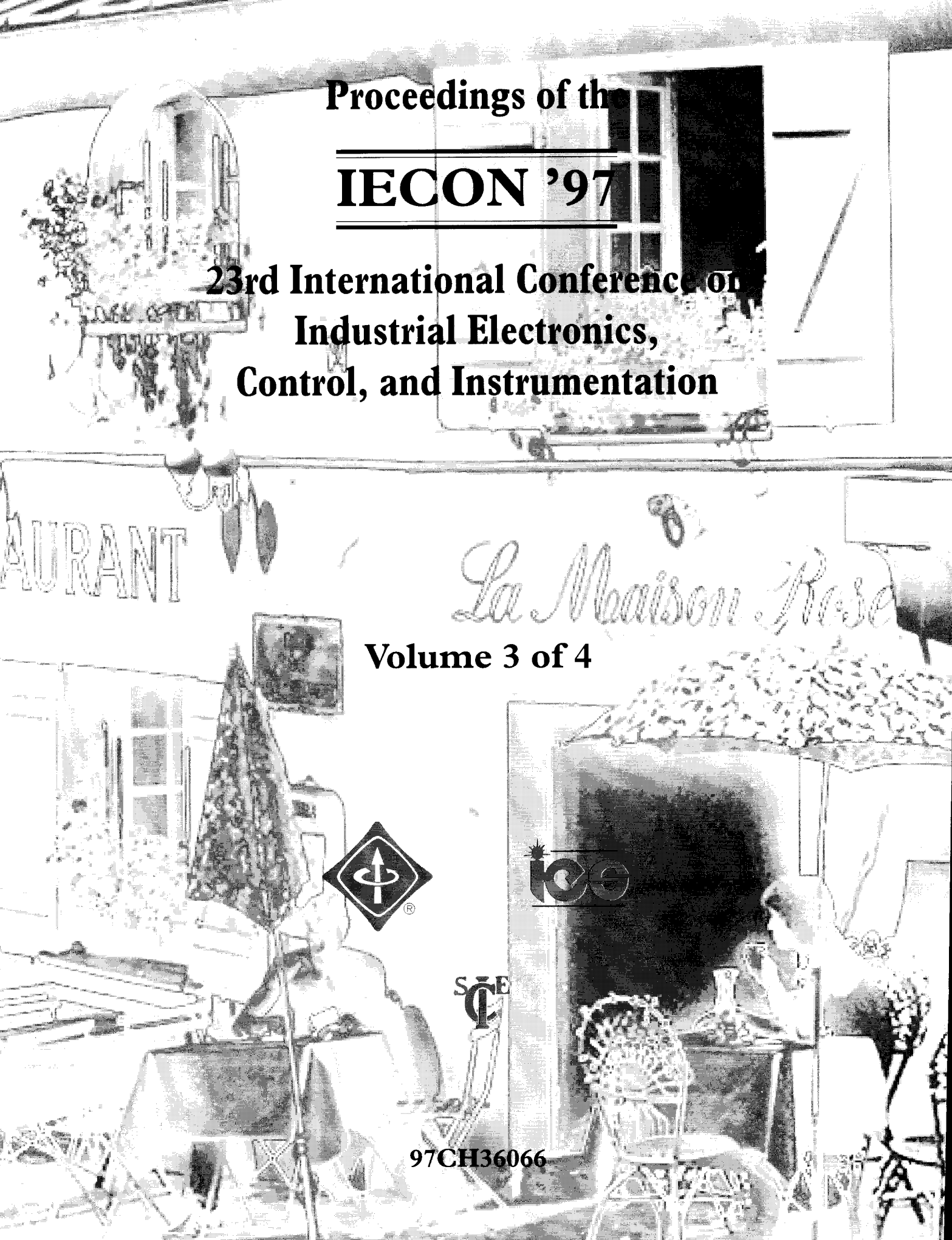
[21] Deller, J. R., Proakis, J. G. and Hansen, J. H., "Discrete-time Processing of Speech Signals," Macmillan, New York, NY, 1993.

[22] Assaleh, K. T., Mammone, R. J., Rahim, M. G. and Flanagan J. L., "Speech Recognition Using the Modulation Model," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. **2**, pp. 664-667, April, 1993.

[23] Rabiner, L. R. and Juang, B. H., "Fundamentals of Speech recognition," Prentice Hall, Englewood Cliffs, NJ, 1993.

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

[25] Mammone, R., Zhang, X., Ramachandran, R., "Robust Speaker Recognition," *IEEE Signal Processing Magazine*, September, 1996.



Proceedings of the

IECON '97

23rd International Conference on Industrial Electronics, Control, and Instrumentation

Volume 3 of 4

97CH36066