TEXT-INDEPENDENT SPEAKER IDENTIFICATION SYSTEM USING PROBABLISTIC NEURAL NETWORK

Taif A. Mehdi

Prof. Mahir K. Mahmood

Electrical Engineering Department, College of Engineering, Al-Mustansiriya University

ABSTRACT

Text-independent closed set speaker identification system is achieved using a Probabilistic Neural Network (PNN) as a classifier and Reflection Coefficients(RC) as a speaker feature. The system is evaluated with a database consisting of 28 speakers(21 male and 7 female). Each speaker has three totally different sentences, the first is used for training and the rest are considered as a testing sentences. The system correctly identified all the database speakers when tested with noise free speech for the two test sentences. For 30 and 20 dB SNR noisy speech, the performance is almost unchanged.

Keywords: Artificial Neural Network (ANN), Automatic Speaker Recognition (ASR), Cepstral Analysis, Probabilistic Neural Network (PNN), Speech Processing, Voice Biometric.

الملخص

يقدم هذا البحث محاكاة لنظام تعريف المتكلم غير المعتمد على النص و من النوع المغلق. أستعمل النظام معاملات الانعكاس (RC) كمعلم للمتكلّمِ ، والمصنف كَانَ شبكة عصبيةَ احتمالية (PNN). قُيّمَ النظام بقاعدةِ بيانات تشْملُ 28 متكلمِ (21 ذكر و7 إناث). قرأ كُلّ متكلّم ثلاثة جُمَلِ مختلفةِ كلياً، الأولى استعملت للتدريب والبقيةِ اعتبرت جُمَل اختبار. تمكن النظام من تعريف جميع المتكلمين في قاعدةِ البيانات بصورة صحيحة عندما أختبر بالكلام النقي لجملتي الاختبار. ل30، 20 dB كان أداء النظام مقارب للنتيجة السابقة.

I. INTRODUCTION

Automatic Speaker Recognition(ASR) (also called Voice Biometric)is the use of a machine to extract, characterize and recognize the information about speaker identity from a spoken phrase[1]. Banking by telephone, telephone shopping, database access services, information services, security control for confidential information areas, and remote access to computers are applications of Automatic Speaker Recognition. Speaker recognition has two branches: identification and verification. Speaker verification task is to verify the claimed identity of person from his voice. In speaker identification there is no identity claim and the system decides who the speaking person is [2]. Open-set speaker identification decides to whom of the registered speakers unknown speech sample belongs or makes a conclusion that the speech sample is unknown. Closed-set speaker identification decides to whom of the registered speakers unknown speech sample belongs. Depending on the algorithm used for the identification, the task can also be divided into text-dependent and text-independent identification. The difference is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text. The process of speaker identification is divided into two main phases: speaker enrollment and identification phase. Both phases include the same first step, feature extraction. The variations between speakers is called inter-speaker variations and the term intra-speaker represents variations for the same speaker. For speaker recognition, features that exhibits high speaker discrimination power, high inter-speaker variability , and low intra-speaker variability are desired[3]. 1978 Larry L. Pfeifer utilized vowel sound as a basis for extracting speaker characteristic and developed a method of text-independent speaker identification system, and a rate of 95 percent of correct identified speaker (from a database of 20 speakers, 10 male and 10 female) . Twelve reflection coefficients were used as a feature [4]. In 1999 Stephen A. Zahorian described a new neural network algorithm for speaker identification with large groups of speakers. The technique was derived from a technique in which an N-way speaker identification task is partitioned into N*(N-1)/2 two-way classification tasks. In that new approach, two-way neural network classifiers, each of which is trained only to separate two speakers, are also used to separate other pairs of speakers. High recognition rate was gained from that network[5]. In 2003 Mustafa Sarimollaoglu, Serhan Dagtas, Kamran Iqbal and Coskun Bayrak developed a text-independent speaker identification system based on Probabilistic Neural Network (PNN). PNNs supply flexibility and straightforward design make the system easily operable along with the successful classification results. The system was able to correctly identify 96% of the speakers, using 0.8 seconds of test samples from each speaker. 20 Mel-scaled cepstral coefficients (excluding 0th) were used as a feature vector. The database consists of speech samples from 28 adults, 21 male and 7 female[6].

Conclusive model of speaker identification system is shown in Figure (1).

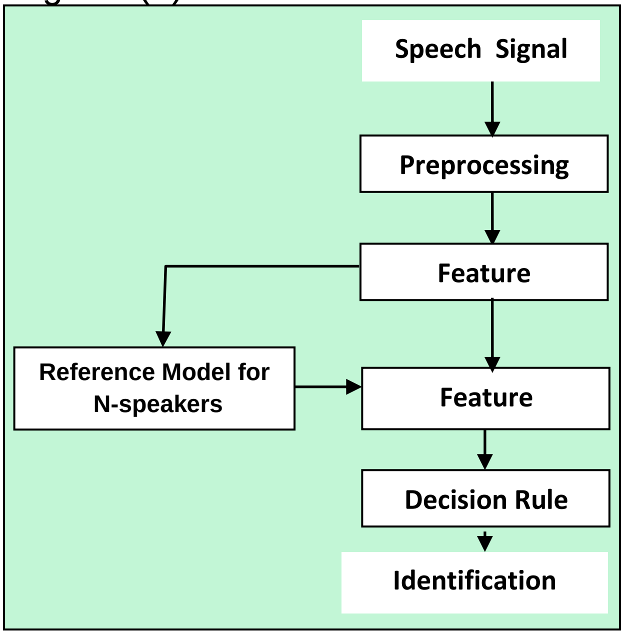

Figure (1) Speaker Identification System

The aim of feature Extraction is to transform raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. Feature Matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the one from a set of known speakers. One of the approaches used is the Probabilistic Neural Network (PNN) [1].

II. SPEECH PROCESSING TECHNIQUES

Since the speech signal is a slowly varying signal or a "quasi-stationary" when examined over a sufficiently short period of time (20-30 milliseconds)[7], this leads to the useful concept of describing human speech signal, called "short-term analysis", where only a portion of the signal(frame) is used to extract signal features one at a time.

Speech production can be modeled by the so-called "source-filter" model. The voice source is either a periodic pulse stream or uncorrelated white noise, or a combination of these[1].

It can be noted that speech signal is predicted as a linear combination of the previous *p* samples. Therefore, the speech production model is often called linear prediction (LP) model or the autoregressive model.

Before the system calculates the features, the speech signal is segmented(with each segment lasts for 20-30 milliseconds) [7] into overlapping(with 20-50% from the frame length) (to ensure that the samples in the last period in previous segment, will not be lost or attenuated very much and, therefore uncounted) frames. Theses frames are windowed through a appropriate window like "Hamming Window"[8]. Then the needed techniques are applied resulting in coefficients (each coefficients vector resulting from only one frame) like Linear Prediction Coefficients(LPC), Reflection Coefficients(RC) , Linear Prediction Cepstral Coefficients(LPCC) and Mel-Frequency Cepstral Coefficients (MFCC)….etc.

When the vocal tract is modeled with the lossless tube model[7], at each tube junction, part of the wave is transmitted and the remainder is reflected back, The reflection coefficients are the percentage of the reflection at these discontinuities. If Levinson-Durbin's algorithm[9] is used to solve the LPC(Linear Prediction Coefficient) equations[1], the reflection coefficients κ_m are the intermediate variables in the recursion.

III. PROBABLISTIC NEURAL NETWORK(PNN)

Probabilistic Neural Network is one of many neural networks[10], their general use is in classification problems. Their advantage over other methods is flexibility and the straightforward design.

Training time which increases substantially with increasing population size is not a disadvantage for PNNs. The PNNs implement window estimator by using a mixture of Gaussian basis functions. If a PNN for classification in K classes(in speaker recognition K referred to the number of speakers) is considered, the probability density function $f_i(x_p)$ of each class k_i is defined by:

$$f_i(x_p)=\frac{1}{(2\pi)^{d/2}\sigma^d M_i}\sum_{j=1}^{M_i}\exp(-\frac{1}{2\sigma^2}(x_p-x_{ij})^T(x_p-x_{ij})),i=1,2,....K........(1)$$

where x_{ij} is the j-th training vector from class k_i, x_p is the p-th input vector, d is the dimension of the speech feature vectors, and M_i is the number of training patterns in class k_i. Each training vector x_{ij} is assumed a centre of a kernel function, and consequently the number of pattern units in the first hidden layer of the neural network is given as a sum of the pattern units for all the classes. The variance σ^2 acts as a smoothing factor, which softens the surface defined by the multiple Gaussian functions.

The PNN used in the system is shown in Figure (2). This model has two layers: radial basis layer and competitive layer. In the first layer(the number of its neurons is Q), vector distances between the input vector p and each row of the input weight matrix IW1,1 are calculated and then multiplied by the bias b(which is a vector all its elements have the same value as given by Equation(3)with length Q). Activation function of the radial basis layer is given as:

$$radbas(n) = e^{-n^2} \quad ....(2)$$

radbas function produces its maximum of 1 where the input p is identical to w , and the output decreases as the distance between p and w increases.

Bias *b* allows the sensitivity to be adjusted and defined as:

$$b^1 = \frac{\sqrt{-\ln 0.5}}{\sigma} \quad ....(3)$$

The spread of radial basis function (σ) represents the typical distance between input vectors. Weights of the two layers (*IW*1,1 , *LW*2,1) are set to the matrices which are formed from the training vectors and the target vectors, respectively. The number of classes of input data, K, is equal to the number of users trained in the system. Q is the number of input/target pairs. Each target vector has K elements one of which is 1 and the others are 0. No training is needed in the construction of PNN in contrast to other neural networks. Input feature vectors are directly as e weight matrix *IW*1,1. This property saves processing time for training beside more neurons are required. In the testing phase, feature vectors extracted from a test sample are entered to the PNN as inputs. In the presence of an input vector, distances between the input and the training vectors are calculated and multiplied by *b* to form n1. *radbas* function designates a number close to 1 where input is close to a training vector. An input vector may be close to several training vectors, which causes a1 to have several elements close to 1. In the second layer, a1 is multiplied by the *LW*2,1. This operation sums the groups of elements which correspond to each of the K classes. Then, *compete* function (which assigns 1 for the largest value, and 0 for all the others) is applied. Finally, occurrence of 1 in the kth row of the output a2 means that the input vector *most probably belongs* to the kth class i.e., the kth speaker.

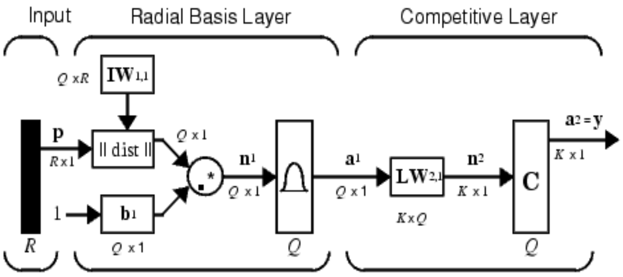

Figure (2) PNN Architecture and radbas function

IV. SIMULATION RESULTS

This section presents the results obtained from the simulation of the closed set text-independent speaker identification systems on personal computer. The simulation has been performed on the Pentium IV personal computer of 2.6 GHz CPU speed, using COOL edit pro.2000 as a speech editor and MATLAB software package (version 7) to simulate the system.

In this system, firstly the RC are calculated then used as an input features to PNN. The system is tested first with original speech data(noise free speech) and then an Additive White Gaussian Noise AWGN is added with signal-to-noise ratios(SNR)of 10,20 and 30 dB. PNN database consists of 28 speakers (21 males and 7 females). The number of speakers in the database is chosen in order to make the program execution time reasonable. All speakers read three sentences (one is used for training and the two others are for testing). The sentences were totally different. The recording is done by means of COOL edit software. Each sentence is of approximately 3-seconds long, and the recording is performed in normal room conditions with a sound card. The speech is sampled at 16 KHz with 16-bits/sample.

A. Simulation of the PNN System

The continuous speech signal is sectioned into frames of N samples with adjacent frames overlapping of L samples (L<N), The chosen values are N=320 samples(which is 20 ms) and L=120 samples.

A frame windowing is done using a Hamming window, and reflection coefficients(RC) are used as input to the network. Prediction order of RC was 18, 22, 26 and 30 and is denoted by "p".

To study the influence of number of frames applied to the PNN(i.e. the length of training and testing sentences)on the system, the number of frames are changed, as: 380,340,300,260,220,180,140,100 and 60. These number of frames are equal to: 4.7575, 4.2575, 3.7575, 3.2575, 2.7575, 2.2575, 1.7575 and 1.2575 seconds respectively. For cases where the number of required frames exceeds the sentence length, the same RC vectors were repeated to get the desired number. The same number of frames were used for training and testing.

During training, the RC vectors obtained from each speaker are used directly as an input weight matrix *IW*1,1, while the 1's in matrix *LW*2,1 are correctly positioned in correspondence with each speaker RC vectors .

During the testing, the RC vectors will be treated separately, i.e. each vector with length (p*1) is applied to PNN. The result of this vector is a number representing the speaker number.

When all the vectors from test utterance are applied, the system will count the number of times in which each speaker is appeared (for the present test utterance) and the speaker with largest number of appearance will be considered as the correct speaker .

The term "test 1" refers to the test with the second sentence in the database; and the term "test 2" refers to the test with the third sentence in the database. The system is trained with the 1st sentence.

Changing the order "p" and test sentences, the results for testing noise free speech for the PNN system are shown in Tables (1),(2),(3),(4),(5),(6),(7)and(8). The results for testing noisy speech for the PNN system(with p=30 and spread=0.1) are shown in Tables (9) ,(10) and (11).

Table(1)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=30 and test 1

Number of Frames	Spread		
	0.05	0.1	0.15
380	28	28	28
340	28	28	28
300	28	28	28
260	28	28	28
220	28	28	28
180	28	28	28
140	28	28	28
100	25	25	24
60	16	14	15

Table(2)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=30 and test 2

Number of Frames	Spread		
	0.05	0.1	0.15
380	28	28	27
340	28	28	27
300	28	28	27
260	28	28	27
220	27	28	28
180	28	28	28
140	27	28	26
100	21	21	21
60	18	17	15

Table(3)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=26 and test 1

Number of Frames	Spread		
	0.05	0.1	0.15
380	28	28	28
340	28	28	28
300	28	28	28
260	28	28	28
220	28	28	28
180	28	28	28
140	27	27	27
100	24	25	24
60	14	19	18

Table(4)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=26 and test 2

Number of Frames	Spread		
	0.05	0.1	0.15
380	27	27	27
340	27	27	27
300	27	28	26
260	27	28	28
220	27	27	27
180	27	28	27
140	27	27	27
100	24	24	21
60	21	20	19

Table(5)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=22 and test 1

Number of Frames	Spread		
	0.05	0.1	0.15
380	28	28	28
340	28	28	28
300	28	28	28
260	28	28	28
220	28	28	28
180	28	27	27
140	27	27	27
100	25	25	24
60	16	15	18

Table(6)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=22 and test 2

Number of Frames	Spread		
	0.05	0.1	0.15
380	27	28	27
340	27	28	25
300	27	28	25
260	27	28	26
220	27	28	27
180	26	27	27
140	26	26	27
100	22	23	20
60	18	17	16

Table(7)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=18 and test 1

Number of Frames	Spread		
	0.05	0.1	0.15
380	28	28	28
340	28	28	27
300	28	28	27
260	28	28	28
220	28	28	28
180	28	28	27
140	26	27	27
100	23	24	24
60	14	15	15

Table(8)
Number of correct identified speakers for 28 speakers with noise free tested speech with p=18 and test 2

Number of Frames	Spread		
	0.05	0.1	0.15
380	27	27	25
340	27	28	25
300	27	27	26
260	27	28	26
220	27	27	26
180	26	27	25
140	25	24	25
100	21	22	22
60	22	20	19

Table(9)
Number of correct identified speakers for 28 speakers with noisy tested speech with SNR=10dB, p=30 and spread=0.1

Number of Frames	test1	test2
380	1	0
340	1	0
300	1	0
260	1	0
220	1	0
180	0	0
140	1	1
100	0	1
60	1	1

Table(10)
Number of correct identified speakers for 28 speakers with noisy tested speech with SNR=20dB, p=30 and spread=0.1

Number of Frames	test1	test2
380	26	26
340	26	24
300	25	23
260	26	25
220	26	25
180	26	23
140	24	19
100	20	17
60	11	14

Table(11)
Number of correct identified speakers for 28 speakers with noisy tested speech with SNR=30dB, p=30 and spread=0.1

Number of Frames	test1	test2
380	28	27
340	28	27
300	28	28
260	28	28
220	28	27
180	28	27
140	28	28
100	25	21
60	17	19

V. Conclusion

The PNN is an efficient method when used with the Reflection Coefficients. The number of frames taken from the speech signal is a crucial factor in the system performance, and a minimum of that number is required to make the system works properly. The prediction order is an important factor. For the PNN system used in this study, if the two previous factors are adjusted in a right manner, the network "spread" has a little impact on the system performance. Finally, the system performs well with SNR of 20 and 30 dB.

References

1. Taif A. Mehdi," Text-Independent Speaker Identification System ", M.Sc. Thesis, College of Engineering, University of AL-Mustansiriya, 2009.
2. Jayant M. Naik, "Speaker Verification: A Tutorial " , IEEE Communication Magazine, pp.42-48, January 1990.
3. Joseph P. Campbell, Jr." Speaker Recognition: A Tutorial ", Proceedings of the IEEE, Vol. 85, No. 9, pp.1437-1462, September 1997.
4. Larry L. Pfeifer, "New Techniques For Text-Independent Speaker Identification ", IEEE International Conference on Acoustic, Speech and Signal Processing, ICASSP'78, Vol. 3, pp. 283-286, April 1978.
5. Stephen A. Zahorian, "Reusable Binary-Paired Partitioned neural Network For Text-Independent Speaker Identification ", IEEE International Conference on Acoustic, Speech and Signal Processing, ICASSP'99, Vol. 2, pp. 849-852, March 1999.
6. M. Sarimollaoglu, S. Dagtas, K. Iqbal and C. Bayrak, "A Text-Independent Speaker Identification System Using Probabilistic Neural Network", http://bayrak.ualr.edu/symsel/mustafa/docs/CCCT-2004.pdf.

   was valid at 17-5-2007.

7. L. R. Rabiner and Ronald W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, New Jercy, 1978.
8. R. E. Ziemer, W. H. Tranter and D. R. Fannin, "Signals And Systems: Continuous And Discrete", Prentice Hall, New Jersy, 1998.
9. John G. Proakis, "Digital Communications", McGraw-Hill International Editions, 1989.
10. Simon Haykin, " Neural Network- a comprehensive foundation " Prentice Hall, 1998.