(JZS) Journal of Zankoy Sulaimani, December 2008, 10(1) Part A (61-66)
گۆڤاری زانکۆی سلێمانی ،کانونی یەکەمی 2008 بەرگ 10(1) بەشی A لاپەرە (61-66)

Speaker Identification Using Backpropagation Neural Network.

***Suzan A. Mahmood,** College of Science, Computer Department, Sulaimany University, Kurdistan Region / Iraq
Loay E. George, College of Science, Computer Department, Baghdad University / Iraq.

Abstract

In this paper, a neural-based system for speaker identification is investigated. The linear predictive coefficients have been used as a set of descriptors for the speaker data. This set of descriptors is fed to a back propagation neural network for the purpose of speaker identification. Also, it is found that the process of segmenting the audio data into slices has great effect on the identification results, the number and length of slice has major influence on the success of the identification process.

Keywords: Leaner prediction coding, Speaker identification, backpropagation NN.

Introduction

In a world where authentication and privacy are taking a lot of our daily efforts, it is becoming more and more important for us to prove our identity to different systems every day so that we access required and useful services. A common user has at least 3 to 4 passwords or identification IDs that he has to use on a daily basis ranging from ATM pin numbers, to Internet access passwords. While sophisticated systems are developed, discussed and debated every day, other more natural ones are emerging as an alternative to password policy or in some cases as supplementary. These might include fingerprints, face, eyes, and speaker identification [1]. Previous researches relied heavily on the Hidden Markov Model in Speaker recognition [2]. In this work starting with processing the voice signal to make it suitable for the second part, the identification part being based on a neural network, The LPC coefficients have been used as input to a neural network for speaker identification, with the process of over lapping segments of audio data shows high performance of speaker identification .

System Layout

Figure (1) presents the layout of the neural based system. The lifecycle of the processing passes through two stages: training stage and recognition stage.
In the first stage of the design, the speech is appropriately processed to be input to the neural networks. By this we imply feature extraction achieved through modeling the human vocal tract using linear predictive coding which is then converted to the more robust cepstral coefficients using gausses elimination algorithm. The second stage of the design is to train the system for different utterances of the words. These utterances should constitute a good sample set of the various conditions and situations in which the word may be pronounced [3].

E-mail: suzanara80@yahoo.com
**Cited from her Ph.D. Thesis*

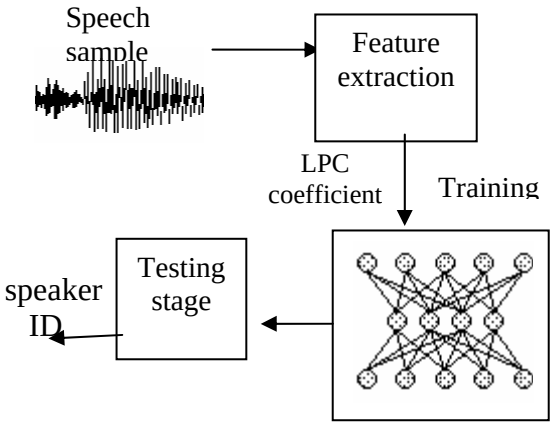


Figure (1) the layout of the proposed neural network based speaker identification system.

This training was implemented on back propagation neural network by using the back propagation training algorithm with momentum and variable learning rate. The last stage in this project is the testing. The system was tested under different conditions: noisy and clean environments, speakers who trained the system and new speakers.

Feature Extraction

Speech acquisition begins with a person speaking into a microphone. This act of speaking produces a sound pressure wave that forms an acoustic signal [4]. The microphone receives the acoustic signal and converts it to an analog signal that can be understood by an electronic device. Finally, in order to store the analog signal on a computer, it must be converted to a digital signal [3].

Since the vocal tract moves mechanically slowly, speech can be assumed to be a random process with slowly varying properties [5]. Hence, the speech is divided into overlapping frames of 40ms. The speech signal is assumed to be stationary over each frame and this property will prove useful in the following steps.

To minimize the discontinuity of a signal at the beginning and end of each frame, each frame was windowed to increase the correlation of the linear predictive coding (LPC) spectral estimates between consecutive frames. The windowing tapers the signal to zero at the beginning and end of each frame.

Linear Predictive Coefficient (LPC)

This captured signal by the microphone contains information in a form not suitable for pattern recognition. However, it can be represented by a limited set of features relevant for the task. These features more closely describe the variability of the phonemes that constitute each word. The feature measurements of speech signals are typically extracted using one of the following spectral analysis techniques: filter bank analyzer, LPC analysis or discrete Fourier transform analysis [5]. Since LPC is one of the most powerful speech analysis techniques for extracting good quality features and hence encoding the speech signal at a low bit rate, it was selected it to extract the features of the speech signal .

A well-known speech coding strategy is to assume that each speech sample can be estimated from linear combination of previous speech samples as described in the following equation:

$$\hat{X} = \sum_{k=1}^{p} a_k X(n-k) \quad \ldots\ldots\ldots\ldots\ldots (1)$$

where, X(n) is the nth speech sample, a_k are p-order linear prediction coefficients (LPC), which can represent the speech signal. The way to find LPC was implemented using the autocorrelation method [2]. The criteria of finding the

values of a's coefficient is the least square error:

$$E_n = \sum_m \left[X_n(m) - \sum_{k=1}^{p} a_k X_n(m-k) \right]^2 \quad \ldots (2)$$

Neural Networks Implementation

Neural networks attempt to mimic some or all of the characteristics of biological neurons that form the structural constituents of the brain [6]. In this paper, the back propagation neural network was adopted since it has been successfully applied to many pattern classification problems including speaker recognition and our problem has been considered to be suitable.

The structure of the back propagation neural network consists of three layers as it is shown in Figure(2): first layer has 210((no. LPC order) * (no. block -1)) input neurons which are fully connected to the hidden layer. The last layer is the output layer consisting of 4 neurons whose output uses to binary encoding ID of 10 speakers. All three layers are fully feed forwarded trained.

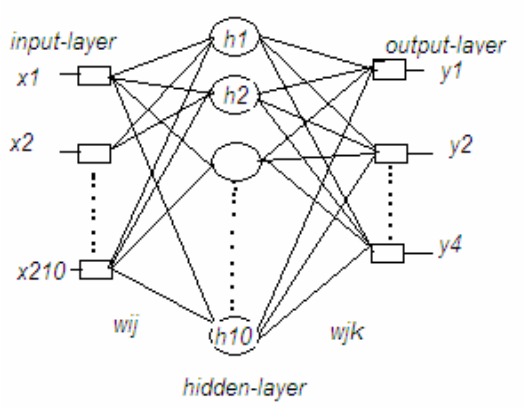


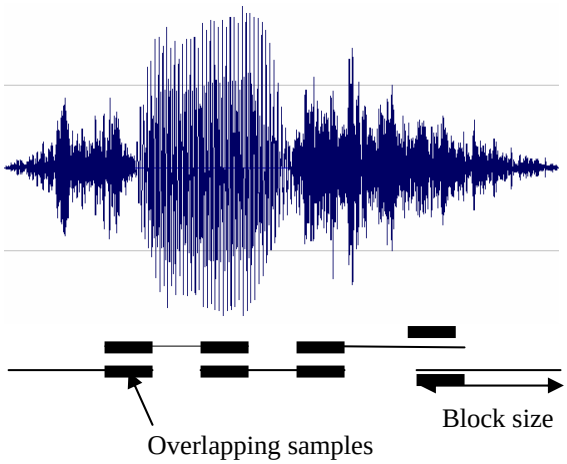
Figure (2): Multi –layer perceptron

In this work a lot number of hidden nodes have been tested to reach the best training results. Also the effects of learning rate value were tested and tabulated.

Segmentation

The segmentation of the samples into a specific number of overlapping slices is shown in Figure (3). Various lengths of slices have been tested to determine the best block length.

Some part of each block (n) is overlapped with the block (n-1) and block (n+1).



Figure(3) Overlapped segmentation

In this paper implemented the following two equation to calculate the overlapping ratio to solve the problem of variant length of the word :

$$N_b = \left\lfloor \frac{N_s}{S} \right\rfloor \quad ,\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

$$r_o = \frac{N_s}{SN_b} \quad ,\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

where, N_s is the number of samples, N_b number of blocks, r_o is the overlapping ratio, and S is block size.
The signal is related to the innovation through the linear difference equation and using Gaussian elimination method to solve it.

Test Results

A set of tests have been conducted on 10 words spoken by 10 speakers. In the experiment, the audio samples were recorded in an office environment at 11.025 kHz sampling rate, 16 bit and single channel. The utterance is obtained from 10 local speakers (5 females and 5 males). Each speaker was asked to pronounce each word 10 times. Therefore, the total utterances are 100 utterances for each word.
6 samples from each spoken utterance were used as training set. The 15th order autocorrelation analysis was utilized to determine the LPC coefficients (i.e, the elements of the feature vector). The LPC analysis was performed for every 30-50ms speech frame.
As we can see from Figure (4), show the part of LPC coefficient for one person pronouncing the word (digit 2) 8 times, the coefficient peaks tend to occur at the same locations. .

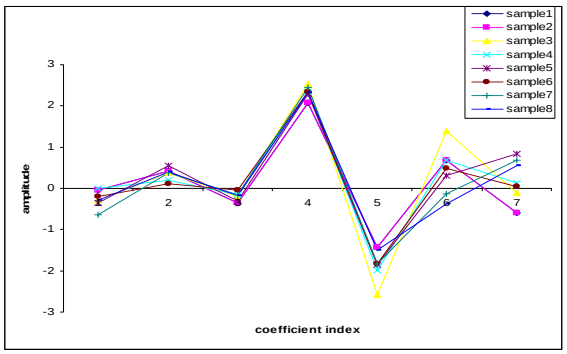
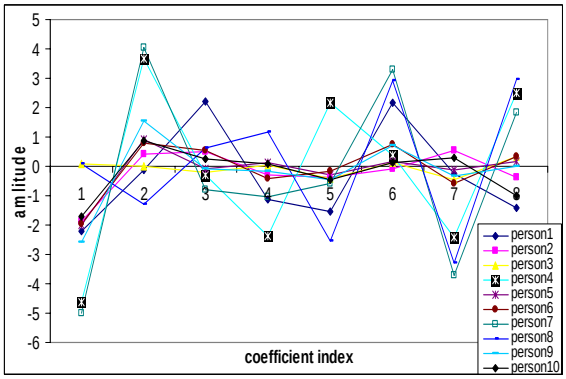


Figure (4) Part of LPC coefficient for person1 and word (two) pronounced six times

While in Figure(5) the same word for 8 persons the coefficient peaks tend to occur at the different locations. These are the formant locations. When used properly, the LPC coefficients emphasize the location of the formants in the frequency spectrum.



Figure(5) Part of LPC coefficient for 10 persons for word (two).

They are also greatly influenced by the glottal shape and vocal cord duty cycles.
It is found that the identification rates of all eight trained words by neural network approach 100%. The effects of learning rate value were tested and tabulated in Table(1).

Table (1) Identification rate with variant block size and learning rate.

Leaning rate	Block size	Rate (%)
0.01	400	65
0.05	400	60
0.1	400	62
0.01	512	55
0.05	512	57
0.1	512	58
0.01	1024	90
0.05	1024	75
0.1	1024	77

And the identification rates (in percentage) for all tested data are shown in Table (2), it is clear that for block size 1024 , learning rate 0.01 and 20 hidden nodes speaker (1/male) and speaker (9/female) have a high identification rate while speaker(4/male) and speaker (10/female) less than 85%. However, despite that the average correct predictions for each person were still very good. Also %50 overlapping segment was tested as it was implemented, the results of the identification shown in Table (3).

Table(2) For block size 1024 and learning rate 0.01

Speaker ID	Identification rate %		
	Hidden 10 nodes	Hidden 15 node	Hidden 20 nodes
1	**80**	**82**	**98**
2	**82**	**83**	**90**
3	**85**	**87**	**88**
4	**86**	**90**	**82**
5	**84**	**85**	**90**
6	**80**	**81**	**90**
7	**78**	**82**	**87**
8	**65**	**78**	**92**
9	**73**	**78**	**98**
10	**70**	**70**	**80**

Table(3) Identification rate for %50 overlapping with learning rate 0.01and 20 hidden nodes

Speaker ID	Identification rate %
1	**85**
2	**82**
3	**67**
4	**65**
5	**70**
6	**70**
7	**65**
8	**84**
9	**80**
10	**72**

Conclusions

LPC speech feature and a back propagation neural network are appropriate to use for a speaker identification system.

The best average identification rate is over 90% for 10 speakers with ANN. In the comparison between ANN with variant parameters like learning rate and number of hidden nodes as an identification engine, rate with 0.01, 20 hidden nodes, and block size 1024 samples shows a great powerful nonlinear recognition performance.

In comparison to the previous works, this study revealed more accurate results because they used in their analysis %50 overlapping segments.

To improve system capability, recording the speech in non noisy environment for training the neural and appropriate length of speech duration should be selected for speaking sentence since it can cover more personal characteristics than using each tone in all utterances.

References

1. Marek Szala "Two-level pattern recognition in a class of knowledge-based systems" , *Knowledge-Based Systems*, **2002,** 15(1), *95-101*.
2. *Amarin Deemagarn, Asanee Kawtrakul* " Thai Connected Digit Speech Recognition Using Hidden Markov Models", *International Conference Speech and Computer*, **2004**, 9(1),209-213.
3. Choubassi, M. M. El, El Khoury, H. E., Jabra Alagha, C. E.,Skaf, J.A.and Al-Alaoui, M.A. "Arabic Speech Recognition Using Recurrent Neural Networks", *IEEE Intl. Symp. Signal Processing and Information Technology ISSPIT*, **2003**, 3(1), 336-340.
4. Campbell,J. P.Jr. , "Speaker Recognition",*PROCEEDINGS OF THE IEEE*,**1997**, 85(9), 284-366.
5. Farell, K., Mammone,R & Assaleh, K."Speaker Recognition Using Neural Networks and Conventional Classifiers" *IEEE Transactions on Speech and Audio Processing*, **1994**, 2(1), Part II, 194-205.
6. Chularat Tanprasert, " Text-dependent Speaker Identification Using Neural Network On Distinctive Thai Tone Marks", *NECTEC Technical Journal*, **2000**,1(6), 123-129.

ناسینەوەی ووتە بێژ بە بەکار هێنانی تۆری دەماری بۆدواوە بلاوبوونەوە.

سوزان عبدالله محمود، کۆلیجی زانست، بەشی کومپیوتەر، زانکۆی سلێمانی، هەرێمی کوردستان / عێراق.
لوئەی ادور جورج، کۆلیجی زانست، بەشی کومپیوتەر، زانکۆی بغداد / عێراق

پوختـــە

تۆێژینەوەکەلە سەر دروست کردنی سستمێکی ناسینەوەی ووتە بێژە بەشێوەیەکی زیرەکانە. وە بە بەکار هێنانی داتای تایبەت کە بەندە بە پێشبینی راستەوانەی ووتە کە ، وە ئەم داتایە ئەدرێت بە تۆری دەماری لە جۆری دواوە بلاوەبوونەوە. وا دەرکەوت لە تۆێژینەوەکەدا کەبەش کردنی شە پۆلەکانی دەنگ بۆ چەند برگەیەك ئە نجامی ناسینەوەکە باشتر ئەبێت وە هەروەها ژمارەی برگەکان و درێژیان کارێکی راستەوخۆ گەورەیان هەیە لە پرۆسەی ناسینەوە.

تتميز المتحدثين باستخدام الشبكة العصبية ذات نشر أسترجاعي.

سوزان عبدالله محمود، كلية العلوم، قسم الكومبيوتر، جامعة السليمانية ، اقليم كوردستان / العراق .
لؤي ادور جورج، كلية العلوم، قسم الكومبيوتر، جامعة بغداد / العراق.

الخلاصة

يهدف البحث الى بناء منظومة تتميز المتحدثين و تتعرف على المتكلمين بدقة مستخدما أنظمة ذكية . تم أستخدام البيانات ذات الخصائص المعتمدة على التنبؤ الخطي لتغذية الشبكة العصبية ذات نشر أسترجاعي . يبين البحث أن تطبيق التجزئة للاشارات الصوتية الى المقاطع قد أدى الى تحسن معدل تعريف الهوية. و كذلك ان عدد المقاطع و أطوالها لها تأثير مباشر و كبير في عملية التمييز.

Received on 1/3/2007 Accepted 23/9/2007

وەرگیراوە لە 2007/3/1 پەسندکراوە لە 23/ 2007/9