مجلة ابن الهيثم للعلوم الصرفة والتطبيقية المجلد24 (1) 2011

معرفة الشخص المتكلم باستخدام تقنية التحول المويجي والشبكات العصبية

سالي علي عبد اللطيف ، انتصار عبد يوسف
قسم علوم الحاسبات ، كلية التربية ، الجامعة المستنصرية

استلم البحث في 21 شباط 2010
قبل البحث في 13 ايار 2010

الخلاصة

معرفة الشخص المتكلم هو آلية للتعرف ذاتيا على الشخص بالاعتماد على معلومات فريدة متضمنة في الموجات الكلامية او الصوتية الصادرة من الشخص المتحدث . هذه التقنية تجعل من الممكن استخدام صوت الشخص للتحقق من هويته ، للسيطرة على الوصول لخدمات اخرى مثل التعامل مع البنوك، او التسوق من خلال جهاز التليفون او الوصول الى بيانات ضرورية، او السيطرة الامنية للتعامل مع المعلومات .

اما الهدف من البحث هو بناء نظام معرفة المتكلم الذاتي لمجموعة محددة من الاشخاص في حالتي النص المعتمد والنص غير المعتمد. ان العمل بصورة عامة يتكون من طورين وهما : طور التدريب ، الذي يتضمن بناء قاعدة بيانات تشمل كل المتكلمين، والطور الثاني هو طور الاختبار او التعرف على المتكلم ، الذي يتضمن عملية مقارنة ما بين الانموذج غير المعرف مع التقدير قاعدة البيانات ، لتحديد المتكلم . ان تقنية التحول المويجي كانت قد انتشرت في معظم تطبيقات معالجة الاشارة الرقمية وقد أدى دورا مهما في معالجة الاشارة الصوتية وتحليلاتها ، وخاصة في تقنية معرفة المتكلم وذلك بسبب ادائه الاقوى فيما يتعلق بالتحليلات المتعددة الانتشار النظام المقترح يتشكل من ثلاث مراحل وهي :المرحلة الاولى وهي مرحلة التهيأ للمعالجة وفيها يتم تقطيع الاشارة الى اطر عديدة وكل واحد من هذه الاطر سوف يضرب بـ Hamming window

اما المرحلة الثانية فهي مرحلة استخلاص الخواص وفيها يتم استخلاص الصفات المميزة لكل كلمة مدخلة باستخدام تقنية (التحول المويجي) وان النتيجة من هذه المرحلة هو متجه خواص. وفي المرحلة الاخيرة وهي مرحلة التصنيف وفيها يتم استعمال (متجه الخواص) المنتج من المرحلة السابقة بوصفه مدخلا" للشبكة العصبية . القيم الناتجة تبين ان الخواص الصوتية تكون فعالة جداً في معرفة المتكلم ، وان الخوارزمية المقترحة هي فعالة فيما يتعلق بـ (تقليل العمليات الحسابية، و تقليل الزمن المستغرق في التنفيذ).

Speaker Identification Using Wavelet Transform And Probabilistic Neural Network

S. A. Abdul- latef , E. A. Yosife
Department of Computer Science, College of Education, University of Al-Mustansiriya

Received in Feb. 21 2010
Accepted in May 31 2010

Abstract

Speaker identification is the process of automatically identify who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speakers voices to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. The goal of this project is to build an automatic speaker identification system for a closed set Text-Dependent & Text-Independent. This generally will include two main phases: The Training Phase , which is used to built the speakers database and the Identification (Testing) Phase , which is used to compare the unknown model with the models stored in the speakers Database .The wavelet transform is diffused into most digital signal processing applications. It is plays very important role in speech signal processing and analysis, and mainly in speaker identification because of its superior performance when used particularly in multi-resolution analysis. The proposed system constructs from three stages, the first stage is the Preprocessing stage, in which the speech signal is separated into many frames, and each frame is multiplied by (Hamming Window). In feature extraction stage, which is the second stage, the discriminative features of each spoken words are extracted by using the DWT technique, the resultant of this stage is the feature vector for each speaker. In the third stage, which is the classification stage, the feature vector of each speaker is used as an input to the neural network. The results show that phonetic features are powerful for speaker identification and the proposed algorithm is efficient concerning the minimizing of the calculation operations and reducing the execution time.

Introduction

The human speech conveys different types of information. The primary type is the meaning of words, which speaker tries to pass to the listener. But the other types that are also included in the speech are information about language being spoken, speaker emotions, gender and identity of the speaker. The goal of automatic speaker recognition system is to extract, characterize and recognize the information about speaker identity [1]. That deals with the proposed algorithm for Text-dependent and Text-Independent speaker identification system. Speaker recognition system attempt to recognize a speaker through measurements of specifically individual characteristics arising in speech signal [2].The proposed speaker identification system based on wavelet and networks will be presented. The wavelet analysis technique was used for feature extraction. Wavelet transform has been successfully applied to the processing of non-stationary speech signal [3].In the discrete case, filters of different cutoff frequencies are used to analyze the signal at different scales. The signal is passed through a series of high pass filters to analyze the high frequencies, and it is passed through a series of low pass filters to analyze the low frequencies. The resolution of the signal, which is a measure of the amount of detail information in the signal,is changed by the filtering

operations, and the scale is changed by upsampling and downsampling (subsampling) operations [4]. The basis of the neural networks is the neuron cell in the human brain. They study the trend of the input data and the output data of training set, which is fed to the network, and iteratively, estimate a multi-dimensional surface an analog function which is a very close approximation of the system being studied [5] . Artificial Neural Networks (ANNs), which have gained prominence in the area of pattern recognition, have several properties that make them attractive for speech recognition. These include a relatively simple implementation, inherently parallel algorithm (making parallel implementation of a natural progression), robustness to noise and self-learning ability. We have chosen the PNN in our system.

However, most of the classifiers have their own disadvantages due to complex distribution of the feature vectors. Probabilistic Neural Network (PNN)is one of promising classifiers because it is based on well-established statistical principles derived from Bayes' decision strategy and non-parametric kernal based estimators of probability density functions, further more PNN can classify samples in testing set with 100% correct rate (PNN) was used for classification procedure [6],[7].In the proposed method, which is called Discrete Wavelet Transform and Neural Networks (DWTNN), there is no need for time alignment because words will be transformed to a constant length vector.[4]

- Speaker Identification System Structure

Speaker identification is a computationally expensive task and requires a large amount of computations to identify the unknown speaker. In this work, we analyze the main speaker identification component. A Speaker Identification system is normally divided into two procedures, Enrolment Procedure and Testing Procedure. where each one can be divided into four "subsystems" or, in other words, has to accomplish four tasks: digitize the spoken utterance , divide it into frames and compute feature of each frame , and collect all features in one vector , as a feature vector , for any spoken utterance . [8] Then, classify each vector as belonging to a specific speaker with a neural network, and finally, give the neural network's outputs for each frame and determine who the speaker is.

- Training Procedure

During this procedure, the user needs to register him/her to the system. In other word, the user may provide the system with a set of utterances so that it can build his/her speech model and use it as a reference later. The procedure builds an initial reference template for a speaker by capturing the identification utterance [8]. However, in this work the reference template is used to train the system to get out the best rule to identification. It's important to obtain a good enrolment template for high performance system. The first stage is the Pre-processing subsystem, which prepare the training speech samples for the next modules. The next stage is the feature extraction subsystem, which aims to extract the information about the user conveyed in the training speech samples. And the final stage is the classification subsystem. which aims to build a suitable classifier that can efficiently distinguish between speakers.

- Testing Procedure

During this procedure the user provides a set of utterances, so as to authenticate him/her self to the system, for gain access to certain resources. In common methods the systems match the presented speech samples with the already recorded user's speech model in order to known who the speaker is.

- The Proposed Speaker Identification Algorithm

1. Input the signal.
2. Sampling speech data to 11,025 Hz, 16 bit Accuracy.
3. Max Frame=30.
4. Calculate the number of frame through the following equation:

$$\text{Number of Frame} = \frac{\textit{Length of Signal}}{\ldots\ldots} \quad \ldots\ldots (1\text{-}1)$$

5. If (Number of Frame=Max Frame) THEN go to step 8.

6. If (Number of Frame > Max Frame) THEN use only the 30 frames in the middle of signal and go to 8.

7. If (Number of Frame < Max Frame) THEN make overlapping between two adjacent frames.

$$\text{Overlapping rate} = 1 - \frac{\text{No. of frame}}{\text{Max no. of frame}} \qquad \ldots\ldots\ldots\ldots(1\text{-}2)$$

8. Multiply each frame by Hamming Window, which has the following form .

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (0 <= n <= N\text{-}1) \quad \ldots\ldots\ldots(1\text{-}3)$$

9.Calculate the Discrete Wavelet Transform for each output frame .

10. Compute the power (energy) through the following equation:

$$P_{norm} = \frac{\sum_{i=1}^{n} S_i^2}{n} \qquad \ldots\ldots\ldots\ldots(1\text{-}4)$$

Where

S_i : Is the element of coefficient set.

n : Is the number of elements.

11. Compute the variance for each feature vector through the following equation:

$$Var = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (1\text{-}5)$$

Where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \ldots\ldots\ldots (1\text{-}6)$$

n = the number of element of signal, x = signal.

12. Initialize all of the net parameters.

13. Estimate the target and input to the diagonal matrix, while input in the feature extraction from the net. Here there are speeches of 30 speakers with 5 words for each speaker.

14. Train the PNN net with the Target & Input.

- Preprocessing

The Preprocessing stage includes four cases: Sampling, Framing and Windowing, as below:-

* Sampling: The process of digitizing sound (broken down the sound into basically identical elements) is called sampling. This method allows a sound, which consist of an analog signal, to be transformed into digital data (i.e. bits and bytes).

* Framing: In this step, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M<N). The first frame consists of the first N samples. The second frame being M samples after the beginning of the first frame, and overlaps it by N-M samples. Similarly, the third frame being 2M samples after the beginning of the first frame (or M samples after the beginning of the second frame) and overlaps it by N-

2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N=256 (which is equivalent to ~ 30 msec).

* Windowing : If we can limit the signal to some finite range of times we have a physical releasable filter. One way to do this is to multiply the signal by some function that is non - zero only in finite range of time. Such a function is called a window. The advantage of multiplying each frame by window is to minimize the signal discontinuities at the beginning and end of each frame. And we notice that all records of the speakers did not start in the same beginning and did not finish in the same end, the numbers of samples in every reading is not similar even if we take it in the same person. So we suggest multiplying the Hamming window by the signal to avoid mismatch between the signals and the limited features of every speakers. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as w (n), 0 < n < N –1, where N is the number of samples in each frame, then the result of windowing is the signal

y1(n)=x1(n)* w(n), ……. $0 \leq n \leq N-1$

Where

x1: is the signal in frame 1;

w(n) : is the Hamming window ;

y1: is the result of multiplying the signal by the Hamming window.

It is clear that the Hamming window does not go to zero at the extremes.

- Feature Extraction

After data has been acquired, it must be pre-processed in order to reduce the extraneous data that is not pertinent to the classification process. The resulting variables of interest are known as features. The Discrete Wavelet Transform (DWT) is a popular technique in the field of digital signal processing. We thus introduce a simple feature extraction model based on the result of DWT. In order to parameterize the speech signal, we should first decompose the signal form using the MRA algorithm. The Multiresolution analysis (MRA) algorithm, which decompose a signal into scales with different time and frequency resolution. MRA is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. The fundamental concept involved in MRA is to find the average features and the details of the signal via scalar products with scaling signals and wavelets. The differences between different mother wavelet functions (e.g. Haar, Daubechies, Coiflets, Symlet, Biorthogonal and etc.) consist in how these scaling signals and the wavelets are defined. The choice of wavelet determines the final waveform shape.

The steps of feature extraction are illustrated as follows:

1. Each frame of the spoken words is now expanded using the Discrete Wavelet Transform (DWT) of 7 levels of decomposition.
2. By computing the power (energy) in each segment in each level of decomposition according to the equation (1-4), feature vectors will be obtained thatwould describ the power distribution over the time– frequency plane. This scale power density along every segment describes the power variation in each scale.
3. The result of the previous step is (8 values) of each frame, the variation of (P_{norm}) of each frame is computed according to the equation (1-5).
4. Collect all the variance values in one vector, where this vector represents the feature vector of one word of one person.

- Testing of the Proposed Speaker Identification System:

Now, the proposed method is implemented over the vocabulary of Arabic word, and made all operation on it.

* Data Collection

The data needed for this work was collected by the aid of a Sound Blaster(SB) card that offer possibilities for creating sound (Digitally record and replay sounds) . The speech sample recorded using 11,025 Hz sampling frequency. And saved it in ".wav" format, because this format gives flexibility in dealing with data. Also, we must apply some condition to prevent signal distortion, and these are:

1- Using (sound proof room) for recording samples to prevent reverberation due to reflections from objects such as walls and furniture.

2- Using noise canceling microphone and holding it close to mouth to avoid background noise.

3- Using the same microphone and amplifier in all sessions.

* Evaluation Tests

The experiments were designed to test the Text-Independent and Text-dependent speaker Identification on the proposed system, the search carried out two experiments when using different mother functions with PNN structure:

1 – Minimum time-consuming in feature extraction stage and PNN learning stage.

2- Maximum ability of speaker recognition.First, Preprocessing operating (Sampling, Framing, And Windowing) are run of the speech signal. After these operations, we make the DWT decomposition algorithm of Db4 type on the speech signal frames, and see the deferent levels of frames. The different levels and effective of DWT on frame we implemented up to 7 levels of decomposition. And the number of Coefficients in each level is computed according to this equation:

$$No.\ of\ Coefficient = \text{floor}\left(\frac{n-1}{2}\right) + N \qquad \text{..... (1-7)}$$

Where

N: is the order of the wavelet type, and here is 4.

n: is the length of the input vector to the law pass filter and high pass filter.

The first level consist of two sets coefficients , each one contains (131) coefficient according to the equation (1-7) , the second level contains two sets of coefficients, each one consists of (69) coefficients , the third level includes two sets each one contains (38) coefficients , the fourth level includes two sets each one contain (22) coefficients, the fifth level contain two sets each of which consists of (14) coefficients , the sixth level contain two sets , each one consist of (10) coefficients , the seventh level contain two sets , each one consist of (8) coefficients .Then, after the DWT Decomposition and fined the Coefficients at each level the power normalized for each set was calculated according to the equation (1-4). When the normalized power values are collected in a vector, the new vector yield, and the variance of all values in vector is computed, the produced element represents the power of one Frame. We repeated all these operations on all other frame in the signal , Finally , we receive one vector containing 40 values, which represent one word of one person .When we reaches the last word of last person , we obtained Matrix of (150*40) , which represent the 30

Person and each one spoken 5 words . The PNN is trained by (150*40) Matrix as Data Base , by using the training algorithm as below . As mention, The PNN is of type Supervised, Feedforward , Trained in time linear with the number of patterns and in one pass, and Classification only .it consists of nodes allocated in three layers after the inputs:

1- *Pattern layer*: There is one pattern node for each training example. Each pattern node forms a product of the weight vector and the given example for classification, i.e. each neuron in the pattern layer computes a distance measure between the unknown input and the training case represented by neuron. Where the weights entering a node are from a particular example. After that, the product is passed through the activation function:

$$\text{Exp}\ [(\ \mathbf{x}^T\mathbf{w}_{ki}\text{-}1\ )\ /\ \sigma^2\ ]\ \ldots\ldots\ (1\text{-}8)$$

2- *Summation layer*: each summation node receives the outputs from pattern nodes associated with a given class, i.e. there is one neuron for each class, these neurons sum the values of the pattern layer neurons corresponding to that class to obtain and estimate probability density function of that class:

$$\sum_{i=I}^{NK} \exp[\ (\ \mathbf{x}^T\mathbf{w}_{ki}\text{-}1\ )\ /\ \sigma^2\ ]\ \ldots\ldots..(1\text{-}9)$$

3- *Output layer*: the output nodes are binary neurons that produce the classification decision :

$$\sum_{i=I}^{NK} \exp[\ (\ \mathbf{x}^T\mathbf{w}_{ki}\text{-}1\ )\ /\ s^2\ ] > \sum_{i=I}^{NK} \exp[\ (\ \mathbf{x}^T\mathbf{w}_{kj}\text{-}1\ )\ /\ s^2\ ].....(1\text{-}10)$$

Results and Discussion

Computation and response time usually increases with population size. Pre-processing stage is applied to prepare the speech data, i.e the sound wave is appropriate to extract the feature from it. And then the process will continue to match the test pattern with the remaining reference patterns. In another word, using PNN to Classify the speaker and known that one .The is proposed system (DWT and PNN) is based on the feature extraction algorithm of DWT, then combining them with an intelligent solution of the Classification System. In Text-Dependent status, the proposed Speaker Identification System use the same word in Training and Testing stages, this work deals with 30 speakers (15 Male and 15 Female) , each one of them speak 5 words for training and three words for testing , and three different wavelet transform (Haar, Db4, Db7), get the best result from Db4 as compared with Db7 and Haar. In Text-Independent status, the proposed Speaker Identification System use the different words in Training and Testing stages, And also it deals with 30 speakers (15 Male and 15 Female) , each one of them speak 7 different words for training and testing them by using any other words , wavelet transform with Db4 function is used.Table (1.1) shows the result of applying the Db4 Wavelet function for Text-Independent.

Conclusions

The conclusions that one can draw from this work may be listed as follows:

1- A Discrete Wavelet Transform (DWT) and Probabilistic Neural Network (PNN) architecture for Text-Independent and Text-Dependent was proposed , In principle , Three- stage processing has been used in most cases and has proved particularly suitable. In the pre-processing stage parameters directly related to the process of speech production are obtained from speech signal. In reducing amount of test data or feature extraction stage a one-dimensional feature vector is derived from these parameters. In last stage, Classification stage, these feature vectors are used in training and testing

phases.

2- DWT is a good feature extraction methods, and to increase the efficiency of the algorithm and minimize the system response time, we combine the method with the intelligent solution of the classification, result indicated that by using wavelet decomposition with the classification process using Artificial Neural Network for classifying a close set of speakers and showing that the choice of wavelet type is important to reduce the error of the network. The suggested method in speaker identification gave high efficiency.

3- The feature vectors obtained in this proposed algorithm could not require a large storage area. This fact makes the extension of the number of feature vectors very simple.

4- The (Db4) gave us the best result as compared with (Haar) and (Db7) in the Processing phase in Text-Dependent and Text-Independent

Refrence

1. Reynolds,D.A. (2002), An Overview of Automatic Speaker Recognition Technology, ICASSP, 4072-4075.
2. Gopinath, R. A.; Odegard, J. E. and Burrus, C. S. April(1994), Optimal wavelet representation of signals and the wavelet sampling theorem", IEEE Transactions on Circuits and Systems II. 41(4): 262-277.
3. Flandrin, p. (1990), Wavelets and Related Time-Scale Transforms , in *Proceedings Of* SPIE Conference 1348 : Advanced Signal Processing Algorithms, Architectures. And *Implementations*. Society Of Photo-Optical Instrumentation Engineers.
4. " MULTIRESOLUTION ANALYSIS AND THE WAVELET TRANSFORM", http://users.rowan.edu/~polikar/ ,June , (1995)
5. " Artificial Neural Networks Technology" http://www.dacs.dtic.mil/techs/neural/Networkforclassification.html
6. Patterson, D. (1996)," *Artificial Neural Networks",* Singapore: Prentice Hall. http://www.statsoft.com/textbook/glosc.htm .
7. "*Implementing probabilistic neural networks*", Department of Biophysical and Electronics Engineering, University of Genova – Via all'Opera Pia 11a 16145 Genova, Italy
8. Minh.N.Do, January (2000), "An Automatic Speaker Recognition System " Swiss Federal Institute of Technology (SFIT), Lausanne , Epel. http://www.Icav.epfl.ch/~mindho/asrp-project/.html.

Table (1.1): Proposed System Results with Wavelet function (Db4) In Text-Independent Status

No.	Speaker No.	Program Output	Result
1.	Speaker 1	3	ok
2.	Speaker 2	20	Not
3.	Speaker 3	22	Not
4.	Speaker 4	25	ok
5.	Speaker 5	30	ok
6.	Speaker 6	4	Not
7.	Speaker 7	43	ok
8.	Speaker 8	52	ok
9.	Speaker 9	59	ok
10.	Speaker 10	67	ok
11.	Speaker 11	23	Not
12.	Speaker 12	83	ok
13.	Speaker 13	89	ok
14.	Speaker 14	95	ok
15.	Speaker 15	102	ok
16.	Speaker 16	107	ok
17.	Speaker 17	117	ok
18.	Speaker 18	122	ok
19.	Speaker 19	53	Not
20.	Speaker 20	137	ok
21.	Speaker 21	143	ok
22.	Speaker 22	150	ok
23.	Speaker 23	159	ok
24.	Speaker 24	110	Not
25.	Speaker 25	172	ok
26.	Speaker 26	180	ok
27.	Speaker 27	186	ok
28.	Speaker 28	193	ok
29.	Speaker 29	201	ok
30.	Speaker 30	205	ok

Success =24
Fail= 6
Time =3 minute
Success rate =80 %