Eng. & Tech. Journal, Vol. 29, No. 5, 2011

Speaker Recognition Based on Semantic Indexing

Dr. Mohammed Sahib Altaei*, Khamail Abbas* & Marwa Jawad*

Received on: 18/5/2010
Accepted on: 3/3/2011

Abstract

In this paper, a new pitch extraction method is established to be employed in improving the performance of the eigenvoices problem. This required indexing the pitch of the voice in a document matrix and then mapping the voice documents into preserved semantic features. The proposed voice recognition system was built to be operated in two phases; enrollment and recognition. Closed dataset of different voices belong to different sexes and ages of speakers were enrolled in the first phase. The results of the recognition phase were promising of about 81% for both sexes. This ensures the successful recognition task and the efficiency of the proposed system.

Keywords: pitch period, speaker recognition, eigenvoices.

تمييز اصوات المتكلمين بناءاً على الارشفة الدلالية

الخلاصة

في هذا البحث، أُعتمدت طريقة جديدة لاستخلاص نمط نغمة الصوت من اجل توظيفها لتطوير اداء منظومة تمييز الاصوات المبنية على مسألة القيم الذاتية. وهذا تطلب ارشفة الانماط الصوتية في مصفوفة وثائقية ومن ثم تحويل الوثائق الصوتية الى صفات معنوية محفوظة. بُنيت المنظومة المقترحة لتعمل بطورين هما: التجميع والتمييز. تم تجميع اصوات مختلفة تعود لمتكلمين ذوات اجناس واعمار مختلفة، كانت نتائج طور التمييزواعدة حيث اعطت درجة تمييز تقدر بحدود 81% لكلا الجنسين وهذا يؤكد نجاح عملية التمييز وكفاءة المنظومة المقترحة.

Introduction

Voice is a combination of physiological and behavioral biometrics. The features of an individual's voice are based on the shape and size of the appendages (e.g., vocal tracts, mouth, nasal cavities, and lips) that are used in the synthesis of the sound. These physiological characteristics of human voice are invariant for an individual, but the behavioral part of the voice of a person changes over time due to some factors; age, medical condition (such as a common cold), and Emotional state, etc. These factors affect the way of pronouncing a text, So that the voice recognition system is called text dependent or independent [4]. In general, voice recognition is biometric modality that uses an individual's voice for recognition purposes; voice is not very distinctive and may not be appropriate for some applications [1]. A text-dependent voice recognition system is based on the utterance of a fixed predetermined phrase, such as that used in the present work. A text-independent voice recognition system recognizes the speaker independent of what he/she speaks. such system is more

* College of Science, University of Al Nahrain / Baghdad

https://doi.org/10.30684/etj.26.5.9
University of Technology-Iraq, Baghdad, Iraq/2412-0758

difficult to design than a text-dependent system but offers more protection against fraud. A disadvantage of voice based recognition is that speech features are sensitive to a background noise; such system should uses filters in a preprocessing stage [5].

The voice recognition may perform two tasks: identification or verification. A voice *identification* system gets a test utterance as input. The task of the system is to find out which of the training speakers made the test utterance. So, the output of the system is the name of the training speaker, or possibly a rejection if the utterance has been made by an unknown person. For a system which does a *verification* of an utterance, the input is the speech signal to be verified as well as the name of the trained speaker who is to be verified. The expected result is a yes/no-decision: The acceptance of the test utterance if it does originate from the proclaimed speaker, and vice versa [11]. In voice identification, a training phase is required. For example; valid users of the system need to be enrolled. During the enrollment procedure, the system 'learns' the person it is supposed to identify. Voice samples of the user are required for this training phase. During the later identification process, the system compares another recorded utterance (called test data) to the training ones. The desired output of the system is the name of one of the training speakers, or a rejection if the test utterance stems from an unknown person. The enrolled dataset may be open or closed. If the system is provided with the information that all possible test utterances belong to one of the persons that have been learned by the system that means a 'closed set' of training speakers. If a test utterance may be originating by a person that has not been shown to the system before, and the system is able to include the new voice samples, i.e. update the dataset, then an 'open set' of speakers is found [11].

Problem Statement

The conventional voice recognition methods extract the short-time acoustic features from the voice signal and Gaussian mixture models or neural network models are trained to estimate the distribution of the feature vectors in higher dimensional space [2]. The advantage of such modeling is its simplicity and fastness, but it is still making a weak recognition score. Eigen problem is then employed to capture the higher level knowledge present in the speech. Eigenvoices enhanced the performance of the voice recognition systems but the recognition score does not improved well [3].

We thought that the improvement of the eigenproblem ability for voice recognition needs to find a feature mapping such that the distance measure between any two voices is preserved as much as possible. In turn, the use of accurate features make the computations are complex and the slot time is longer. The late of decision making is an actual problem should be faced in the field of interest. In this paper, we address this challenge by indexing the voice signal in the pitch space and then mapping the pitch documents into eigen space to produce a semantic features used for the purpose of voice recognition.

Related Work and Contribution

Recognition of voices has obtained increasing attentions in recent years. Most studies were based

on extracting some recognizable features in the frequency domain, since they are more robust than the features of the time domain. Literatures states there are different recognizable features are adopted or suggested. In the following sections we describe the head research lines and our contribution in the field of study:

Related work

Eigenvoice based methods have been shown to be effective for fast speaker adaptation when only a small amount of adaptation data is available. In [1], the principal component analysis (PCA) employed to find the most important eigenvoices. An adapted speaker recognition model found by the kernel eigenvoice method resides in the high-dimensional kernel-induced feature space, which cannot be mapped back to an exact preimage in the input speaker supervector space. The segmental eigenvoice method in [2] has been providing rapid speaker adaptation with limited amounts of adaptation data. In this method, the speaker-vector space is clustered to several subspaces and PCA is applied to each of the resulting subspaces. A soft-clustering method is proposed in which each element in a speaker vector can be assigned to more than one cluster. Then, those elements far apart from any of the clusters are removed. The experiments showed 3.3% average improvement when only 10 utterances are used. Also, Eigenvoice (EV) speaker adaptation in [3] has been shown effective for fast speaker adaptation when the amount of adaptation data is scarce.

The maximum *a posteriori* eigen-decomposition (MAPED) is suggested in [4] to improve the eigenvoices speaker adaptation. The linear combination coefficients for eigenvector decomposition are estimated according to the MAP criterion. By incorporating the prior decomposition knowledge. It is shown MAPED is able to achieve better performance than maximum likelihood eigen-decomposition (MLED) with few adaptation data. The adaptation of covariance matrices of the hidden Markov model (HMM) is exploited in the eigenvoice framework. In [5], the conventional Maximum Likelihood Eigen-Decomposition (MLED) method (ML) criterion and suffers from the unrealistic assumption made by HMM on speech process, so alternative schemes showed more efficiency to improve the performance of the voice recognition. The proposed method in [6] is an EM procedure based on a novel maximum likelihood formulation of the estimation problem which is similar to the mathematical model underlying probabilistic principal components analysis. It is enable to extend eigenvoice/EMAP adaptation in a natural way to adapt variances as well as mean vectors. It differs from other approaches in that it does not require that speaker dependent or speaker adapted models for the training speakers be given in advance. Accordingly, the method is applied directly to large vocabulary tasks even if the training data is sparse in the sense that only a small fraction of the total number of Gaussians is observed for each training speaker [6].

Real-life speaker verification systems are often implemented using client model adaptation methods is discussed in [7], since the amount of data available for each client is often too low to consider plain maximum likelihood methods. While the *Bayesian maximum A Posteriori*

(MAP) adaptation method is commonly used in speaker verification, other methods have proven to be successful in related domains. An experimental comparison between three well-known adaptation methods, namely MAP, *maximum likelihood linear regression*, and *eigenvoices* is reported. All three methods are compared to the more classical *maximum likelihood* method, and results are given for a subset of the *1999 NIST speaker recognition evaluation* database.

Contribution

In this paper, we propose a new framework to discover the relational rules for the voice recognition based on eigen problem using short time temporal feature "pitch". This framework consists of a formal model definition and a three stages algorithm for recognition. We modify the classical way of eigen problem (that be requisite input whole the speech signal into eigen computation) by input just the pitch period, so that the created recognition system becomes feasible (fast and accurate) for recognizing a query voice from huge dataset. We also, propose a suitable similarity measure incorporating the quantitative meaningful features (eigen value) with the qualitative features (eigen vectors) to get a potential recognition decision.

Eigen Theory

The mathematics of the eigenproblem are relatively recent but the real-world knowledge of the concepts have been known for a lot longer. In the following the achievements history and mathematical representation of eigenproblem are explained:

Idea Behind Eigen Problem

For old age, the idea of talking to one another using a string stretched tightly between two cans has been tried by children. The vibrations in the voice will transfer to the string. Basically a string that you strike will vibrate and that will produce tones and overtones. The oscillating shape that a specific vibrating string takes and the associated pitch of the tone are characteristic (eigen in German) to the string. These depend on the length and cross-section of the string, on the material of which the string is composed and on the tension on the string. Hence the *natural* shape of the vibrations and the *natural* pitch of the tones are termed natural or eigen modes, or eigenvibrations, and natural or eigen frequencies, respectively [7]. The string (or any natural object) has its own system of eigenvibrations and eigenfrequencies. In mathematics these are referred to as eigenfunctions and eigenvalues. The eigenfunctions associated with strings are simple sine called harmonic oscillations (while it becomes a composition of sines at different phases for human vocal tracts). If we consider the displacements of the string at discrete points then we have a vector of values; the eigenvector [8] [9]. There are many different numerical methods to estimate the eigenvector and eigenvalue, the most reliable and robust one is the power method [10] that used in the present work.

Mathematical Representation

If $[A]$ is a $n \times n$ matrix, then $[X] \neq 0$ is an eigenvector of $[A]$ if $[A][X] = \lambda[X]$, this is eigen equation where λ is a scalar and $[X] \neq 0$. The scalar λ is called the eigenvalue of $[A]$and $[X]$ is called the eigenvector corresponding to the eigenvalue λ, to find the eigenvalues of a $n \times n$ matrix

[*A*], we can adapt the eigen equation to be as follows.

$[A][X] = \lambda [X]$

$[A][X] - \lambda [X] = 0$

$[A][X] - [I][X] = 0 \quad \ldots (1)$

$([A]-[\lambda][I])[X] = 0$

Now for the above set of equations to have a nonzero solution, i.e.,

$$det([A]-[I])=0 \quad \ldots (2)$$

This left hand side can be expanded to give a polynomial in λ solving the above equation would give us values of the eigenvalues. The above equation is called the characteristic equation of [*A*]. For a [*A*] *n*×*n* matrix, the characteristic polynomial of *A* is of degree *n* (*n* roots) given by eq.(2). It should for [*A*] to identify all the following eigen theorems:

$$\lambda^n+c_1\lambda^{n-1}+c_2\lambda^{n-2}+\ldots+c_n=0 \quad \ldots (3)$$

4.3-Eigen Theorems

1. If [*A*] is a *n* × *n* triangular matrix – upper triangular, lower triangular or diagonal, the eigenvalues of [*A*] are the diagonal entries of [*A*] .

2. if [*A*] is a singular (noninvertible) matrix then λ= 0 is an eigenvalue of [*A*].

3. Both [*A*] and $[A]^T$ have the same eigenvalues.

4. Eigenvalues of a symmetric matrix are real.

5. Eigenvectors of symmetric matrix are orthogonal, only for distinct eigenvalues.

6. $|\det(A)|$ is the product of the absolute values of the eigenvalues of [*A*] .

Proposed Approach

The concept of multi-stage query processing and metric index structures have been used to model the proposed approach. We claim that these stages can beneficially be combined and that, through the combination, a significant speed up and efficient voice recognition system can be achieved. The proposed voice recognition system operated in two phases: enrollment and recognition as Fig (1) shows.

The enrollment is an offline phase in which the voice features of all the dataset are collected in a codebook file for purpose of recognition. Whereas the online recognition goes to compare the recognition features of the query voice with that stored in the codebook, then it makes a recognition decision based on the comparison. In the following, we demonstrate the preprocessing stage that aims to priory preparing the voice signal, and indexing stage that make to archive the voice feature, the last stage is mapping the archive into semantic features for purpose of features comparison based recognition. The following subsections explain the sequenced stages of the proposed approach.

Preprocessing

One of the most important characteristics of the voice signal is the regularity of its general shape, this includes the pitch of the signal. The pitch is related to the tonic pattern of the voice, it is useful for voice recognition. Pitch is the periodic portion of the signal that repeats itself with same shape and different amplitude. Fig (2) shows a voice signal of a spoken utterance and its pitch period of such voiced utterance. Because the noise is usually mixing with the signal in most the frequency bands, the voice samples should pass through a multi-bands filter before extracting the pitch period. Such that, the preprocessing stage consists of two steps; filtering and pitch

extraction, both are explained with details in the following subsections:

Step 1: **Voice filtering**, the noise is naturally found in the speech signal due to the digitization process or external effects such as another voice source. Such noise can potentially deviate the recognition precision. Two filters are used to discard the noise from the signal, they are;

i. Zero-crossing filter: which responsible on discarding the noise existing in the high frequency band. Let F_i represents the successive frames of the voice signal, then F_i is noisy and should be discarded if the number of transitions (N) across the offset ($offset$=127) is greater than 0.3 of the frame size (w=100 sample), otherwise F_i is useful signal. The mathematical representation of such filter are given as follows:

$$F_i = \begin{cases} noise & \text{if } N \geq 0.3 \times w \\ \text{Signal} & \text{otherwise} \end{cases} \quad ...(5)$$

Where,

$$N = \sum_{i=0}^{w-1} T_i \quad ...(6)$$

$$T_i = \begin{cases} 1 & if\ F_i > Offset\ and\ F_{i+1} < Offset \\ & or\ F_i < Offset\ and\ F_{i+1} > Offset \\ 0 & Otherwise \end{cases} .....(7)$$

ii. Average power filter: The main function of this filter is to discard the unvoiced regions existing in the voice since the signal in such region has low amplitude, which make the noise almost dominants on the behavior of the frame. This is achieved by computing the average power ($\overline{F}$) of current frame (F_i), and then check if $\overline{F}$ less than a threshold (T_r) then F_i is noisy and should be discarded, otherwise Fi is accepted as useful signal. The mathematical relationships given as:

$$F_i \Rightarrow \begin{cases} unvoiced & \text{if } \overline{F} < T_r \\ voiced & \text{Otherwis} \end{cases} \quad ... (8)$$

$$\overline{F} = \frac{1}{w} \sum_{i=x}^{x+w-1} A_i \quad ... (9)$$

Step 2: **Pitch extraction**

The pitch is extracted using a new proposed method, which consists of following five steps;

i. Extract the fluctuations (which are the difference between the current value and next one in the speech signal), store them in a vector V.
ii. Detect the max positive peak in the V, and determine the position (S_1) of its first peak value.
iii. Detect the second max positive peak, and determine the position (S_2) of its first peak value.
iv. Store the values in between the positions S_1 and S_2 in a vector P as the pitch period.
v. Up/down sampling the vector P to make its number of elements equal to m, this is done by adding an average between any two values to complete its length (i.e. $L=|S_2-S_1|$ becomes m).

Indexing and Mapping

The eigenproblem uses the numerical linear algebra as a basis for information retrieval. Thus, we employed eigenproblem as a tool to index the pitch period in order to solve the voice recognition problem. The procedure of the indexing is carried out by coding the pitch period as a sequence of samples, see Fig (3). Then, the coded pitch can be understood as a normalized vector of an m-dimensional space, where m denotes the number of pitch values

(attributes). Let the symbol B denotes $m \times n$ term-document matrix related to m term (pitch keywords) in n documents (voice samples). So that, the (i,j) element of the term-document matrix A represent the pitch value of i^{th} term in the j^{th} voice sample

The eigenproblem based features mapping involves the eigenvalue and eigenvector of A, which still is very memory and time consuming operation. The use of just the pitch instead of the whole voice signal in the indexing leads to reduce the computation time and needed memory size, such that the operation being faster even at large data collection. The mapping of the non-square matrix B by the eigenproblem of B^TB (where T refers to the transpose superscript) using power method can be obtained very effectively. The indexed pitch features in the B are now mapped into new semantic ones represented by eigenvalue and eigenvector for each pitch sample as Fig (4) shows.

Data Enrollment and Recognition

The dataset is of closed set data type, the enrollment phase includes collect the semantic features of whole the voices samples in a database file called codebook. The eigenvalues are stored as the first column in the codebook, while the eigenvector are stored after that in m-columns. Since the set of points (eigenvector) in the recognition is more descriptive than the single point (eigenvalue), such that both eigenvalue and eigenvector are employed together in the voice recognition task. Hence, the recognition phase including a quantitative and qualitative comparison for the semantic features between the query voice and the voice samples stored in the codebook as shown in Fig (5).

In order to increase the precision of making the recognition decision, variant weights (w_1 and w_2) are assigned to the contribution of both eigenvalue and eigenvector to issue the recognition decision. Decision making is based on the percent (C_j) of the j^{th} similarity measure (S_j) as follows:

$$S_j = w_1\left|\lambda^q - \lambda_j^c\right| + w_2\left|\frac{1}{R}\sum_{i=1}^{R} X_i^q - X_i^c\right| \quad \ldots(10)$$

$$C_j = \frac{S_j}{\sum_{j=1}^{10} S_j} \quad \ldots(11)$$

Where, the superscript q and c refer to the query and codebook references, S_j is the j^{th} similarity measure between the query voice and j^{th} voice in the codebook, λ^q and λ^c are the first recognition semantic features (eigenvalues) of the query and codebook voices sequentially, while X^c and X^c are the second recognition semantic features (eigenvectors) of the query and codebook voices.

Experiment and Evaluation

To test the general applicability of the proposed approach, the voice recognition system was built to distinguish the query voice in between dataset. All processing were implemented in visual basic 6.0 and the experiments were run on a computer with Dell 1.7 GHz processor and 2 GB main memory under windows operating system. The recognition system was implemented as described in section (5). In the following an explanation about input/output details and the evaluation of the proposed system:

Input voice dataset

The voices dataset are recorded by media player under windows operating system with an

attributes of 22 *kHz* sampling rate, mono, and 8 *bit* resolution. The recorded dataset contains 50 speech wave files taken by 10 speakers; male, female, and children, they speaking the same utterance 5 times. The recording time of each speech (wave) file was about 2 *second*, the collection of the files required about 2 *MB* of RAM, i.e. 41 *kB* for each.

Output results

The extracted pitch was found taking different shapes when different speakers are pronounce same utterance, whereas the shape of the pitch was approachly remaining similar whenever the same speaker says same utterance at different records. Fig (6) shows the shapes of the pitch for two sample voices at two different records.

It is found that L=200 *samples* length of the pitch period is exhibit both male and female voices. Such that if the length of the pitch belong to a specific voice is greater than 200 *sample* then the pitch should be down sampled to be 200 *sample* length, also when the pitch is less than 200 *sample* then the pitch should be up sampled to make its length 200 *sample*. It is found the rank value (R=5) of a document matrix in the eigen computations showed good results for eigen voice computations. Fig (7) shows the recognition scores of the proposed system versus voice samples at which w_1=0.3 and w_2=0.7, it is noticed that the average estimation of the recognition score reaches to about 81% for whole the ten speakers contributed in the dataset.Power method is one of the effective methods used to find the largest eigenvalue in an absolute sense. This eigenvalue needs to be distinct, such that a scaling step will be included in the eigenvalues computations. Also, a tolerance error will govern the stop condition of finding the optimal eigenvalues. The algorithm of the power method is given as follows:

1. Assume a guess $[X^{(0)}]$ for the eigenvector in $[A][X] = \lambda[X]$ equation. One of the entries of $[X^{(0)}]$ needs to be unity.
2. Find $[Y^{(1)}] = [A][X^{(0)}]$
3. Scale $[Y^{(1)}]$ so that the chosen unity component remains unity. $[Y^{(1)}] = \lambda^{(1)}[X^{(1)}]$
4. Repeat steps (2) and (3) with $[X] = [X^{(1)}]$ to get $[X^{(2)}]$.
5. Repeat the steps 2 and 3 until the value of the eigenvalue converges.

If E_s is the pre-specified percentage relative error tolerance to which you would like the answer to converge to, keep iterating until

$$\left|\frac{\lambda^{t+1}-\lambda^{t}}{\lambda^{t+1}}\right|\times 100 \le E_s \qquad \ldots (4)$$

Where the left hand side of the above inequality is the definition of absolute percentage relative approximate error, denoted generally by E_s A pre-specified percentage relative tolerance of $0.5\times10^{2-m}$ implies at least m significant digits are current in your answer. When the system converges, the value of λ is the largest (in absolute value) eigenvalue of $[A]$.

Evaluation

In the current computer implementation of eigen problem based indexing and mapping, a careful computation treatment was found needed especially with that related to the variables precision and the normalization of eigen values. Thereby, the numerical experiments indicate quite optimistic availability of the proposed algorithm for automated voice recognition system.

It was shown the recognition results vary with the amount of the dependency of the recognition parameters, where the experiment showed the recognition results was greatly depending on eigen vector than eigen value, and the recognition score was relatively increased by increasing the dependency on eigen vector. This refers to that the descriptive information carried by eigen vector is clearly shown and comparable than that of eigen value. The reason behind that is because the eigen vector is a set of points which can pictures more detailed information through its behavior. In correspondence, same detailed information were embedded in the eigen value, which cannot be shown in the recognition. As a result, one can see the eigen vectors of different voices take different shapes as Fig (8) shows, and it were approachly similar for same voices.

It is noticeable that the sum of the overall differences between the eigen vectors belong to different voices was large and vice versa. This indicates the similarity between the voices. Since the similarity measure used to distinguish between two voices was descriptively simple (just MSE), the good results referred to high robustness of the eigen vector to recognize the voice samples. In comparison, the eigen values showed an acceptable results less than that of eigen vectors, because of the eigen value is an individual value of limited variety. In spit of the eigen value carried the same information related to the meant voice, it cannot describe the difference between different voice samples well.

It seen that the shapes of eigen vectors for different voices were distinguishable since the difference between them are relatively great, which give the ability for the eigen vector to recognize the voices. It found the shapes of eigen vector differ with the sex of speaker. The eigen vectors of male speaker was absolutely different than that of female speakers. The distinguishing between the female and child voices was found decreasing comparing with that of male voices, but it is still giving respect recognition score. It thought the shape of the eigen vector varying with the tone variety of the voice, which in turn varying with the voice sharpness (i.e. vary with sex of speaker), which of course refers to the variety with the frequency of the voice.

Conclusions and Further Work

The ability of the eigen vector to recognize different voices with acceptable precision score indicates the efficiency of the eigen problem to perform the voice recognition task. The successful application of both the indexing and mapping of voice features ensures that the problem of voice recognition exhibits the eigen problem. For further work, one can use the singular value decomposition as a semantic indexing method to improve the results of the present research. The most interest establishment is that the stable behavior of the semantic features leads to make robust recognition decision.

References

[1]B. Yegnanarayana and S. P. Kishore, “AANN: An alternative to GMM for pattern recognition,” Neural Networks, vol. 15, no. 3, pp. 459–469, Apr. 2002.

[2]B. S. Atal, “Automatic speaker recognition basedon pitch contours,” J. Acoust. Soc. Amer., vol. 52,no. 6, pp. 1687–1697, 1972.

[3] F.Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in Proc. IEEE Int. Conf. Acoust.,Speech, and Signal Processing, Orlando, Florida,USA, May 2002, vol. 1, pp. 141–144.

[4] D. A. Reynolds and R. C. Rose, "Robust text in dependent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Automatic

[5]-Anil K. Jain, Fellow, IEEE, Arun Ross, Member, IEEE, and Salil Prabhakar, Member, IEEE Invited Paper ,"An Introduction to Biometric Recognition"

[6] J. Mariethoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *Proc.of ICSLP*, 2002, pp. 581–584.

[7] J. T. Kwok, B. Mak, and S. Ho, "Eigenvoice speaker adaptation via composite kernel PCA," in *NIPS 16*, S. Thrun, L. Saul, and B. Sch¨olkopf, Eds. MIT Press, Cambridge, MA, 2004..

[8]." A New Eigenvoice Approach To Speaker Adaptation" Chih-Hsien Huanga, Jen-Tzung Chiena and Hsin-min Wangb a Department of Computer Science and Information Engineering, Cheng Kung University, Tainanb Institute of Information Science, Academia Sinica, Taipei E-mail: acheron@chien.csie.ncku.edu.tw, jtchien@mail.ncku.edu.tw, whm@iis.sinica.edu.tw,2003

[9]B. Kolman and D. Hill, Introductory Linear Algebra with Applications, 2nd edition,Prentice Hall, 2001.

[10]"Eigenproblems in Pattern Recognition " Tijl De Bie, Nello Cristianini, and Roman Rosipal,2001

[11] Sadaoki Furui NTT Human Interface Laboratories, Tokyo, Japan , "Speaker Recognition"2002

Review

1-" Kernel Eigenvoice Speaker Adaptation'Brian Mak, Member, IEEE, James T. Kwok, and Simon Ho

*2- "Improvement of Eigenvoice-Based Speaker Adaptation by Parameter Space Clustering",Shutaro Tanji*1, *Koichi Shinoda*1, *Sadaoki Furui*1, *and Antonio Ortega*2

3-"A Comparative Study of Two Kernel Eigenspace-based Speaker Adaptation Pitch extractionMethods on Large Vocabulary Continuous Speech Recognition"Roger Hsiao and Brian Mak

4-"A New Eigenvoice Approach To Speaker Adaptation"*Chih-Hsien Huanga, Jen-Tzung Chiena and Hsin-min Wangba* Department of Computer Science and Information Engineering, Cheng Kung University, Tainan Institute of Information Science, Academia Sinica, Taipei

5-"Discriminative Speaker Adaptation with Eigenvoices" *Jun Luo, Zhijian Ou, Zuoying Wang* Department of Electronic Engineering Tsinghua University, Beijing, China

6-"Maximum Likelihood Estimation Of Eigenvoices And Residual Variances For Large Vocabulary Speech Recognition Tasks" *P. Kenny, G. Boulianne And P. Dumouchel* Centre de recherche informatique de Montr´eal (CRIM)

7-"A Comparative Study of Adaptation Methods for Speaker Verification"

Johnny Mari´ethoz Samy Bengio October 21, 2002

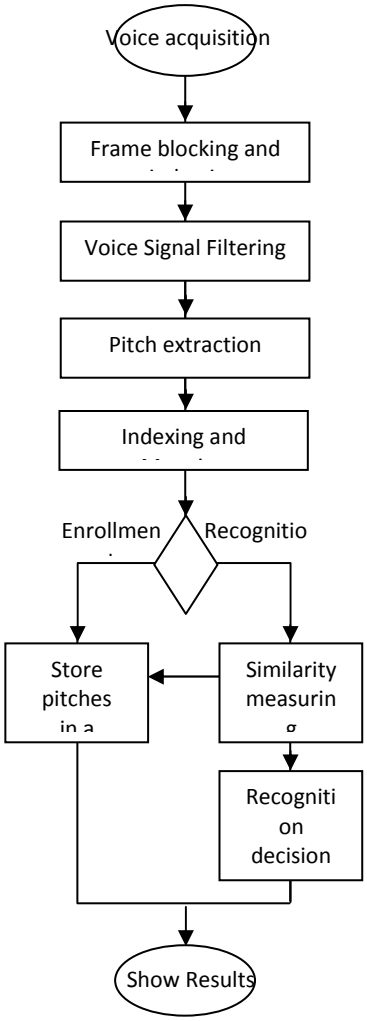
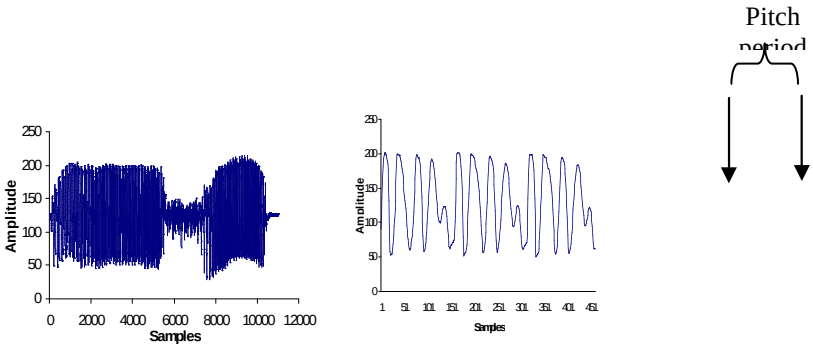


Fig (1) Sequential stages of the proposed voice recognition system.



(a) Voice signal **(b) Pitch period**

Figure (2) The voice signal and pitch period of voiced utterance.

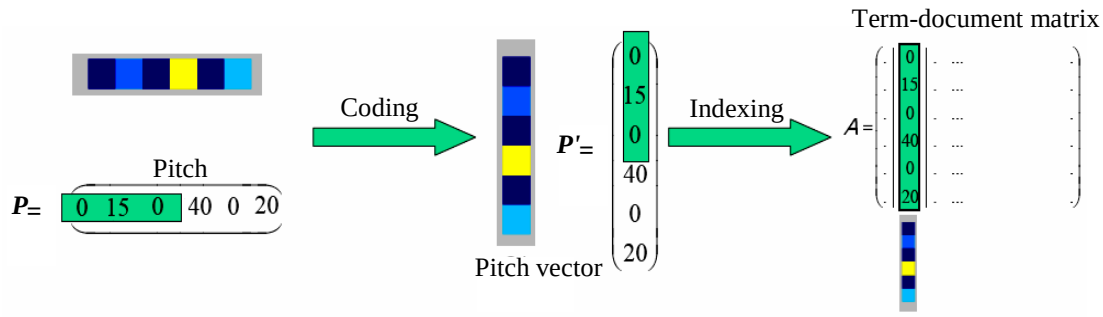


Figure (3) Indexing of the pitch vector into term-document matrix.

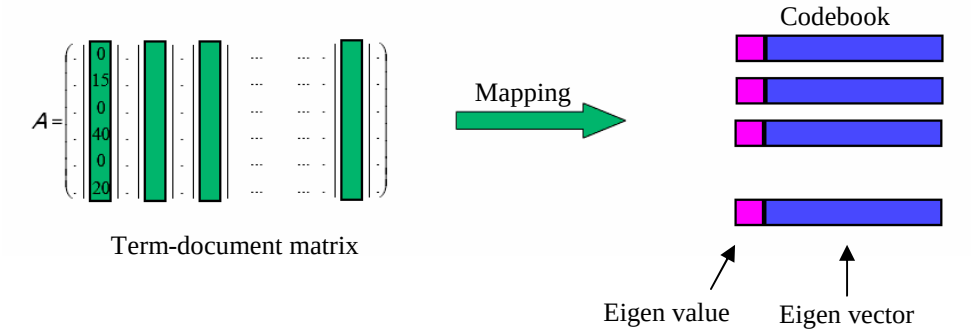


Figure (4) Mapping the term-document into codebook.

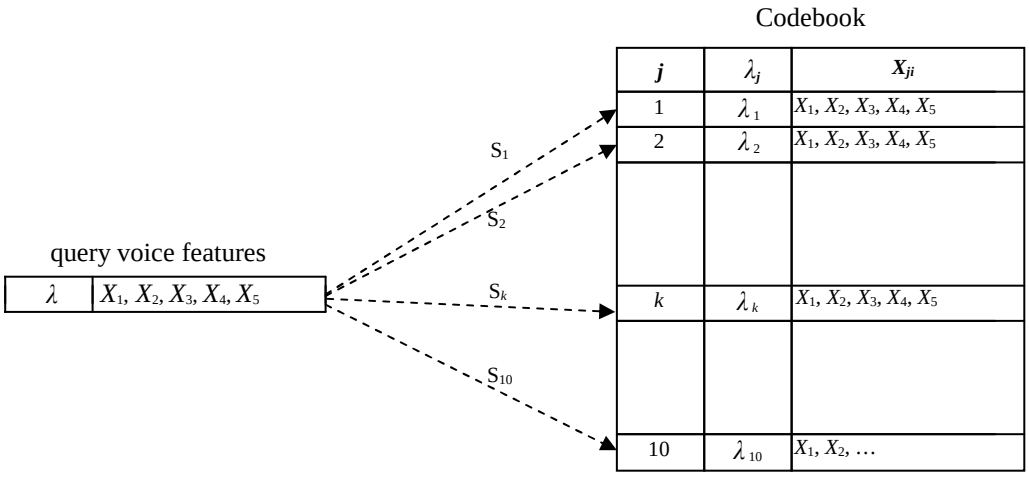


Figure (5) Similarity measuring between the query and codebook voices

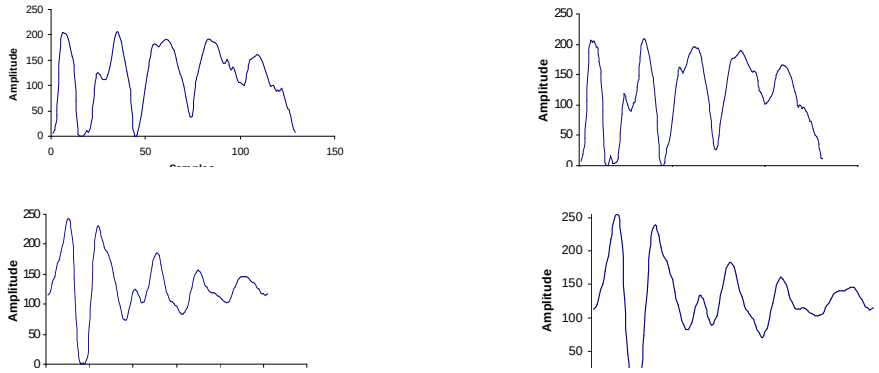


Figure (6) the pitch period of two different voices

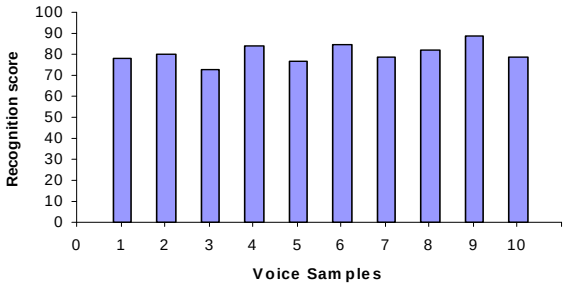


Figure (7) The recognition score of the ten tested voices.

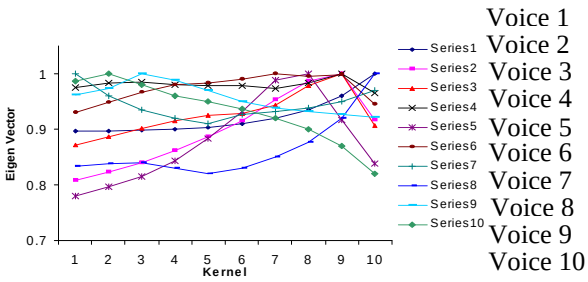


Figure (8) Normalized eigen vectors of the ten tested voices.