Formants of Speech for Biometric Identification System

Binsu C. Kovoor
Assistant Professor, Information Technology Department, School of Engineering, Cochin University of Science and Technology, CUSAT (P.O.) Kochi, India, PIN 682022
Mobile no. 0091 9847788551

ABSTRACT

In hands free security system, speech is the widely used bio-feature. Extracted formant frequencies of human speech are utilized for the identification of a person. In this study, a prototype system was developed and tested. The advantage of the system is that it can be used without the knowledge of the end users. The system was developed in four modules. In the first module the speech samples are recorded. The formant frequencies are extracted from the speech samples and recorded in a knowledge base in the second module. The autoregressive (AR) models are built in the third module. The testing is carried out in the fourth module. The influence of the envelop threshold and orders of AR model is examined to identify the optimum performance of the system.

Key words: Biometrics, speech, sound, auto-regression, formants.

Corresponding Author: Binsu C. Kovoor

INTRODUCTION

Biometrics as used in security and access control applications means measurable physical characteristics of a person that can be checked on an automated basis. Biometrics security system based on voice data is an emerging reliable method and is getting highly popularized. Speaker recognition, or voice recognition is the task of recognizing people from their voices and such systems extract features from speech, model them and use them to recognize the person from his/her voice. The human speech contains numerous discriminative features that can be used to identify the speakers. While authentication of a speaker using text dependent data is available in the published literature, the authentication based on text independent voice data available is very limited.

Broadly speaking, attempts at Automated Speech Recognition fall into two categories - a knowledge-based approach, in which knowledge about speech from the domains of linguistics and phonetics is used to construct a set of rules which is in turn used to interpret the acoustic input signal, and a ‘pattern -matching’ approach in which a priori knowledge about speech is largely ignored and techniques of pattern classification are applied to the input signal [1]. The
knowledge-based approach is able to exploit a body of knowledge about speech and, in particular, the relationship between features extracted from the speech signal and higher level linguistic representations (e.g. phonemes, syllables). However, this relationship, which is very complex, is still far from being understood, largely because of the enormous variability of signals interpreted by the brain as representing the same linguistic units. Of the knowledge that is available, it is not yet clear how best to represent or use it in a computational framework.

The pattern-matching approach makes no attempt to use this kind of knowledge, but gathers its 'knowledge' of the speech signal in a statistical form using examples of speech patterns. Powerful mathematical techniques are available which are guaranteed to optimize the technique and these ensure that the approach is surprisingly successful, even though almost all knowledge of speech production, perception and the speech signal is ignored. Furthermore, these techniques are applicable to patterns at any level and hence can be used to optimally decode other representations of speech signals (e.g. phonetic segments, words), thus providing a coherent framework for speech recognition and understanding.

A speaker identification biometric security system based on pattern matching approach has been implemented in MATLAB in which the speaker identification is carried out using formant analysis of text independent data. The influence of the parameters namely, envelope threshold and order of autoregressive (AR) model has been studied.

**PATTERN CLASSIFICATION**

**Training Phase**

**Recognition Phase**

Fig 1: Speaker Identification Process
The process of speaker identification can be divided into two main phases, namely, the training phase or enrolment phase and the testing phase or identification phase. Fig. 1 shows a block diagram of a pattern classifier in training and recognition modes. During the speaker enrolment phase or the training mode, speech samples that contain the discriminating features are collected from the speakers, and they are used to train the model. When performing recognition of isolated words, each different word is regarded as a class and many examples of each class are used to build a model for the class, and these models are subsequently stored. Thus at the end of the training phase, a speaker database is created containing the name and enrolled features of the intended users. In recognition mode, a pattern of unknown class is compared with each model and classified according to the model to which it is 'closest'.

**FEATURE SELECTION**

Feature selection is of great importance in speech recognition, as accuracy is highly dependent on the type and number of features used. Speaker identification has been done successfully using formant frequencies in the human speech. This technique consists of extracting a small number of representative features as an efficient means of characterizing the speaker-specific features.

Formants are those frequency ranges which emerge from the mouth and nose with the greatest relative amplitude. Formants may be recognized as the resonant frequencies of the vocal tract. For all voiced sounds including vowels, it is usually sufficient to look at the two lowest frequency formants to recognize the phoneme. Those formants are labeled F1 and F2. A vowel is a relatively long-lasting and unchanging part of the sound signal. Vowels almost always carry the greatest energy in the speech signal, because during vowel phonation the vocal tract in the human being is open to its maximum possible range. The vowels are probably the easiest speech segment to recognize in a spectrogram. The vowels are characterized using different set of formants in the spectrogram. In this paper, vowel recognition has been implemented through the analysis of formants. Vowels will almost always have four or more distinguishable formants; sometimes there are more than six. Vowels and its corresponding formant regions are shown in Table 1.

| Vowel | Main formant region          |
|------|-----------------------------|
| u    | 200 to 400 Hz               |
| o    | 400 to 600 Hz               |
| a    | 800 to 1200 Hz              |
| e    | 400 to 600 and 2200 to 2600 Hz |
| i    | 200 to 400 and 3000 to 3500 Hz |
VARIABLES OF THE STUDY

Envelope Threshold

The voice signal is analyzed by breaking up the word or series of words into syllables [2]. Though the words have consonants, the vowel sounds make up the majority of the syllables. After smoothing, there is a clear definition of the peaks. The goal is to choose a correct threshold amount to differentiate the numerous peaks while at the same time being able to keep the peaks for each and every signal. Because the signal is normalized prior to being run through the enveloping functions, that threshold will not change for different input volumes. All of the vowel sounds are preserved by the signal and most of what is cut off is a consonant.

Order of autoregressive (AR) system

The autoregressive model is one of a group of linear prediction formulae that predict an output of a system based on the previous outputs and inputs. In order to extract features from the speech signal, it is necessary to construct a reasonable model of human speech production system. Modeling process is usually divided into two parts: the excitation modeling and the vocal tract modeling [3]. Autoregressive (AR) model is used to determine the characteristics of the vocal system as in Eq. 1. An approach based on the vocal tract model given in the literature [3][4] has been implemented in this study.

\[
s(n) = b_0 u(n) + \sum_{i=1}^{i=p} a_i s(n-1) \tag{1}
\]

where \( u(n) \) is a normalized excitation signal, \( b_0 \) is the gain of the excitation signal and the coefficients \( a_1, a_2, \ldots, a_i \) are the weights for previous sound samples \( s(n-1) \).

The model evaluate the formants or resonant frequencies of the vocal system. (i.e. the peaks in the frequency response). One conjugate pole pair is required to produce each formant and one formant is expected in each 1 kHz band or so. The order of AR model is given by Eq. 2.

\[
\text{Order of AR model} = \left\{ \begin{array}{ll} \text{sampling frequency/2} \\ + \text{‘adjustment factor’} \end{array} \right. \tag{2}
\]

Hence, by varying the order of AR system, the sampling frequency in the system is varied. The adjustment factor is fixed arbitrarily.

IMPLEMENTATION

Details of Data Set

In this study, the utterance of 5 vowels namely, a,e,i,o,u, has been recorded from six different users. Each vowel has been recorded five times and used for the training phase and the formants corresponding to each of the vowels have been extracted. The names of speakers and average values of formants extracted from the vowel sound have been stored in the database, thus creating a model database.
The words selected for classification purpose are dog, cat, meet, hello and dish. The speaker has been asked to read the selected words and the sound signals have been recorded in separate files. The formants of the vowel in each of these words have been extracted and compared to the corresponding data in the speaker database.

**Working Principle**

The identification of the speaker from the sound signal involves several steps. The formants are extracted from the sound signals [5]. The formants of each of the vowels are distinct and the magnitude lies between different ranges (Table 1). The extracted formant values are compared with the ranges and vowel in the sound signal is identified. The formants extracted from the input signal are then compared with the formants of the vowel stored in the database. The magnitude of the difference between the formants extracted from the input signal and the formants stored in the database is computed. The speaker in the database, who has minimum difference between the formants of input signal and the stored value of the formants, is identified. If the difference is less than the envelope threshold value, the speaker is declared as identified.

**Performance measurement parameters**

In the biometric security system for speaker identification (BiSySI), the influence of two parameters, namely, envelope threshold and order of the AR model has been studied. The envelope threshold has been varied between 0.01 and 1.0 while the order of the AR model has been varied between 0 and 20. The success score rate (SSR), which is the percentage of true enrollees permitted, false acceptance rate (FAR) which is the percentage of impostors permitted and false rejection rate (FRR) which is the percentage of true enrollees declined has been computed to evaluate the performance of BiSySI. The variation of SSR, FAR and FRR have been plotted and compared. The performance of the system, BiSySI, is reckoned to be better, if SSR assumes highest magnitude and both FAR and FRR takes lowest magnitude.

**RESULTS AND DISCUSSION**

The biometric security system implemented in MATLAB uses formants extracted from the voice signals for speaker identification. The influence of the parameters, namely, envelope threshold and order of the AR model has been studied. The SSR, FAR and FRR have been computed and compared for evaluating the system. The voice signals of five words spoken by the different speakers have been considered as the input signal to system.

**Effect of envelope threshold**

By varying the value of the envelope threshold between 0.01 and 1.0, the influence of the value on the performance of the system has been studied and is as depicted in Fig. 2.
Fig 2: Influence of the envelope threshold in BiSySI (order of AR model=15)

When the value of envelope threshold is varied between 0.05 and 0.70, the SSR is found to be between 80 and 93 percent while FAR varied between 3 and 17 percent and the FRR between 3 and 10 percent. It was also observed that the magnitude of SSR is found to be as low as 0 percent when the envelope threshold lies outside the range between 0.05 and 0.70. Hence, it is recommended that the magnitude of envelope threshold in the BiSySI shall be set between 0.05 and 0.70 to obtain favorable results. The success score rate of the proposed system based on formant analysis will be as high as 80 percent or better, when envelope threshold is set between 0.05 and 0.70.

Effect of order of AR model

Coming to the influence of the order of AR model, the SSR was found to be between 80 and 93 percent and the FAR is found to be between 3 and 13 percent for an order variation between 9 and 15. The FRR was also found to be between 3 and 13 percent. The influence of order of AR model in BiSySI is as given in Fig. 3. The favorable limit of the magnitude of order of AR model is found to be between 9 and 15.
System performance based on selected words

The average of SSR, FAR and FRR has been computed corresponding to each of the selected word, over the favorable ranges of envelope threshold and order of AR model and is as given in Fig. 4. While the lowest magnitude of SSR is found to be 80 percent, the greatest magnitude of SSR is found to be 92 percent, corresponding to the word hello. The reason for this can be interpreted as the presence of two vowels in the word hello while all the other words selected were containing only one vowel. This indicates that an increase in the magnitude of average SSR is expected when words or sentences containing multiple vowels are used as input signal.

Interaction of parameters on system performance

The interactions of parameters, namely, order of AR model and envelope threshold on SSR, FAR and FRR is given in the 3D plots shown as Fig. 5, 6 and 7 respectively. From Fig. 5, the peak is found to be when the order of AR model takes a value of 14 and envelope threshold takes a value of 0.7. From Fig. 6, it is clear that the lowest magnitude of FAR will be when the order of AR model is 11 and Fig. 7, shows that the lowest magnitude of FRR is when the
order of AR model is 14. These interaction diagrams will help to choose the appropriate value for the order of the AR model and the envelope threshold for meeting the desired requirements.

Fig 5: Influence of order of AR model and envelope threshold on SSR (data shown only in the favorable range)

Fig 6: Influence of order of AR model and envelope threshold on FAR (data shown only in the favorable range)

Fig 7: Influence of order of AR model and envelope threshold on FRAR (data shown only in the favorable range)
CONCLUSION

A biometric security system for speaker identification (BiSySI) has been developed using formant analysis. The text independent voice signals are used as the input of the BiSySI. The algorithm for formant analysis together with speech signal processing has been implemented in MATLAB. The influence of parameters such as envelope threshold and order of AR model has been studied. It is found that the most favorable results are expected when envelope threshold takes a value between 0.05 and 0.70 and order of AR model takes a value between 9 and 15. The magnitude of verification success score rate lies between 80 and 93 percent over the favorable range. The performance of BiSySI improves with inputting the voice signals of words containing more than one vowel. However, lot more studies are to be carried out to identify the uncertainties involved in the speaker identification system. The proposed BiSySI can be used for the applications such as voice dialing, telephone shopping, database information services, voice mail, security control and remote access.

REFERENCE

[1] R. Lawrence and J. Bing-Hwang, Fundamentals of Speech Recognition, New Jersey: Prentice Hall, 1993.

[2] L.R. Rabiner and R.W. Schafer, Digital processing of Speech Signals, India, Pearson Education, 1990.

[3] J. P. Campbell, Speaker Recognition: a Tutorial, Proc. of the IEEE, 85, 1437-1462(1997).

[4] George Tzanetakis and Perry Cook, Musical Genre Classification of Audio Signals, IEEE Transactions On Speech And Audio processing, 10( 5), 2002, pp. 293-302.

[5] T. Kinnunen, T. Kilpelainen and P. Franti, “Comparison of clustering algorithms in speaker identification” in SPC2000, Proc. IASTED Int. Conf. Signal Processing and Communications, Marbella, Spain, 2000, pp. 222-227.