Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 9 (8): 472-479, 2014
ISSN 1819-3579 / DOI: 10.3923/tasr.2014.472.479

Speaker Identification Using Bayesian Algorithm

Hasan Bjaili, Khalid Daqrouq and Rami Al-Hmouz
King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding Author: Hasan Bjaili, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

The majority of the computations during the process of speaker identification originate from the likelihood computations between the feature vectors of the unknown speaker and the models in the database. The identification time depends on the number of feature vectors, their dimensionality, the complexity of the speaker models and the number of speakers. The main objective of this study is to use Bayesian algorithm in speaker identification with different features extraction methods. Three methods are used to extract the essential speaker features based on Discrete Wavelet Transform and Linear Prediction Coefficient (LPC). The results showed that Bayesian algorithm gives excellent performance in case of minimum signal fluctuations. It has been shown that Bayesian classifier achieves a better recognition rate (90.93%) with the Wavelet Packet (WP) and Average Framing Linear Predication Coding (AFLPC) feature extraction method. It is also suggested to analyze the proposed system in Additive White Gaussian Noise (AWGN) and real noise environments; 58.56% for 0 dB and 70.52% for 5 dB.

Key words: Speaker identification, bayesian, biometrics, simulation

INTRODUCTION

Biometrics refers to the study of the human beings that measures human behavioral and physiological data. Behavioral biometrics are based on the way, human beings do things such as gait, signature, blinking and lip movement, etc. while physiological biometrics are based on measuring a person's physical characteristics such as finger prints, iris pattern, retina patterns, palm prints, facial features and hand geometry, etc. Speaker recognition is the one of the vital biometric recognition which is the process of recognizing a person by his/her voice. There are six different types of speaker recognition which are: Speaker verification, speaker identification, speaker classification, speaker segmentation, speaker detection and speaker tracking (Campbell, 1997).

Speaker identification process may be either text-dependent or text-independent. In the text-dependent process, the speaker trains the system by uttering specific keywords and repeats them at the recognition phase, whereas the text-independent does not depend on a specific keywords.

The main focus of this study is speaker identification (recognition) that can be classified as either physiological or behavioral depending on the concentration of the system. If the concentration is on the focal tract then it is classified as physiological and if the concentration is on the way of speaking, it is considered behavioral.

Speaker identification may split in two additional categories, closed-set and open-set. During the closed-set, the speaker utterance is compared against all speaker models in the database. The system will then return the ID of the closest match. No speaker will be rejected during the process

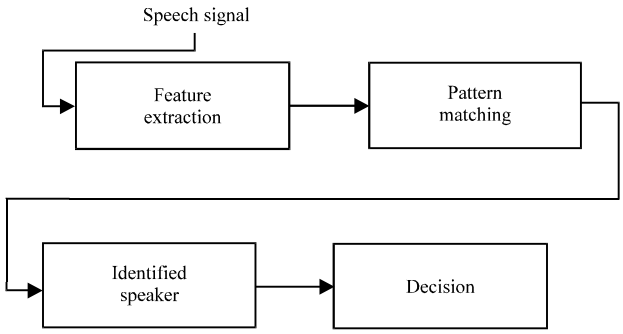

Fig. 1: Process of speaker identification

of the closed-set mode. The open-set identification is basically a closed-set identification adding to it a speaker verification capability. The closest match will then be verified and it is matched the speaker will be granted rights to the system.

In the identification task Fig. 1, the system has trained models for a certain amount of client speakers and the task is to determine which one of these models best matches the current speaker. Speaker identification could be used in adaptive user interfaces. For example, a car shared by a couple could recognize the driver by his/her voice and adjust the seat accordingly. This particular application concept belongs to the more general group of speaker adaptation methods that are already employed in speech recognition systems (Kuhn *et al*., 2000).

The preprocessing, feature extraction, speaker modeling and decision logic modules comprises the speaker identification system. In the preprocessing step, the input signal is adjusted and prepared with the appropriate characteristics for the next step. Preprocessing covers the conversion of analog speech signal into digital form, filtering, framing, endpoint detection and windowing. In the feature extraction part the input utterance is converted into a set of vector. Those vectors are used characterized the speakers' speech properties. Irrelevant features are removed in the front-end process. After, the relevant features are highlighted during the front-end process, they are used to create a speaker model. This process is called speaker modeling. Then, those models are stored in a database for future usage. Matching the unknown speaker sample with the correct speaker using the reference models database is done by the decision logic which is the last step in the speaker identification system, it is the decision will be made based on the maximum posteriori probability calculated using the unknown speaker feature vectors and the speaker reference model best selected (Farhood and Abdughafour, 2010).

Naive Bayesian classification is the simplest classification algorithms and yet effective and efficient. It is used when the dimensionality of the inputs is high. The use of Bayesian in speech processing has gained attention (Zweig and Russel, 1998; Maina and Walsh, 2011a, b; Jiang and Deng, 2001; Vogt and Sridharan, 2004; Meuwly and Drygajlo, 2001; Shiota *et al*., 2012; Jiang *et al*., 1999; Gauvain and Lee, 1996; Khanteymoori *et al*., 2008).

In (Maina and Walsh, 2011a, b) the study presented a comparison of two variations of Bayesian algorithms for joint speech enhancement and speaker identification. In both algorithms they make use of speaker dependent speech priors which allowed them to perform speech enhancement and speaker identification jointly.

The application of Bayes algorithm is factor scoring to speaker verification was presented in (Jiang and Deng, 2001). Bayesian theorem has also been implemented as a robust methodology

for forensic automatic speaker recognition (Meuwly and Drygajlo, 2001). It has been shown that the approach gives an adequate solution for the interpretation of the recorded speech as evidence in the judicial process. In (Shiota *et al*., 2012) an integrating model structures based on the Bayesian framework for speech recognition has been proposed. The speech recognition experiment demonstrated the proposed method could automatically estimate reliable posterior distributions of model parameters and an adequate posterior distribution of model structures.

In this study, different feature extraction methods will be implemented, mostly based on Discrete Wavelet Transform and Linear Prediction Coefficient (LPC) techniques for text-independent speaker identification systems. The investigation procedure is based on feature extraction and voice classification. In the classification phase, Bayesian algorithm is proposed and the performance of Bayesian algorithm for speaker identification will be measured with different feature extraction methods for clear and noisy voice signals.

NAIVE BAYESIAN CLASSIFIER

Observing the feature of AFLPC methods reported in (Daqrouq and Al-Azzawi, 2012) and using these features statistics, the likelihood of each feature as begin a verified speaker can be determined.

Let the class C_i be a speaker class and let D be data that provides information about C_i, then:

$$P(C_i \mid D) = \frac{P(D \mid C_i)P(C_i)}{\sum_j P(D \mid C_j)P(C_j)}$$

The first step is to estimate $P(D \mid C_i)$ usually referred to as the likelihood function. This is achieved using training speaker-feature samples. Features are collected by applying Discrete Wavelet Transform (DWT) and Forward Discrete Wavelet Transform (FDWT) on speaker signals. Figure 2 shows histogram for three different speakers' features. It is clearly shown that using speaker features statistics don't give abundant distinctions among speaker classes. Therefore, applying GMM leads to low classification rate because speaker-feature-distributions can be modeled as single Gaussian distribution but no discrimination among classes.

Accordingly, we build likelihood function for each feature per speaker; it was found that most features can be modeled as Gaussian distribution as shown in Fig. 3.

For each feature, there would be a probability score to each speaker-class. Therefore, let f be a feature in features' vector. Then:

$$P(C_i \mid f_k) = \frac{P(f_k \mid C_i)P(C_i)}{\sum_j P(f_k \mid C_j)P(C_j)}$$

and:

$$P(C_i \mid f_1 f_2 \ldots f_N) = \frac{P(f_1 f_2 \ldots f_N \mid C_i)P(C_i)}{\sum_j P(f_1 f_2 \ldots f_N \mid C_j)P(C_j)}$$

where, N is the total number of features in AFLPC. Under the assumption of conditional independence, we reach to.

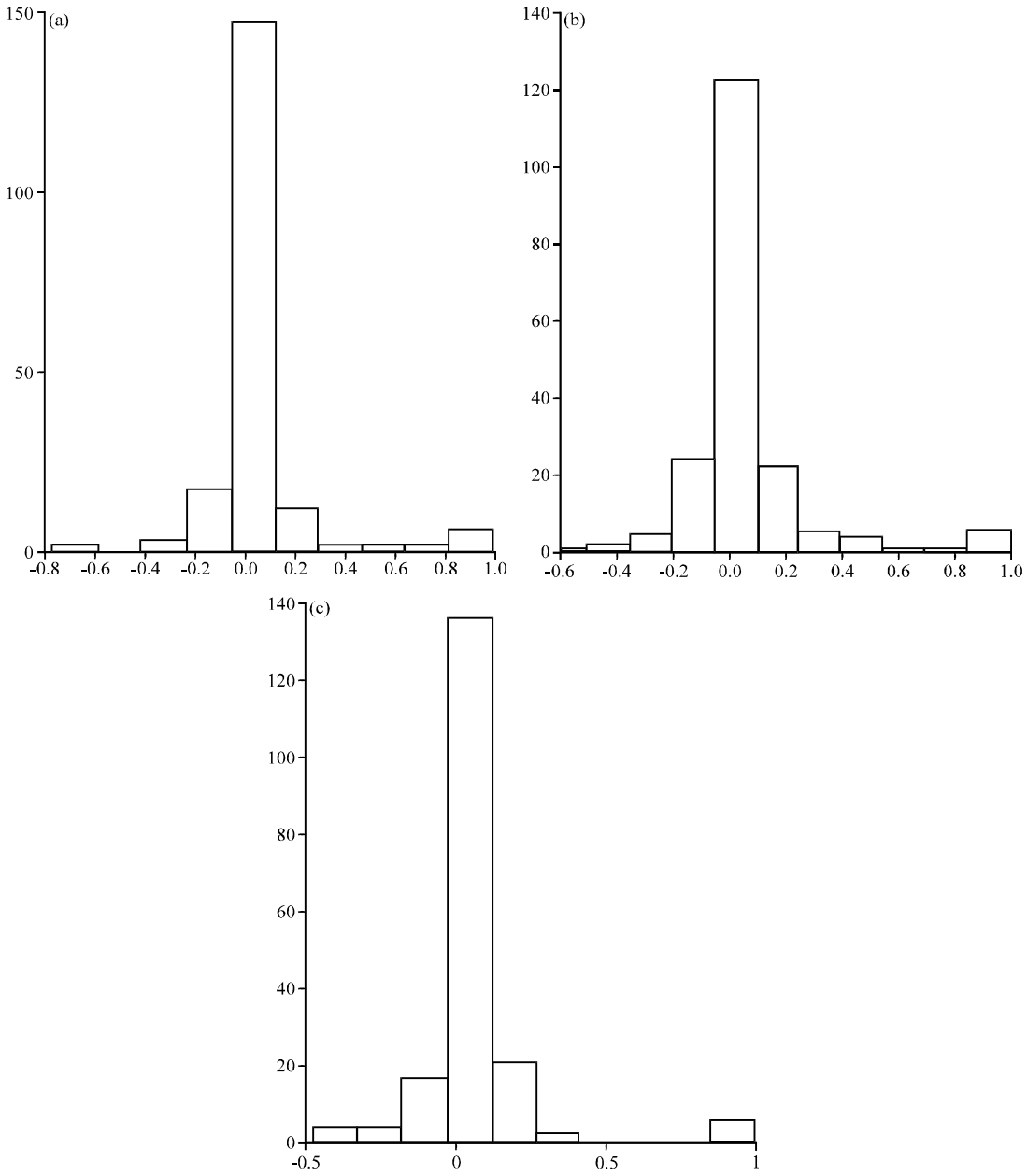

Fig. 2(a-c): Speaker-feature distributions, (a) Feature distribution of class 1, (b) Feature distribution of class 12 and (c) Feature distribution of class 32

Posteriori $P(C_i | f_1\ f_2\ ...\ f_N)$ is computed throughout Bayesian fusion based on all features' probabilities in speaker signal AFLPC. The P(C) is the prior probability about the speaker-class C_i; it was assumed that all classes are equally likely (P(C) = 1/50). $\Sigma_j\ P(f_1\ f_2\ ...\ f_N | C_j)P(C_j)$ is a normalization term. The Maximum A Posteriori probability (MAP) of Ci is used to estimate the speaker class Ci that maximizes $P(C_i | f_1\ f_2\ ...\ f_N)$:

$$Argmax_i \{P(C_i | f_1\ f_2\ ...\ f_N)\}$$

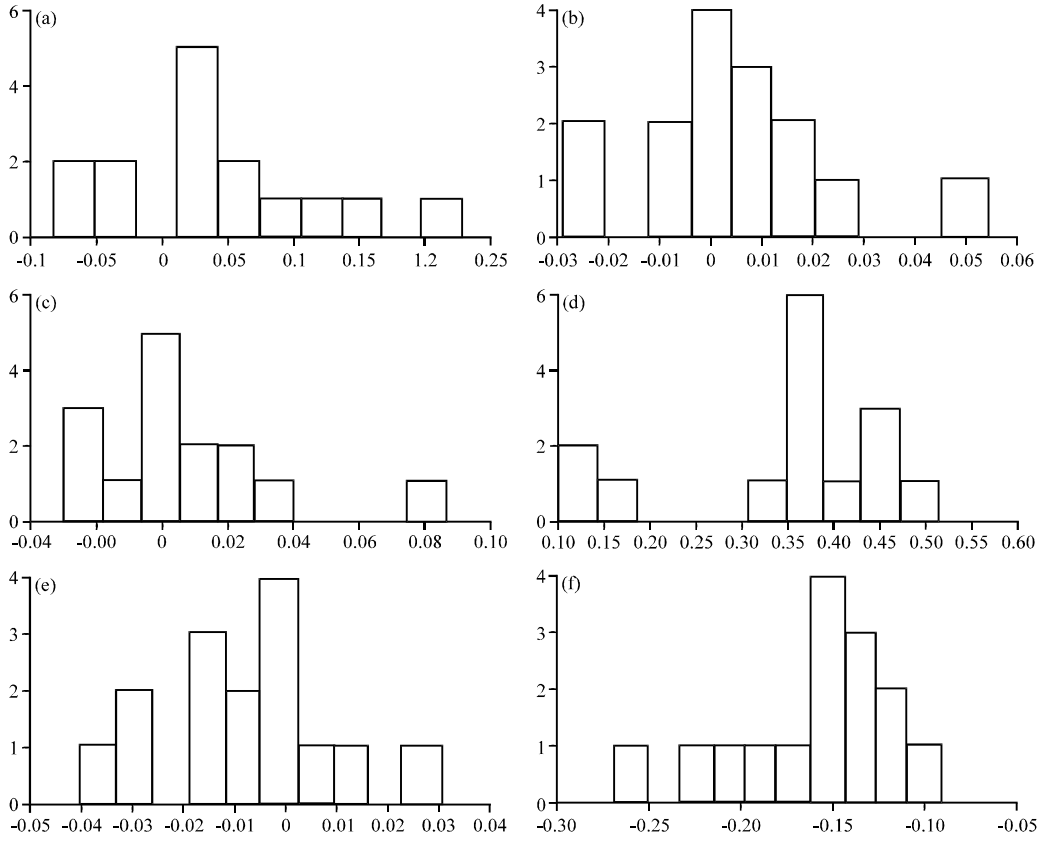

Fig. 3(a-f): Feature distributions, (a) Feature number 4 distribution of class 1, (b) Feature number 156 distribution of class 10, (c) Feature number 190 distribution of class 24, (d) Feature number 156 distribution of class 41, (e) Feature number 190 distribution of class 24 and (f) Feature number 190 distribution of class 2

Similarly, features from different approaches can be combined to produce probability-scores to each speaker. The essence of this approach is a way to combine different methods in a probabilistic manner, one method produces features that are different from other methods as well as features from different methods suffer from independent noise. As a result of the combined methods, features will be more descriptive and noise can be eliminated in the process of fusion.

RESULT AND DISCUSSION

To examine the presented text-independent speaker identification system, a testing database was created from the Arabic language. The recording environment is a normal office setting via., PC-sound card with original frequency of 4 kHz and a sampling frequency of 16 kHz. These utterances are Arabic spoken digits from 0 to 14. Each speaker also distinctly reads 30 sec worth of different Arabic texts ten separate times.

In the experiments, several feature extraction methods were analyzed to expose the efficacy of the proposed system. The following experiment investigates the proposed method in terms of the recognition rate.

Table 1: Comparison between different feature extraction methods

Feature extraction methods	Recognition rate (%)
Discrete Wavelet Transform (DWT) LPC average framing at level-5	88.10
Wavelet Packet (WP) LPC average framing at level-2- (WPLPCF)	90.93
Method 1+method 2	93.55
LPC with 30 coefficient	48.19
DWT with LPC calculated for each sub-signal with 30 coefficient	81.85
WP with LPC, 30 coefficient	84.48
Energy calculated for WP sub-signals	71.85

Table 2: Recognition rate with white Gaussian noise

	Recognition rate (%)	
Feature extraction methods	0 dB	5 dB
Discrete Wavelet Transform (DWT) LPC average framing at level-5 (DWTLPCF)	67.34	77.80
Wavelet Packet (WP) LPC average framing at level-2 (WPLPCF)	63.89	75.77
Method 1+method 2	69.67	79.10
LPC with 30 coefficient	21.53	35.20
DWT with LPC calculated for each sub-signal with 30 coefficient	81.85	63.65
WP with LPC, 30 coefficient	50.78	59.20
Energy calculated for WP sub-signals	44.56	55.12

Table 3: Recognition rate with real babble noise

	Recognition rate (%)	
Feature extraction methods	0 dB	5 dB
Discrete Wavelet Transform (DWT) LPC average framing at level-5	55.23	72.54
Wavelet Packet (WP) LPC average framing at level-2	51.23	69.13
Method 1+method 2	55.68	74.42
LPC with 30 coefficient	20.08	25.17
DWT with LPC calculated for each sub-signal with 30 coefficient	55.39	63.65
WP with LPC, 30 coefficient	39.14	50.19
Energy calculated for WP sub-signals	35.67	40.52

Table 1 shows a comparative study of different feature extraction methods and Bayesian algorithm classifier is utilized. The best recognition rate selection obtained was 90.93 for WPLPCF.

Another experiment was conducted to assess the performance of the system in noisy environments. Table 2 and 3 summarize the results of the speaker identification corresponding to White Gaussian Noise WGN and real noise (restaurant noise, which seems like babbling) with the Signal-to-Noise Ratio (SNR) of 0 and 5 dB references, respectively.

In both noisy conditions, the best Bayesian recognition rate was obtained for DWTLPCF. The reason for DWT's success over WP is that the feature vector is obtained from level 5 (depth 5), where the sub signals are filtered in lower depth than in WPLPCF at level 2. It is shown that the use of the Eigen vector in conjunction with WPLPCF can improve the robustness of an identification system.

CONCLUSION

The Bayesian algorithm in speaker identification with different features extraction methods has been successfully implemented. Seven methods are used to extract the essential speaker

features based on Discrete Wavelet Transform and Linear Prediction Coefficient (LPC). Experimental results showed both DWT and WP linked with AFLPC are suitable for the feature extraction method. However, WP resulted in better performance in terms of recognition rate. As a comparison with other methods, WPLPCF produced a higher recognition rate. The experimental results revealed the proposed AFLPC technique with DWT can accomplish better results for a speaker identification system in an AWGN environment; 67.3% for 0 dB and 77.8% for 5 dB. It can also be concluded that that Bayesian algorithm gives excellent performance in case of minimum signal fluctuations.

ACKNOWLEDGMENT

This study was funded by the Deanship of Scientific Research (DSR), king Abdulaziz university, Jeddah. The authors, therefore, acknowledge with thanks DSR technical and financial support.

REFERENCES

Campbell, Jr. J.P., 1997. Speaker recognition: A tutorial. Proc. IEEE., 85: 1437-1462.

Daqrouq, K. and K.Y. Al Azzawi, 2012. Average framing linear prediction coding with wavelet transform for text-independent speaker identification system. Comput. Electr. Eng., 38: 1467-1479.

Farhood, Z. and M. Abdulghafour, 2010. Investigation on model selection criteria for speaker identification. Proceedings of the International Symposium on Information Technology, Volume 2, June 15-17, 2010, Kuala Lumpur, Malaysia, pp: 537-541.

Gauvain, J.L. and C.H. Lee, 1996. Bayesian Adaptive Learning and Map Estimation of HMM. In: Automatic Speech and Speaker Recognition: Advanced Topics, Lee, C.H., F.K. Soong and K.K. Paliwal (Eds.). Kluwer Academic, Boston, USA., ISBN-13: 9780792397069, pp: 83-107.

Jiang, H., K. Hirose and Q. Huo, 1999. Robust speech recognition based on a Bayesian prediction approach. IEEE Trans. Speech Audio Process., 7: 426-440.

Jiang, H. and L. Deng, 2001. A Bayesian approach to the verification problem: Applications to speaker verification. IEEE Trans. Speech Audio Process., 9: 874-884.

Khanteymoori, A.R., M.M. Homayounpour and M.B. Menhaj, 2008. A bayesian network based approach for data classification using structural learning. Proceedings of the 13th International CSI Computer Conference on Advances in Computer Science and Engineering, March 9-11, 2008, Island, Iran, pp: 25-32.

Kuhn, R., J.C. Junqua, P. Nguyen and N. Niedzielski, 2000. Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process, 8: 695-707.

Maina, C.W. and J.M. Walsh, 2011a. Approximate Bayesian robust speech processing. Proceedings of the 45th Asilomar Conference on Record of the Signals, Systems and Computers, November 6-9, 2011, Pacific Grove, CA., USA., pp: 397-400.

Maina, C.W. and J.M. Walsh, 2011b. Joint speech enhancement and speaker identification using approximate Bayesian inference. IEEE Trans. Audio Speech Lang. Process., 19: 1517-1529.

Meuwly, D. and A. Drygajlo, 2001. Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). Proceedings of the Workshop on Speaker Odyssey-The Speaker Recognition, June 18-22, 2001, Crete, Greece, pp: 145-150.

Shiota, S., K. Hashimoto, Y. Nankaku and K. Tokuda, 2012. A model structure integration based on a Bayesian framework for speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, March 25-30, 2012, Kyoto, Japan, pp: 4813-4816.

Vogt, R.J. and S. Sridharan, 2004. Bayes factor scoring of GMMs for speaker verification. Proceedings of the Workshop on Odyssey: The Speaker and Language Recognition, May 31-June 3 2004, Toledo, Spain, pp: 173-176.

Zweig, G. and S. Russell, 1998. Speech recognition with dynamic Bayesian networks. Proceedings of the 15th National Conference on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conference, July 26-30, 1998, Madison, WI., USA., pp: 173-180.