Journal of Computer Science 3 (5): 304-309, 2007
ISSN 1549-3636

Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform

[1]Muzhir Shaban Al-Ani, [2]Thabit Sultan Mohammed and [3]Karim M. Aljebory
[1]Amman Arab University, P.O.Box-2234, Amman-Jordan-11953,
[2]Al-Zaytoonah University, P.O. Box- 130, Amman 11733-Jordan
[3]Al-Isra private University, P.O. Box- 22,23, Amman 11622-Jordan

Abstract: In speaker identification systems, a database is constructed from the speech samples of known speakers. The approach implemented in this paper is hybrid, where the wavelet transform and neural networks are used together to form a system with improved performance. Features are extracted by applying a discrete wavelet transform (DWT), while a neural network (NN) is used for formulating the system database and for handling the task of decision making. The neural network is trained using inputs, which are the feature vectors. A criteria depends on both false acceptance ratio (FAR) and false rejection ratio (FRR) is used to evaluate the system performance. For experimenting the proposed system, a set of 25 randomly aged male and female speakers was used. Results of admitting the members of this set to a secure system were computed and presented. The evaluation criteria parameters obtained are; FAR=14.5% and FRR=24.5%

Key words: Speaker identification, speaker recognition, discrete wavelet transform, multi-valued neural networks

INTRODUCTION

Speaker identification has been a wide and attractive area of research. Many works based on speech features, were proposed. In a speaker recognition system there are three important components; the feature extraction component, the speaker models and the matching algorithm. Feature extraction drives a set of speaker-specific vectors from the input signal. Speaker model is then generated from these vectors for each speaker. The matching procedure performs the comparison of the speaker models[1]. Some of recent works on speaker identification depend on classical features including *cepstrum* with many variants[2], sub-band processing technique[3-6], Gaussian mixture models (GMM)[7], linear prediction coding [8,9], wavelet transform[10-12] and neural networks[11-13]. In[14], an overview of several modeling techniques is given. In[11], a hybrid approach of wavelet transform and neural networks is adopted, where the sounds heard over a chest wall, not an uttered ones, are classified such that they can be used for diagnosing pulmonary diseases. This same hybrid approach together with a number of other approaches are studied in[12] and their performances are compared for phoneme recognition uttered by a single speaker.

In this study, we consider a hybrid approach, where the feature extraction component is performed using discrete wavelet transform (DWT), while the speaker modeling and speaker matching components are both performed using neural networks. Our trend is motivated by the fact that wavelet transform offers fine approximation characteristics compared with other spectral analysis techniques; such as discrete cosine transform (DCT). The possibility of introducing a selective zeroing of the coefficients is another merit of wavelet transform. With wavelets, it is possible to analyze a signal at several levels of resolution, making it possible to capture transient, high-frequency bursts with poor frequency resolution and also slowly varying characteristics with high-frequency resolution. Therefore, it is possible to trade off frequency resolution for better time resolution (for analyzing transients) and time resolution for better frequency resolution (for analyzing slow variations), a facility not afforded by the short-time Fourier transform[15]. Spoken sentences by a relatively large society of random speakers were used in this work to form the database of the system. The diversity of such society had imposed a challenge on the performance of the system, the training process, necessary input data used to train the neural networks, and the choice of features to be extracted. The selection of the features varies from application to another and it is desirable that dissimilar acoustic vectors would be clearly separable from each other (forming separate clusters). However, detailed

Corresponding Author: Muzhir Shaban Al-Ani, Amman Arab University, P.O. Box-2234, Amman, Jordan, 11953

analysis of feature vectors does not support this assumption, where it is found that the distribution of the feature vectors can be considered more or less as of a continuous probability distribution rather than a set of data clusters[16]. In accordance with this, we in our work consider a vector composed of a set of features without concentrating specifically on a certain feature.

Concepts of speaker identification systems: Speaker identification systems may be classified into two categories based on their principle of operation.
Text-dependent systems, which make use of a fixed utterance for test and training and rely on specific features of the test utterance in order to affect a match.
Text-independent systems, which make use of different utterances for test and training and rely on long-term statistical characteristics of speech for making a successful identification.

Text-dependent systems require less training than text-independent systems and are capable of producing good results with a fraction of the test speech sample required by a text-independent system. The pitch period or fundamental frequency of speech varies from one individual to another; pitch frequency is high for female voices and low for male voices. This suggests that pitch might be a suitable parameter to distinguish one speaker from another, or at least to narrow down the set of probable matches[17]. This concept of speaker identification is adopted in this paper. The analysis of the frequency spectrum of the test utterance provides valuable information about speaker identification. The spectrum contains both pitch harmonics and vocal-tract resonant peaks, making it possible to identify the speaker with a high probability of being correct. The vocal-tract filter parameters (filter coefficients) can also be used to good effect for speaker identification. This is due to the fact that different speakers have different vocal-tract configurations for the same utterance, depending on their physical and emotional conditions, as well as whether the speaker is a native or non-native speaker[9].

In any text-dependent speaker identification system, an important decision is the choice of test utterance. The source-filter model is most accurate at representing voiced sounds, such as the vowels. Vowels have a definite, consistent pitch period. The vocal-tract configuration for vowel-utterances exhibits a clear formant (resonant) structure. The frequency spectrum corresponding to vowel-utterances therefore contains a wealth of information that can be used for speaker identification. In general, it is difficult to guarantee a hundred percent recognition even with the best speaker identification approaches.

Generally speaking, two parameters may be used to describe the overall performance of a speaker-identification system.

A false acceptance: Which occurs when the system incorrectly identifies an unregistered individual as an enrolled one, or when one registered individual is mistaken for another. The FAR (False Acceptance Ratio) is the ratio of the number of false acceptances to the total number of trials. The value of FAR can be reduced by setting a strict low threshold.

A false rejection: Which occurs when the system incorrectly refuses to identify an individual who is registered with the system. The FRR (False Rejection Ratio) is the ratio of the number of false rejections to the total number of trials. Setting the threshold to a liberal high value can minimize the value of FRR. The requirements for low FAR and FRR are seen to be conflicting and both parameters cannot be simultaneously lowered. However, a low FAR is vital for good speaker identification systems and most systems are biased for good FAR performance at the expense of FRR.

Spectral analysis using wavelets: The spectral analysis tool, which were used in this work is the wavelet transform (WT).The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently [18,19]. The DWT analysis can be performed using a fast, pyramidal algorithm related to multirate filterbanks. The main process performed by this algorithm is a number of successive highpass and lowpass filtering of the time domain signal and is defined by the following equations [12, 20]:

$$Y_{high}(k)=\Sigma\ x(n)\ g(2k-n) \quad (1)$$

$$Y_{low}(k)\ =\Sigma\ x(n)\ h(2k-n) \quad (2)$$

where Y_{high} (k), Y_{low} (k) are the outputs of the highpass (g) and lowpass (h) filters, respectively after subsampling by In this work, the 9th level wavelet obtained for each sampled speech input is a vector composed of 22 coefficients serving as the model for the speaker.

The matching and decision making processes: The ability of neural networks to accumulate knowledge about objects and processes using learning algorithms

Table I: Classification of identified and misidentified speakers according to classes

Class	Male			Female		
	up to 20	21-40	above 40	up to 20	21-40	above 40
No. of Samples	175	250	210	130	150	100
Correct Identification	82 (46.8%)	173 (69.1%)	142 (67.6%)	77 (59.2%)	90 (60%)	55 (55%)
False Rejection	58 (33.2%)	51 (20.4%)	40 (19%)	33 (25.4%)	37 (24.6%)	30 (30%)
False Acceptance	35 (20%)	26 (10.4%)	28 (13.4%)	20 (15.4%)	23 (15.4%)	15 (15%)

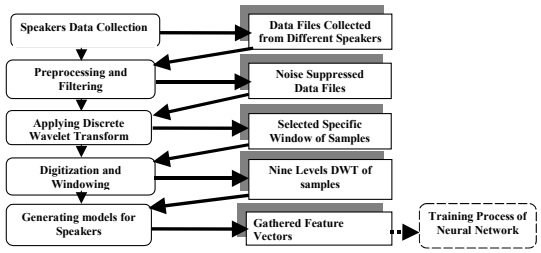

Fig. 1: Proposed system off-line activities

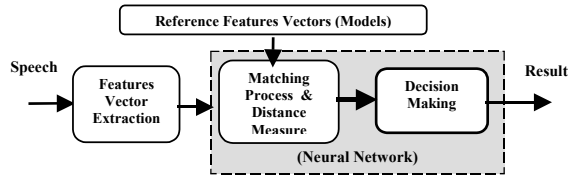

Fig. 2: The online processing for speaker identification

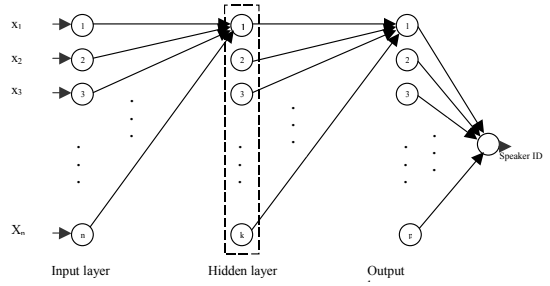

Fig. 3: An MVN-based neural network architecture

makes their application in speech recognition very promising and attractive, where different types of neural networks are successfully used for solving problems in both speaker verification and speaker identification. The type of neural networks, which is adopted in our work is a one based on multi-valued neurons (MVN),. The MVN- based neural network has been chosen depending on the fact that it support multi-valued threshold logic and its ability to implement arbitrary mapping between inputs and outputs described by partially defined multiple-valued function. This type of neural networks is also known by their quick converging learning algorithms. A comprehensive observation of MVN and its theoretical aspects together with its learning and properties are presented in[21].

The proposed system: The proposed system of speaker identification is composed of two main phases; first is an off line processing to generate a model (pattern) matching data file. This implies number of sequential steps as shown in the block diagram of Fig. 1. Once the samples are collected, preprocessing is applied to remove unwanted data as well as the redundant noise. Then it is converted to digital forms and stored in data files. A rectangular window is applied to limit the data used to a specific period. The data patterns of the different samples collected before are used to train the neural networks. The second phase of the system, whose steps are shown in Fig. 2, implements a strategy of speaker identification. This phase of the system applies a model matching approach that compares average features derived from test data with the collection of the stored average speaker's templates which are built in during the training process.

Based on the mathematical properties of MVN and their learning policy, we propose the NN structure. The general structure of the MVN-based neural network used for identification is presented in Fig. 3.

In the above figure, the input layer is composed of *n* neurons corresponding to *n* input values (*n =22*, which are the values contained in the ninth level DWT). The output layer contains 25 neurons representing a set of p speakers, where *p=25*. A hidden layer is used with eleven neurons *(k =11)*, where this number of neurons is found suitable. Throughout the experimentation, we have found that any increase of the hidden neurons amount does not improve the results, results may get worse for a smaller amount of the hidden neurons. This scheme complies with the set of the adopted number of

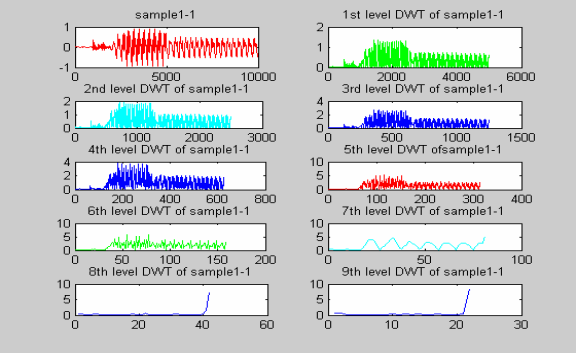

Fig. 4: The reference sample1_1 (S_{11}) and the corresponding wavelet application

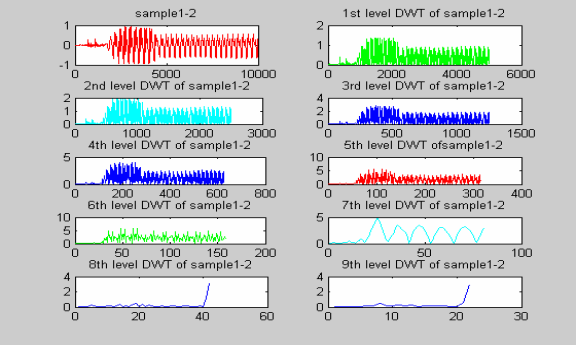

Fig. 5: The processed sample1_2 (S_{12}) and the corresponding wavelet applications

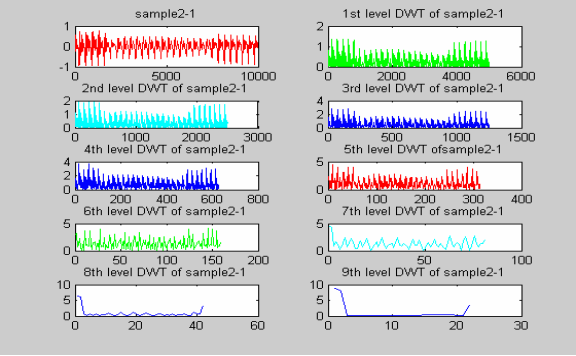

Fig. 6: The reference sample2_1 (S_{21}) and the corresponding wavelet application

speakers belonging to a typical organization (bank customers, club members, e-service subscribers .. etc.). Our goal is to identify a representative speaker related to such an organization. Such society is formed basically of two classes (male and female). These two classes are further classified according to the age into six classes (up to 20 years, 20-40 years and more than 41 years as being a male or a female).

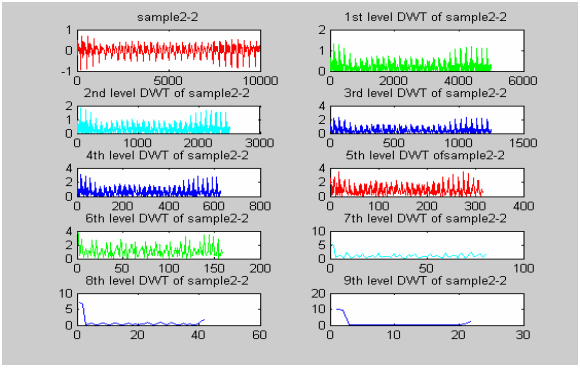

Fig. 7: The processed sample2_2 (S_{22}) and the corresponding wavelet applications

Results and evaluation of system performance: Referring to the structure of the proposed system, a series of processing are applied to the input speech samples, as shown in Fig. 1 & 2. In the digitization and windowing stage of Fig. 1, the acquired samples are passed through a window to truncate the data to a set of (10000) specific values that contain data of high speech entropy. All simulations are performed by using MATLAB 7.0. Figures 4 and 5 are showing the output of processing two speech samples (segments) acquired from speaker 1, who is a male aged above 40 years. We will refer to them as S_{11} and S_{12} being elements of a set S_1 (i.e. $S_{11}, S_{12} \in \{S_1\}$), where $\{S_1\} = \{S_{11}, S_{12}, S_{13}, \ldots\ldots\ldots, S_{1m}\}$ is the set of ($m = 75$) speech samples acquired from speaker 1. While Fig. 6 and 7 present other outputs (S_{21} and S_{22}), which are elements in a set S_2 containing speech samples of another speaker (speaker 2, a female aged 21-40 years). A similarity between the DWT levels (1-9) of Fig. 6 and their counterparts of Fig. 7 can be clearly noticed since they belong to the same speaker. Such form of similarity can also be seen in the spectral analysis presented in figures 4 and 5. Similar to what is performed on the samples S_{11}, S_{12}, S_{21} and S_{22}, elements of many sets ($\{S_1\}, \{S_2\}, \ldots\ldots, \{S_p\}$) belonging to a society of $p = 25$ speakers are used to train the neural network with each set comprising (30-100) elements. The system is then tested with speech samples which are either stored (i.e. from the training set) or samples which are not stored. The non-stored samples may belong to an enrolled speaker or to some other speaker. All the neurons were taught using a learning algorithm based on equation proposed in [22]. The final classification results are given in Table (I). In addition to the values appearing in this table the two parameters of the evaluation criteria (i.e. FAR and FRR) are calculated and found to be: FAR= 14.5% and FRR = 24.5%. These results, when

compared to outcomes from other works, prove to be good, taking in consideration that the speech samples used in our work are not dedicated for certain class of speakers but are gathered from a relatively random society of speakers. We can therefore consider the overall performance of the system is successful and promising.

CONCLUSION

Neural networks and wavelet transform techniques have been used as a hybrid approach for speaker identification, with the intention that a better performance of identification is to be obtained. Through the use of wavelet transform specific properties of speakers are extracted as vectors (patterns) and then subjected to a neural network based on multi-valued neurons. The activation function and learning properties of these neural networks are being invested to widen the threshold of accurate identification. In speaker identification systems, it is a fact that there is no 100% guarantee of accurate results. Our system, with its hybrid structure, gives acceptable results in speaker identification. Results, which can be considered of a good accuracy in spite of the fact that the society of speakers is relatively random. Concentrating on extracting the dominant features of a speech sample is an advantage of using wavelet transform, where it leads to a reduction in the storage capacity. Storage reduction is an important factor when talking about applications through internet and telecommunications. Finally, with a total of 61% correct identification, we can not claim that our system can be directly deployed into practical implementation. Further work can be carried out for improvements, specially on the feature extraction phase.

REFERENCES

1. Furui, S., 1997. Recent Advances in Speaker Recognition, Pattern Rec. Letters. 18: 859-872.
2. Campbell, J., 1997. Speaker Recognition: A tutorial. Proceedings of the IEEE. pp: 1437-1462.
3. Besacier, L., J.F. Bonnastre and C. Fredouille, 2000. Localization and Selection of Speaker-Specific Information with Statistical Modeling. Speech Communications. 31: 89-106.
4. Besacier, L. and J.F. Bonnastre, 2000. Subband Architecture for Automatic Speaker Recognition. Signal Processing. 80: 1245-1259.
5. Damper, R.I. and J.E. Higgins, 2003. Improving Speaker Identification in Noise by Subband Processing and Decision Fusion. Pattern Recognition Letters. 24: 2167-2173.
6. Sivakumaran, P., A.M. Ariyaeeinia and M.J. Loomes, 2003. Subband Based Text-dependent Speaker Verification. Speech Communications. 41: 485-509.
7. Reynolds, D.A, T.F. Quatieri and R.B. Dunn., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. pp: 19-4.
8. Bassam A. Mustafa, B. Y. Thanoon and S.D. Al-Shamaa., 2005. A Database System for Speaker Identificatoin. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan. May 2005.
9. Mohd Saleem, A. M., K. Mustafa and I. Ahmad., 2005. Spoken Word of German Digits Uttered by Native and non Native Speakers. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan. May 2005.
10. Prina Ricotti, L., 2005. Multitapring and Wavelet Variant of MFCC in Speech Recognition. IEE Proceedings on Vis. Image Signal Process., pp: 29-35.
11. Dokur, Z. and T. Olmz., 2003. Classification of Respiratory Sounds By using An Artificial Neural Networks. International Journal of Pattern Recognition and artificial Intelligence. 4: 567-580.
12. Abduladheem A., M.A. Alwan, and A.A. Jassim, 2005. Hybrid Wavelet-Network Neural/FFT Nural Phoneme Recognition. Proceedings of The 2nd International Conference on Information Technology. Al-Zaytoonah University of Jordan, May 2005.
13. Farrell, K.R., R.J. Mammone and K.T. Assalah., 1994. Speaker Recognition Using Neural. Networks and Conventional Classifiers. IEEE Trans. on Speech and Audio Proc. pp: 194-205.
14. Ramachandran, R.P. K.R. Frrell and R.J. Mammone., 2002. Speaker Recognition- General Classifier Approaches and Data Fusion Methods . Pattern Recognition. 35: 2801-2821.
15. Burrus, C., R. Gopinath and H. Guo. 1998. Introduction to wavelets and wavelet Transforms.1st edition. Prentice Hall.

16. Kinnunen, T., I. Kurkkainen and P. Franti, 2002. Is Speech Data Clustered?- Statistical Analysis of Cepstral Features. Proc. IASTED Int. Conf. of signal Processing and Communication (SPC),pp: 222-227, Marbella, Spain.
17. Rabinar, L. and R.W. Schafer, 1993. Fundamentals of Speech Recognition. Prentice Hall.
18. Long, C.J., and S. Datta. 1996. Wavelet Based Feature Extraction for Phoneme Recognition. Proc. Inter. Conf. on Spoken Language Proc. ICLSP'96. Philadelphia. USA. pp: 264-267.
19. Gupta, M. and A. Gilbert. 2001. Robust Speech Recognition Using Wavelet Coefficient Features. Proc. IEEE Automatic Speech Rec. and Understanding Workshop. Italy. pp:445-448.
20. Tzanetakis, George, E. Georg and Perry Cook., 2001. Audio Analysis using the Discrete Wavelet Transform. In. Proc. WSES Int. Conf. Acoustics and Music: Theory and Applications (AMTA 2001) Skiathos, Greece.
21. Aizenberg, I.N., N.N. Aizenberg and J. Vandewalle., 2000. Multi-Valued and Universal Binary Neurons: Theory, Learning, Applications. Kluwer Academic Publishers, Boston / Dordrecht/London.