Speaker Identification using Wavelet Analysis and Modular Neural Networks

Anupam Shukla Ritu Tiwari Hemant Kumar Meena Rahul Kala

Department of Information Technology,

Indian Institute of Information Technology and Management Gwalior,

Morena Link Road, Gwalior, Madhya Pradesh-474010, India

Citation: A. Shukla, R. Tiwari, H. K. Meena, R. Kala (2009) Speaker Identification using Wavelet Analysis and Modular Neural Networks, *Journal of Acoustic Society of India*, 36(1), 14-19.

Abstract

The increase in automation coupled with the increasing security issues urge the need of recognition and verification systems. The use of speech as a means for the same is hence becoming very popular. Speaker identification is the field where we try identify the person or speaker based on the spoken words. This makes use of signal processing techniques for the extraction of features from the speech signal as well as soft computing techniques for the recognition. In this paper we use wavelet analysis to extract useful characteristics of any author. These extracted features are trained using Modular Neural Network (MNN). The MNN consists of three ANNs which are trained individually with the entire data set using Back Propagation Algorithm (BPA). These vary in their architecture as well as the number of epochs used during training. Then a voting mechanism is adopted to integrate the solution of the individual ANNs and give the final ANN. We first record the training data set from a number of speakers with different words. In testing, we make the speaker speak out the same set of words. The features are extracted from and fed into the trained MNN. The MNN tells us the identity of the speaker. We recorded the voice data of

numerous speakers. The features were extracted and used in MNNs. A high performance of 94.7% clearly shows the efficiency of the algorithm.

KEYWORDS: Speaker recognition, Wavelet Analysis, Artificial Neural Networks, Modular Neural Networks, Wavelet Transforms, Speaker identity, Ensemble,

1. Introduction

The use of biometric features for the identification of persons presents a very promising technology because of the ease of use, speedily response and security reasons. Speech is one such biometric feature which is being used over decades for the identification and verification purposes. The problem of speaker identification is to collect the words spoken by the speaker and identify him/her using the various features present in his/her speech. The problem deals with developing mechanisms to acquire the speech, work out the relevant features that may be constant over natural variations and to develop mechanisms to learn and identify these features. The recognition systems make use of historical database which stores the speech spoken by the speaker at the time of registration.

In this paper we have used Wavelet Analysis to extract the various features of the speaker. This is an excellent means of analysis of such type of signals and is advancement over the Fourier analysis or Short Time Fourier Analysis (STFT) [1, 2].

We use Modular Neural Network (MNN) to identify the speaker by the characteristics extracted [3, 4, 5]. The neural networks are an excellent means to learn data and reproduce them whenever needed. They also give very good results to unknown inputs. This is known as the generalizing capability of the ANNs. The ANNs however many times suffer from problems like sub-optimal training, poor generalization, and incapability to handle certain data, etc. The modularity is added to handle these situations. Besides, modularity helps in giving good

performances and speedily training and response when exposed to a high dimensionality of data.

Many samples are recorded for different speakers. Each speaker says a set of words. The features are extracted from these samples and kept as training data for the MNNs. Then whenever a speaker has to be recognized, he must say those words. We would extract the features and feed it into the neural network. Using this we would get the identity of the speaker.

This paper is organized as follows. Section 2 deals with the motivation of the paper. Section 3 and 4 discuss the concept of Wavelet analysis and MNNs respectively. Section 5 talks about the procedure followed. The results are given in Section 6. Section 7 gives us the conclusion remarks.

2 Motivations

Speaker identification is one of the most developing fields of today [6]. A lot of work has already been done in this field using various tools and techniques. The basic aim of all this is to develop a system that can identify the speaker based on the voice characteristics.

A lot of work in this field exists in Hidden Markov Models. These are completely statistical models. These try to predict the expected value of output when the historic data is known. A lot of work also exists in Wavelet and other transforms [7, 8, 9, 10]. Here people have tried to identify the speaker by using transformations and analysis. Neural Networks have been extensively used for the machine learning [11]. They provide a convenient way to train the network and test it with high accuracy. Wavelet Analysis is a very promising feature extraction tool being used for signal analysis these days. When applied to speech signals, the analysis gives us valuable features that are robust against noise as well as minor variations. The sub-optimal or poor performance which might happen with conventional ANN is

handled with the MNNs. These two when combined form an excellent system for speaker recognition.

3 Analysis Technique

In this paper we have used Wavelet transform to extract characteristics. This analysis is advancement over Fourier analysis and Short Time Fourier Analysis (STFT) [12, 13, 14]. Fourier analysis breaks down a signal into constituent sinusoids of different frequencies. In this analysis however while transforming to the frequency domain, time information is lost. Short Time Fourier Analysis is a technique called windowing the signal which maps a signal into a two-dimensional function of time and frequency. It uses Fourier transform to analyze only a small section of the signal at a time. While the STFT compromise between time and frequency information can be useful, the drawback is that once you choose a particular size for time window, that window is the same for all frequencies.

The Wavelet Transform is a windowing technique with variable-sized regions. Wavelet analysis allows the use of long time intervals where we want more precise low-frequency information, and shorter regions where we want high-frequency information. This is shown in figure 1.

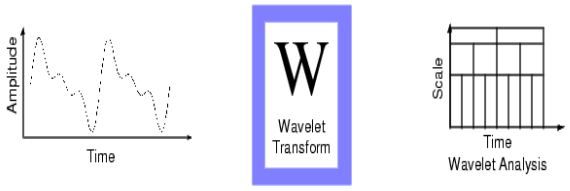

Fig.1: Wavelet Transform

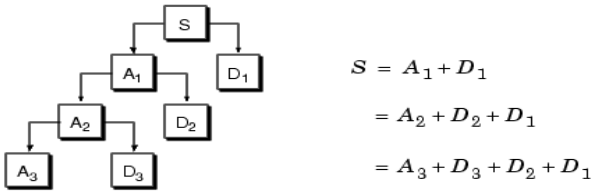

Fig.2: Wavelet Packet Analysis

Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, aspects like trends, breakdown points, discontinuities in higher derivatives,

and self-similarity. In wavelet analysis, a signal is split into an approximation and a detail. The approximation is then itself split into a second-level approximation and detail, and the process is repeated. For n-level decomposition, there are n+1 possible ways to decompose or encode the signal. This is given in fig. 2.

3.2 Modular Neural Networks.

The MNNs are an excellent means of machine learning when high dimensionality or precise decisions are involved [3, 4, 5]. The conventional ANNs have very high power of learning the past data and to generalize it to unknown inputs. This makes the ANNs the choice of most researchers for most of the real life problems. Unfortunately, the ANNs may fail in numerous situations. Many times, especially in the use of Back Propagation Algorithm (BPA), the ANN sub optimizes. Here it gives correct solutions to only a set of inputs. For the other inputs it may fail. The initial weights, training epochs, network architecture, etc. are known to cast a deep impact in the performance of the ANN. Choosing the incorrect value of any of these may lead to poor performance. Further, with the increase in the dimensionality of the ANN, the problem increases. Either the ANN fails to train, or improperly trains itself, or trains itself only for a set of values per input. This results in poor performance.

For these problems, modularity is usually added in the ANNs that lead to the emergence of MNNs. There are various kinds and architectures of MNNs that are commonly used. All of them employ the division of computation into various modules that the ANN is broken into. The task is performed in parallel. Later the results of the various modules are combined or integrated to give the final result. The novelty in use of MNNs lies in the way the computation is broken and integrated in the problem.

The MNN we use in this paper is an ensemble MNN that uses majority rule or a voting mechanism for the purpose of integration of various ANNs. The general architecture

of this system is given in figure 3. This contains different ANNs. Each ANN has a slightly different architecture in terms of number of neurons in the hidden layer. Each ANN is trained to a different number of epochs using the same training data that is common to all ANNs. TO calculate the output, we first apply the input to all the ANNs and calculate the individual outputs. Each ANN returns the output or class that the input may belong to as per its calculations. A voting is done amongst the ANNs and the class that gets the maximum number of votes wins. In case of a tie, the preference is given to the ANNs earlier in the list of ANNs

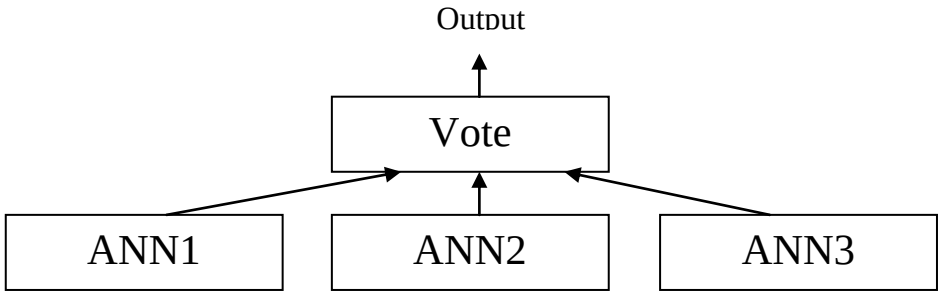

Fig. 3: The Ensemble ANN

4. Procedure

In this section we discuss the procedure followed for the experiment. The basic concept used is that we use Wavelet Analysis to extract features. Using this we can convert any speech into a set of features. These features were used for further processing.

We took various speakers who had to be identified. These speakers were made to say a set of works one after the other. We applied Wavelet transform to each of the words spoken by the speaker. This gave us a series of features extracted. Hence we got a series of features for every word spoken by any speaker. These series of values were extracted from the speech.

This formed a good database for the training of the MNN. Three ANNs were trained for each speaker. Training not only ensured that the ANNs learn the features of each and

every speaker but also that the different ANNs differ to some extent in generating the final output. Then we tested the system. Some random inputs were given to the system. The input consisted of all the words spoken in order by the speaker. The output was collected and the speaker was identified. In the next sections we discuss the details of the Wavelet Analysis and the MNNs.

4.1 Wavelet Analysis.

The wavelet analysis is used for analysis of the given input speech signal. Here the signal refers to a particular word spoken by a particular speaker. The output of the Wavelet analysis is the set of features that were extracted. As stated above, the Wavelet analysis consists of a detail and approximation. An approximation can be further broken down into detail and approximation. Hence we take a series of levels. A level here represents the degree of detail in the analysis.

For this problem we fix the number of details to be 5. These are numbered as D_1, D_2, D_3, D_4, D_5 and A_5. Hence for every word spoken, we extract a total of six features. These features are further used for the MNNs.

4.2 Modular Neural Network.

The purpose of the MNN is to learn the data for any speaker first. The learning is followed by the testing. Here whenever a speaker says the set of words, he is identified by the MNN. Hence using the neural networks we can classify any input sequence to an output class. We discuss the inputs and outputs in the next section.

4.2.1 Input.

We know that every speaker speaks a total of 5 words. Every word is a collection of 6

features extracted. Hence there are a total of 30 features for every speaker. These form the 30 inputs for the neural network. These 30 inputs are given to the neural network at every epoch of the neural network. These are numbered as I_1, I_2, I_3, I_4…. I_{20}

In this paper we do not take the straight values as input. On the other hand we calculate the values of each of the input from the data extracted by the Wavelet Analysis. We have a huge amount of data for every input which is the training data. The general formula is given by equation (1)

$$I_i=(V_i - \text{Min}(V_{ij})) \ / \ (\text{Max}(V_{ij}) - \text{Min}(V_{ij})), \quad \text{for all } j \qquad (1)$$

Here I_i is th i^{th} input of the neural network

V_i is the i^{th} feature extracted from Wavelet Analysis

$\text{Min}(V_i)$ is the minimum of all V_{ij} found in the training data set

$\text{Max}(V_i)$ is the maximum of all V_{ij} found in training data set for all j in data set

In this formula we have basically normalized the inputs and get their ranges between 0 and 1. This can be well worked upon by the neural network.

4.2.2 Output.

In this problem we have created a classificatory output pattern. In this system there are as many numbers of outputs as are the number of classes. The outputs are numbered O_1, O_2, O_3…O_n where n is the number of classes. Each output O_i represents the probability of the output to be the i^{th} class. This varies from -1 to 1. The higher the output, the more will be the probability of the class to which the output belongs to be the class of the output.

Hence in this way we need to find the maximum output and the class corresponding to

it is the final class. For the training phase, the output of the class to which the input belongs is taken as 1, the output of the other classes are taken as -1. Hence if there are a total of 10 speakers, the output for the first speaker considered will always be <1,-1,-1,-1,-1,-1,-1,-1,-1,-1>.

4.2.3 Ensemble.

The majority vote as discussed earlier was used for the purpose of the integration of the various outputs of the individual ANNs to generate the final output. There were 3 ANNs that we used. In case a majority was not formed we preferred the ANN that was the first to be trained over the other two ANNs. Its output was regarded as the final output.

The three ANNs taken had a difference of one neuron in the first and only hidden layer. Each was trained for a different number of epochs with a difference of 10% from each other.

5 Results

In order to prove the algorithm, we recorded data of 20 speakers. They were made to say 5 words each '*ab*', '*is*','*baar*','*aap*'and '*apne*'. The speakers repeated these words and all of them were recorded. The Wavelet Analysis kit of Matlab was used for the Wavelet Analysis. The number of levels was specified to be 6. This gave us 6 different values found from wavelet analysis. These values were recorded for each and every word for each author. The values were then collected into Matlab. The values were processed to form the input of each of the ANN in the MNN. The output class of each of these values was known. The network was trained using these values. In testing we used new data and did the entire process with it again. We extracted its features, processed the inputs and fed it into the each

of the ANN in MNN. The outputs were received. The output array was iterated for all the classes to find the maximum output class. This class was declared as the output.

The results of some the inputs of the wavelet analysis are given in table 1. Each ANN had a total of 1 hidden layer. There were a total of 47, 48 and 49 neurons in the hidden layer of the ANNs.

Table1: the results of the Wavelet Analysis

S.No.	Speaker	Word	D1	D2	D3	D4	D5	A5
1	A	1	1825	3642	7275	14542	29075	1825
2	A	2	1811	3614	7220	14432	28855	1811
3	A	3	1904	3799	7590	15172	30335	1904
4	A	4	1873	3737	7466	14924	29840	1873
5	A	5	1845	3682	7355	14702	29395	1845
6	B	1	1842	3675	7341	14673	29338	1842
7	B	2	1799	3590	7171	14334	28659	1799
8	B	3	1852	3696	7384	14760	29512	1852
9	B	4	1851	3694	7380	14752	29495	1851
10	B	5	1843	3678	7347	14685	29362	1843

We got a performance of 97.5%. A high performance as the result clearly shows that the algorithm works well and gives correct results on almost all inputs that are given.

6. Conclusion

In this paper we proposed the use of Wavelet Transform and MNN for the speaker identification. Wavelet transforms were used for the feature extraction. We extracted six features per word spoken by the author. These characteristics were given as a training data to the neural networks. The neural networks learnt the characteristics and reached the performance goal. When a new data was given as input to the neural network, it could make

out the class to which it belongs. An acceptable performance level of 97.5% clearly show the fact that the algorithm can be used for the identification of the speaker. In this paper we have proposed the use of Wavelet Analysis for the speaker identification. The same theory may also be used for the identification of words. Identification of author using multi-lingual words and the identification of speakers by giving any order of words needs to be done in future.

References

[1] G. A. Papakostas, , D. A. Karras, B. G. Mertzios, and Y. S Boutalis, 2007, An Efficient Feature Extraction Methodology for Computer Vision Applications using Wavelet Compressed Zernike Moments, *ACM International Journal of Information Sciences*, Vol 177, Issue 13

[2] Yuan-Liang Tang and Chih-Jung Hung, 2005, Recoverable Authentication of Wavelet-Transformed Images, *ICGST International Journal on Graphics, Vision and Image Processing*, Vol S11, pp 61-66

[3] Gasser Auda and Mohamed Kamel, 1998, Modular Neural Network Classifiers: A Comparative Study, J*ournal of Intelligent and Robotic Systems*, pp 117–129

[4] Farooq Azam, Biologically Inspired Modular Neural Networks, PhD Thesis, *Virginia Polytechnic Institute and State University*

[5] Albrecht Schmidt, A Modular Neural Network Architecture with Additional Generalization Abilities for High Dimensional Input Vectors, PhD Thesis, *Manchester Metropolitan University*

[6] Joseph P Campbell Jr, 1997, Speaker Recognition: A Tutorial, *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462

[7] C. Ben Amar and O Jemai, 2005, Wavelet Networks Approach for Image Compression, *ICGST International Journal on Graphics, Vision and Image Processing*, Vol. SI1, pp 37-45

[8] C.J. Long and S Datta, 1996, Wavelet Based Feature Extraction for Phoneme Recognition, *Proc. of 4th Int. Conf. of Spoken Language Processing*, pp 264-267

[9] Stephane G. Mallat, 1989, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *674 IEEE Transactions On Pattern Analysis and Machine Intelligence.*, Vol. II, No. 7.

[10] S. Mohanty, S. Bhattacharya, 2008, Recognition of Voice signals for Oriya Language using wavelet Neural Network, *ACM International Journal of Expert Systems with Applications*, Vol 34, Issue 3, pp 2130-2147

[11] Anupam Shukla and Ritu Tiwari, 2007, Fusion of Face and Speech Features with Artificial Neural Network for Speaker Authentication, *IETE Technical Review*, Vol 24, No 5, pp 359-368

[12] Christopher Torrence and Gilbert P; Compo, 1998, A Practical Guide to Wavelet Analysis', Bulletin of the American Meteorological Society, Vol 79, pp 61-78

[13] George Tzanetakis, Georg Essl, Cook Perry, 2001, Audio Analysis using the Discrete Wavelet Transform, *In. Proc. WSES Int. Conf. Acoustics and Music: Theory and Applications* Skiathos, Greece

[14] Hubert Wassner, Geard Chollet, New Sepstral Representation using Wavelet Analysis and Spectral Transformation for robust speech recognition