Speech Recognition using Wavelet Packets, Neural Networks and Support Vector Machines

Purva Kulkarni, Saili Kulkarni, Sucheta Mulange, Aneri Dand, Alice N Cheeran
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai, India.

***Abstract*—This research article presents two different methods for extracting features for speech recognition. Based on the time-frequency, multi-resolution property of wavelet transform, the input speech signal is decomposed into various frequency channels. In the first method, the energies of the different levels obtained after applying wavelet packet decomposition instead of Discrete Fourier Transforms in the classical Mel-Frequency Cepstral Coefficients (MFCC) procedure, make the feature set. These feature sets are compared to the results from MFCC. And in the second method, a feature set is obtained by concatenating different levels, which carry significant information, obtained after wavelet packet decomposition of the signal. The feature extraction from the wavelet transform of the original signals adds more speech features from the approximation and detail components of these signals which assist in achieving higher identification rates. For feature matching Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are used as classifiers. Experimental results show that the proposed methods improve the recognition rates.**

Keywords*— *Wavelet Packet Transform, Feature Extraction, Artificial Neural Networks, Support Vector Machines

I. INTRODUCTION

Speech recognition is the process of extracting and determining information conveyed by a speech signal using computers. Automatic speech recognition methods, investigated for many years have been principally aimed at realizing transcription and human computer interaction systems. Speech recognition is currently used in many real-time applications, such as mobile phones, smart devices, computers, and security systems. However, these systems are far from perfect in correctly classifying human speech into words. Speech recognizers consist of a feature extraction stage and a classification stage. The feature extraction can be considered as a dimensionality reduction process that attempts to capture the essential characteristics of the speech with less memory requirements for signal representation.

There are various techniques for extracting speech features in the form of coefficients such as the linear prediction coefficients (LPCs), the Mel-Frequency Cepstral Coefficients (MFCCs) and the Linear Prediction Cepstral Coefficients (LPCCs) [6]. Classification techniques used in speaker identification systems include Gaussian Mixture Models (GMMs), Vector Quantization (VQ), Hidden Markov Models (HMMs) and ANNs [3], [6], [11], [12].

The MFCCs are the most popular speech features used in speaker identification [7]. The MFCCs for recognition performs better in clean environments, but they are not robust enough in noisy conditions. They are based on the known evidence that the information carried by low frequency components of the speech signal is more than that carried by high frequency components. MFCC analysis method uses the Discrete Fourier Transform (DFT) to transform fixed frames of speech into its corresponding frequency domain thus the disadvantages of the Fourier Transform would also exhibit in these cepstral analysis methods.

In the MFCC technique the input speech signal is divided into fixed size short frames for analysis. Thus in the MFCC method the signal in these short frames is considered to be stationary but in reality the speech spectrum is highly fluctuating and unpredictable. Pitch, tone, phoneme pronunciations differ in accordance with the speakers, thus the algorithm should be able to extract correct features despite of these fluctuations. The MFCCs assume that the speech signal is stationary within a given time frame and therefore lack the ability to analyze the localized events accurately.

To address this problem, a lot of research has been directed towards the use of wavelet based features [4], [5], [14]. The discrete wavelet transform (DWT) has a good time and frequency resolution and hence it can be used for extracting the localized contributions of the signal of interest [10]. Wavelet denoising can also be used to suppress noise from the speech signal and it can lead to a good representation of stationary as well as non-stationary parts of the speech signal. To facilitate this, wavelets transformation is applied to the frames instead of FFT in the conventional MFCC method. Moreover, the wavelet analysis also gives the provision to go upto any resolution as desired by the user which plays a very crucial role in the process of feature extraction in speech recognition.

II. WAVELET PACKET TRANSFORM

The idea behind the Wave Packet transform is to decompose both approximation and detail parts of DWT for next level of analysis. The dyadic DWT involves decomposition of only the so called approximation subspace

978-1-4799-3140-8/14/$31.00 ©2014 IEEE

(Fig. 1). In the wave packet transform the objective is to get around this limitation, so it is intended to decompose the incremental subspace (i.e. detail subspace) as it is done for approximation subspace in DWT (Fig. 2). For example to decompose V_1 into V_0 and W_0, it is also intended to decompose W_0 in the next step.

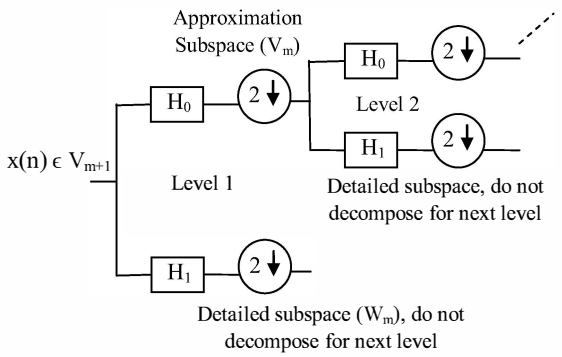

Fig. 1: Structure of Discrete Wavelet Transform

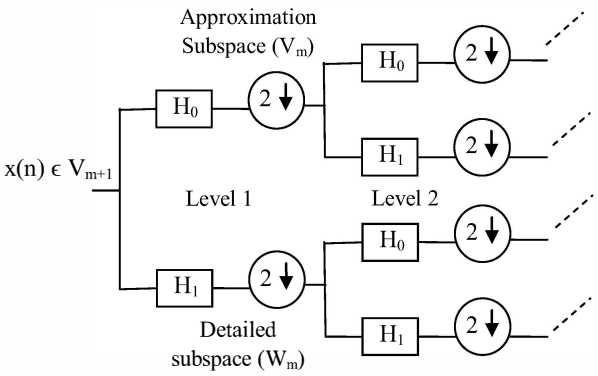

Fig. 2: Structure of Wavelet Packet Transform

The key advantage of wavelet packet decomposition is the flexibility it offers. Instead of confining to the extremes of dyadic decomposition or full subband decomposition, intermediate forms can be chosen. This is particularly useful in feature selection process where one way of attaining compression is by discarding wavelet subbands with very low activity. Either the dyadic decomposition or the subband decomposition may yield subbands that have partial activity. Rather than retaining such a subband in its entirety, it is better to split further to confine the activity to most of a smaller subband and discarding other subbands.

III. METHODOLOGY

Speech recognition is addressed in two steps; feature extraction and classification;

A. Feature Extraction

In the methods used below wavelet packet transform is preferred as compared to wavelet transform. When the speech spectrum is observed a significant amount of information is found to be present in the lower frequencies along with a few higher frequencies. Thus the wavelet packet transform helps in capturing the significant information at both low as well as high frequencies in various subbands, which provides the efficient feature set.

Fig. 3 shows the tree structure of the wavelet packet coefficients. Out of these subbands, it is sufficient to select only those sub-bands that contain maximum information. According to the database used in this paper, following levels are found to be significant:(1,0), (2,2), (3,2), (4,0), (4,1), (4,2), (4,8), (4,12).

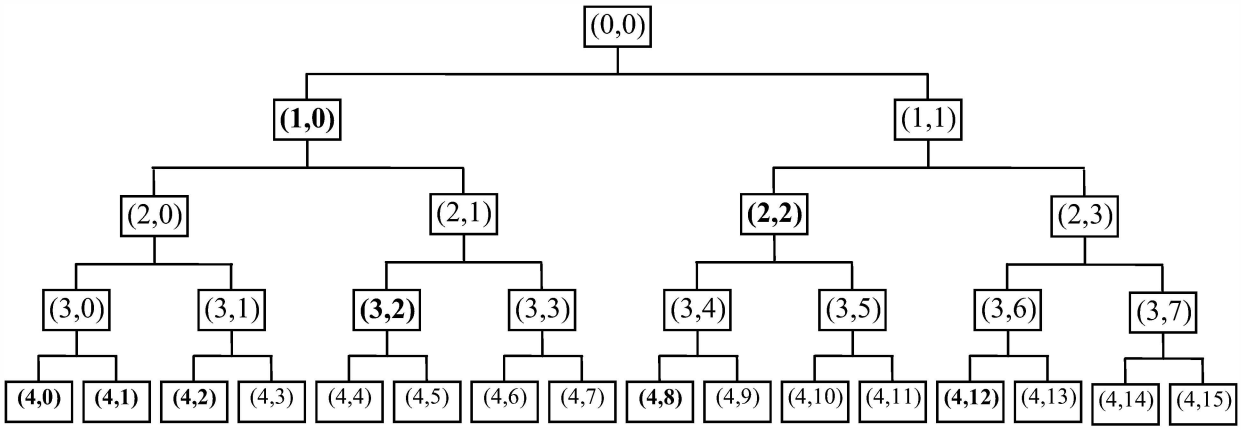

Fig. 3: Wavelet Packet Coefficients Tree

Due to the highly variable nature of speech signals in term of its duration, acoustic information even for a similar word utterance by same speaker at different time, the number of frames per speech pattern will be most likely different from each other [8]. Since number of input to the neural network is fixed, the number of speech pattern frame need to be normalized prior to giving it to the network for training or testing. To solve this problem two techniques are used:

1) Frame Normalization (used in method 1)

There are three possibilities of the number of frames after features extraction which are:

a. Number of frame is equal to the number of input node.

b. Number of frame is less than the number of input node.

c. Number of frame is greater than the number of input node.

Fix the number of frames (f). If the number of frames received (s) after framing is greater than fixed number then chose all the (i*s/f) frames and if it is less than fixed then chose (i*f/s) frames where i =1 to n (n=s) after doing this fill all the empty frames with previous adjacent frame [1], [13].

2) Principal Component Analysis (used in method 2)

The frame size is fixed of duration 10ms which is kept small as to capture maximum information; the number of frames per sound signal obtained is large and different for different sound signals. PCA reduces redundancies in the data and thus reduces the number of frames to a fixed number, which are linearly uncorrelated by using orthogonal transformation. Thus, the requirement of fixed number of data points of the classifier is met and also the data size is reduced, as only significant information is retained [16].

- Method 1

1) Pre-emphasis
2) Framing: to divide in frames of 10-25ms
3) Windowing
4) Wavelet packet transform (this is instead of FFT in MFCC)
5) Filter bank energies of the above mentioned subbands are calculated and logarithm of these energies is operated upon by DCT. DCT provides efficient way of concentrating the feature information in the first few coefficients itself [9], [15].

Steps 3-5 give a feature set of one frame. A feature matrix is constructed with each row representing a feature set of each frame. Thus, the number of rows in the feature matrix is equal to the number frames obtained.

- Method 2

In this method the coefficients obtained in subbands are concatenated. Then to form a feature matrix frame wise feature set is arranged in a row wise fashion [2]. Again the number of rows is equal to the number of frames.

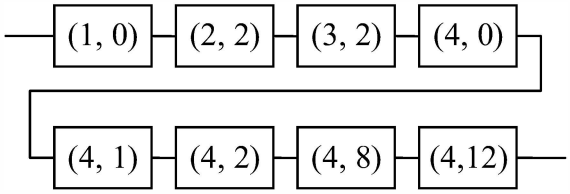

Fig. 4: Feature Set Formation

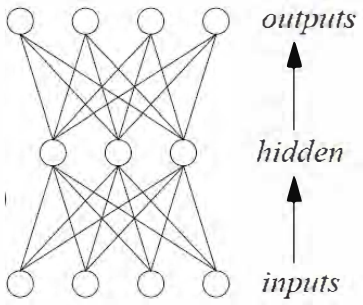

Fig. 5: Three layer Feed-forward network

B. Classification

Once the features are extracted for each frame of the each word, these vectors are used to create a classifier model which is trained to identify different words. The model is made by training the feature vectors as observations. For training, a Gaussian Mixture Model is made for each sound sample. And output of this, which is given in accordance with the Expectation Maximization algorithm, is a row vector having data points equal to the number of rows of the feature matrix which is ultimately given to the classifier for training.

1) Artificial Neural Networks

Artificial neural networks (ANNs) are intelligent systems that are related in some way to a simplified biological model of the human brain. They are composed of many simple elements, called neurons, operating in parallel and connected to each other in the forward path by some multipliers called the connection weights. Neural networks are trained by adjusting values of these connection weights between the network elements. Neural networks have self learning capability, are fault tolerant and noise immune, and have applications in system identification, pattern recognition, classification, speech recognition, image processing, etc. In this work, back propagation based ANN is used for classification.

- Back-propagation Algorithm

A three layer feed-forward neural network with a sigmoidal hidden layer followed by a linear layer is employed in this application for classification. The neural network is trained using the gradient descend based back-propagation algorithm. A momentum term is used in the algorithm to achieve a faster global convergence. In this application, an adaptive back-propagation learning method is employed that is

the learning gain is adjusted during the training to enhance faster and global convergence. The three layer feed-forward neural network architecture for this application is shown in Fig. 5.

2) Support Vector Machines

A classification method is a particular way of constructing a rule, also called a classifier, from the labeled data and applying it to the new data. A binary (two-class) classification problem can be described as follows: for a given set of labeled points (x_i, y_i), x_i are vectors of features and $y_i \in$ (-1, +1) are class labels, construct a rule that correctly assigns a new point x to one of the classes. The vectors x_i in this formulation correspond to objects, and the dimensions of the space are the features or characteristics of these objects. Using labels {0, …., K-1}instead of (-1, +1) a multiclass problem with K classes can be described. Support Vector Machines is a method for constructing a special kind of rule, called a linear classifier, in a way that produces classifiers with theoretical guarantees of good predictive performance (the quality of classification unseen data). The theoretical foundation of this method is given by statistical learning theory [17].

IV. Results and Discussions

All the experiments are performed in MATLAB R2011a. The classifiers are trained with 48 samples. Each word is divided in 25 ms frame for method 1 and in 10 ms frame for method 2. The mother wavelet used is Daub -1, decomposition is done till 4^{th} level, subbands (1,0), (2,2), (3,2), (4,0), (4,1), (4,2), (4,8), (4,12) are used for forming the feature set for method 1 and 2. The classical MFCC is also implemented for comparative evaluation with the proposed methods.

Table I & II show the results of method 1 and method 2 with NN respectively. It is tested with four inputs. First three are from the training set. The fourth one is used for testing the classifier. The values in the respective input column represent the output of the neural network after training.

Table III shows the results of classical MFCC with NN. It is tested with four inputs. First three are from the training set. The fourth is the input not from the training set. The values in the respective input column represent the output of the neural network after training. It is observed that efficiency of proposed methods is better than that of classical MFCC.

Table IV describes the results obtained from testing Method 1, Method 2 and MFCC with SVM as the classifier. Class column represents the original class of the input signals. The model is tested with 6 samples from each class. For e.g. for class one six test inputs were tested out of which four were classified correctly. The output column demonstrates the number of times each class is classified correctly.

For each word, there are 10 sound samples and there are 6 such words. Out of these 10, 8 are given for training and 2 are reserved for testing. In the set of 10 sound samples, 5 samples are of one speaker and rest five of another. Testing is first done by giving one of the sound samples from those 8 samples, then the classifier is tested with a sample from the other 2 sound samples not used for training.

Table V shows the efficiencies of the described three methods using both the classifiers. It is observed that the features extracted using wavelets show better results than MFCC for both the classifiers. Between method 1 and 2, features extracted using method 1 are better as they have advantages of both MFCC and wavelets. Also, method 1 features are concise as compared to method 2 hence the classifier takes lesser time for training and classification. Hence they are more efficient. It is observed that neural network classifier gives better results as compared to SVM for chosen data base.

TABLE I. Proposed Method-1 usiing NN Classifier

Class	Input-1	Input-2	Input-3	Test Input
1	1.0057	-0.0005	0.8794	0.8099
	0.1603	0.2942	0.0569	-0.0303
	-0.0477	0.4085	0.0102	-0.1146
	-0.0360	0.4886	0.5445	-0.3695
	-0.1197	-0.3779	-0.6493	0.3849
	0.1485	0.1375	0.6824	0.3080
2	-0.1110	0.0975	-0.0162	0.1110
	0.9510	1.0781	1.0308	0.6856
	0.8340	0.0450	0.0118	-0.8388
	-0.5943	0.0405	0.0132	0.2548
	0.0512	0.1399	0.076	0.5098
	0.4317	-0.1675	-0.0174	0.2744
3	0.1266	0	0.0147	-0.3546
	0.1653	0	-0.0132	-0.2321
	0.6009	1	0.9573	0.5191
	-0.1723	0	-0.0639	0.1403
	0.2622	0	0.0601	0.1666
	-0.1371	0	0.0595	0.2850

TABLE II. Proposed Method-2 usiing NN Classifier

Class	Input-1	Input-2	Input-3	Test Input
1	0.2076	0.3312	1.0154	0.5976
	-0.0915	0.0934	-0.0254	0.2073
	0.2672	` 0.2471	0.1182	-0.0114
	0.1748	0.2479	0.0008	0.1503
	0.1888	-0.0879	-0.1254	0.0008
	0.1760	0.3143	-0.1236	0.0623
2	0.6588	-0.0609	-0.0832	0.1478
	0.5610	0.3465	-0.3028	0.0097
	0.1252	0.0787	-0.0980	-0.0024
	0.4480	-0.0353	0.1012	0.2378
	-0.0186	0.2838	0.0187	0.0525
	0.2002	-0.0508	0.7400	0.3408
3	0.3082	0.0665	0.2408	0.0689
	-0.0962	-0.1514	-0.2622	-0.1372
	0.4160	0.3993	0.5880	0.3714
	0.2703	0.1432	0.1881	0.1893
	-0.1233	0.2244	-0.1065	0.2489
	0.2798	0.0899	0.2084	0.1024

TABLE III. MFCC USIING NN CLASSIFIER

Class	Input-1	Input-2	Input-3	Test Input
1	0.9968	0.9602	0.9906	2.1430
	0.0014	-0.1232	-0.0774	1.5732
	-0.0040	-0.2082	-0.0082	0.8583
	-0.0003	-0.0122	-0.0889	0.5924
	0.0027	0.0504	-0.0173	0.4268
	0.0024	-0.0686	0.1411	1.9895
2	0	0	0.7866	0.2139
	1	1	0.1450	-1.3874
	0	0	-0.1398	-0.4080
	0	0	0.0263	0.1940
	0	0	0.1581	2.0167
	0	0	-0.2865	0.3057
3	0.1444	-0.1067	-0.083	0
	0.2394	0.9202	-0.228	0
	0.7322	0.5541	0.8873	0
	-0.2637	0.4272	-0.0172	1
	0.2712	0.1175	0.0087	0
	-0.2190	0.7944	0.0267	0

TABLE IV. PROPOSED METHOD AND MFCC USIING SVM CLASSIFIER

Class	Method 1		Method 2		MFCC	
	Input Class	Output	Input Class	Output	Input Class	Output
1	1	1	1	1	1	1
	1	1	1	1	1	1
	1	1	1	1	1	1
	2	1	2	1	2	1
	2	1	2	1	3	1
	3	0	3	1	3	1
2	2	2	2	2	2	2
	2	2	2	2	2	2
	2	2	2	2	2	2
	1	2	1	2	1	2
	3	0	3	2	1	2
	3	0	3	0	3	2
3	3	3	3	3	3	3
	3	3	3	3	3	3
	3	3	3	3	3	3
	1	3	1	3	1	3
	1	3	1	3	2	3
	2	0	2	0	2	3

TABLE V. RECOGNITION PERFORMANCE

	MFCC	**Method 1**	**Method 2**
SVM	50%	73%	61.67%
ANN	66.67%	91. 67%	66.67%

REFERENCES

[1] Adam T.B., Salam, and Gunawan, "Wavelet based cepstral coefficients for neural network speech recognition," IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2013," pp. 447-451, 2013.

[2] Cherupalli, Pranav Kumar and Gari, Jaya SankarKottareddy, "A Wavelet based Feature Extraction for Voice-Lock systems," TENCON 2005 2005 IEEE Region 10, pp. 1-4, 2005.

[3] Gandhiraj R., P.S. Sathidevi," Auditory-based Wavelet Packet Filterbank for Speech Recognition using Neural Network", Proceedings of the 15thInternational Conference on Advanced Computing and Communications, pp.666-671, 2007.

[4] Hsieh, C-T Lai, E and Wang, Y-C, "Robust speech features based on wavelet transform with application to speaker identification," IEE Proceedings-Vision, Image and Signal Processing, vol. 149 (2), pp. 108-114, 2002.

[5] Jong B.C. , "Wavelet Transform Approach For Adaptive Filtering With Application To Fuzzy Neural Network Based Speech Recognition", PhD Dissertation, Wayne State University, 2001.

[6] Kesarkar, Manish P, Feature extraction for speech recognition, Electronic Systems, EE. Dept., IIT Bombay, 2003.

[7] Klautau, Aldebaro, "The MFCC", Technical report, Signal Processing Lab, UFPA, Brasil, 2005.

[8] Maorui, Bai and Mingming, Feng and Yuzheng, Zheng, "Speech Recognition System Using a Wavelet Packet and Synergetic Neural Network," Measuring Technology and Mechatronics Automation (ICMTMA), pp. 453-456, 2010.

[9] Patil, HA Basu, TK, "Comparison of subband cepstrum and Mel cepstrum for open set speaker classification," India Annual Conference, 2004. Proceedings of the IEEE INDICON, pp. 35-40, 2004.

[10] Polikar, Robi, "The Engineer's Ultimate Guide To Wavelet Analysis-The Wavelet Tutorial", available at http://www. public. iastate. edu/rpolikar/WAVELETS/WTtutorial. html. Accessed on 15th Dec. 2013.

[11] Polur P.D. and G. E. Miller, " Experiments With Fast Fourier Transform, Linear Predictive and Cepstral Coefficients in Dysarthric Speech Recognition Algorithms Using Hidden Markov Model", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 13, No. 4, pp. 558-561, 2005.

[12] Pullella D., "Speaker Identification Using Higher Order Spectra", Dissertation of Bachelor of Electrical and Electronic Engineering, University of Western Australia, 2006.

[13] Salam, Md and Mohamad, Dzulkifli and Salleh, Sheikh, Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters", International Arab Journal of Information Technology (IAJIT), vol. 8 (4), 2011.

[14] Sarikaya R., "Robust And Efficient Techniques For Speech Recognition in Noise", PhD Dissertation, Duke University, 2001.

[15] Sarikaya, Ruhi and Pellom, Bryan L and Hansen, John HL, "Wavelet packet transform features with application to speaker identification," IEEE Nordic Signal Processing Symposium, pp. 81-84, 1998.

[16] Smith, Lindsay I, A tutorial on principal components analysis, Cornell University, USA, vol. 51, pp. 52, 2002.

[17] Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, 2nd edition, 1998.