Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 8, August 2014, pg.63 – 71

RESEARCH ARTICLE

Speech Recognisation System Using Wavelet Transform

Ankita Chugh
Department of Electronics and Communication
PDM College of Engineering for Women, Bahadurgarh
Haryana, India
Ankita16chugh@gmail.com

Poonam Rana
Department of Electronics and Communication
PDM College of Engineering for Women, Bahadurgarh
Haryana, India
jaglanpoonam@gmail.com

Suraj Rana
Department of Electronics and Communication
MRIEM, Rohtak
Haryana, India
rana.suraj@gmail.com

***ABSTRACT:** To develop speech recognition system with low word error rate using wavelet transform through pattern recognition approach. The aim of this paper is to make intelligent system that can recognize the speech signal. This includes also how the feature extracted from the speech signal using Discrete Wavelet Transform and then Dynamic Time Warping is used for pattern matching from the stored database of stored pattern to recognize the test word.*

***Keywords:** Speech recognisation, dynamic time warping, discrete wavelets transform, short time fourier transform, fast fourier transform*

I. INTRODUCTION

Modern technology is advancing in the direction of better man-machine interaction. Initial steps for human-machine communications led to the development of keyboard, the mouse, the trackball, the touch-screen, and the joystick. . However none of these communication devices provides the ease of use devices provides the ease of use of speech, which has been the most natural form of communication between humans for centuries. This calls for the development of a speech recognition system that can be added to a machine to accept spoken commands. Speech recognition by machine refers to the capability of a machine to convert human speech to a textual form, providing a transcription or interpretation of everything the human speaks while the machine is listening. Speech recognition is the classification of spoken words by a machine. The words are transformed into a format that a machine can understand then matched in some way against a template or dictionary of previously identified sounds. There are several issues when developing a speech recognition system. One is to determine if the system will be for a single user or many different users. The first type of system is called a speaker dependent system and is much easier to develop because system only has to determine what a single user has uttered. The template or database consists of signals recorded by the same user that is going to use the system. The second is a speaker independent system. The speech recognition is divided into two stages one is training stage and another is recognition stage. In training stage speech features are extracted and saved to make a reference template. The recognition phase may be divided further into two stages .The first one is the feature extraction stage wherein short time temporal or spectral features are extracted. The second one is the classification stage wherein the derived parameters are compared with stored reference parameters and decisions are made based on some kind of minimum distortion rule. For feature extraction, some kind of transformation is used which can give time-frequency analysis of the speech signal, Short Time Fourier Transform, Linear Predictive Coding are a few of them. Wavelets can also be used in creating a speech recognizer. A wavelet is a wave of finite duration and finite frequency. They have the ability to capture localized features of a signal and act in much the same way as the Fourier transform acts with sine’s and cosine’s. Because of this good localization of features, wavelets can be very useful in speech recognition. The wavelet transform is a technique that processes data at different resolutions and scale. The output of the wavelet transform is a set of approximation coefficients and a set of detail coefficients. By taking the wavelet transform of the previous ones approximation coefficients, more and more octaves can be generated. In this work, Discrete Wavelet Transform is used for feature extraction.

II. BACKGROUND

Problems in recognizing speech include noise, speaker variations, and differences between the training and testing environments, such as the microphones used **[1].** One way of dealing with this is to adapt the recognition system’s internal model (i.e. Hidden Markov Model weights). Another is to normalize the new speech to conform to the training data. Variations with different speakers mean that speaker dependent systems usually do better than speaker-independent ones, since the former uses the speaker for training. Dynamic Time Warping, or a similar algorithm, is necessary because of the non-uniform patterns of different speech signals. Also, different speakers will more than likely say the same words at different rates. This means that a simple linear time alignment comparison, such as the root square mean error, cannot be used efficiently. One way to do speech recognition is phoneme-based indexing **[2].** A phoneme is a basic sound in a language, and words are made by putting phoneme together. One method is to consider the trip hone, a set of three phonemes where a phoneme is considered with its left and right neighbors **[3].**Therefore, this method identifies speech based on its component phonemes. We are not trying to match a spoken word to a word list, but rather output the phonemes detected. For example, if the user says the word “pocket”, our system should output “p”, “ah”, “k”, “eh”, and “t”. Our approach includes the wavelet transform, shown in figure 1 **[4].** This figure shows that 1-dimensional signals broken into two signals by low-pass and high-pass filters. The down samplers (shown as an arrow next to the number2) eliminate every other sample, so that the two remaining signals are approximately half the size of the original. As this figure shows, the low-pass (approximate) signal can be further Decomposed, giving a second level of resolution (called an octave). The Number of possible octaves is limited by the size of the original signal, though a number of octaves between 3and 6 is common. Wavelets express signals as sums of wavelets and their dilations and translations. They act in a similar way as Fourier analysis but can approximate signals which contain both large and small features, as well as sharp spikes and discontinuities. This is due to the fact that wavelets do not use a fixed time frequency window. The underlying principle of wavelets is to analyze according to scale.

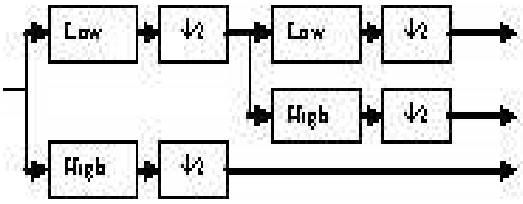


Fig. 1 Discrete Wavelet Transform

III. LITERATURE SURVEY

Many different methods**, algorithms, and mathematical** models have been developed to help with speech analysis and speech recognition. This section points out advances and techniques that have been and are being applied to the speech recognition process.

One method of feature extraction for phoneme recognition proposed by **Long and Dutta [5]** is to transform a signal by choosing the best-suited wavelet basis for the given problem. This is known as Best-basis algorithm and results in adaptive time-scale analysis. The goal is to find a basis, which can most uniquely represent a signal in the presence of other known classes. They used two separate dictionaries as their library of basis, one containing wavelet packets and the other containing smooth localized cosine packets. The most suitable basis is chosen by picking the one that gives minimum entropy from all of the others. Wavelet Packets are a subset of the wavelet transform, and offer greater flexibility for the detection of oscillatory or periodic behavior. The training features for a feed-forward neural network was obtained using the best-basis paradigm, and a dictionary was chosen for each phoneme by a minimum cost function. Five nodes for the neural network classifier were used after they determined that this was a suitable number. Their method was tested on a few phonemes taken from the same user but uttered in different words.

Gouvea et al. [6], design procedures to improve the accuracy of speech recognition systems in noisy environments, as well as normalizing speech signals to account for different speakers. They used recording from the 1995 ARPA Hub 3 which contained recordings of speech in both clean and noisy environments. The 1995 ARPA Hub 3 task was designed to test speech recognition systems for a variety of recording conditioned. Different environments were used for recording as well as different microphones. Initially, signals were classified as clean or noisy using the difference between the minimum and the maximum values of the zeroth order cepstral coefficient. The minimum value of the zeroth-order cepstral coefficient is a measure of the noise in the signal, and the maximum value of the zeroth-order cepstral coefficient is the measure of signal itself. The difference between these is then a measure of the signal to noise ratio. Cepstral coefficients are a result of using the Fourier transform on the spectral magnitudes of the signal. These coefficients are often used as an input to hidden Markov models. The signal classified as “clean” was processed differently from those that were classified as “noisy”. Codebook dependent cepstral normalization was used to attempt to estimate the noise and filter that would best represent the reference static. To help with speaker normalization, a warping function was found by looking the Gaussian mixture models of each speaker compared to a model made for a prototype speaker. An optimal warping function is then found for each speaker using this method**.** Hidden Markov Models were created for a generic speaker based on the optimal warping function. With these techniques, the word recognition rate was lessened, especially for noisy speech.

Hauptman [7] proposed a system to recognize speech that would get its information from closed-captioned television. The television data would be used for training a speech recognition system. Recognizing speech typically involves models for acoustics, language, and pronunciations. The acoustic model often uses neural networks (NN) and/ or Hidden Markov Models. These approaches require accurate training data, generated by the laborious process of humans listening to speech and typing the words. This work is challenging, since transcribers sometimes misspell words, insert extra words, and leave out other words, leading to a word error rate (WER) of 17% for prime time news programs. Other problems with analyzing speech are silences and extraneous noise made by the speaker.

Ganapathiraju et al [8] used a syllable-based system for large vocabulary continuous speech recognition. A large vocabulary is typically larger than 1000 words. Continuous speech is like having a normal conversation. There is no stopping after each sound or word but rather a constant utterance by a user. An example of continuous speech would be dictation where complete sentences and ideas are given without pause. Continuous speech is more difficult to recognize because there are no obvious start and end points of the phoneme or words. The speech recognizer is constantly running, listening for sounds to interpret. A syllable-based system uses a longer time frame, which should model the variations in pronunciations. The performance of this system is compared to using a tri-phone system. The decision to use syllables instead of phonemes is based on the fact that a lot of words tend to run into each other during speech, and a lot of phonemes get deleted when people speak. For example, a sentence starting as “Did you get …” could be heard as the first two words merged into the third word as “jh y u g eh”. Because of this, the syllable may be a more stable unit to work with for speech recognition. The syllable based and tri-phone systems were both based on a standard large vocabulary continuous speech recognition system developed from a commercial package, HTK. HTK stands for Hidden Markov Model Toolkit, and was developed at the Speech Vision and Robotics Group of the Cambridge University Engineering Department. It is portable toolkit for building and manipulating hidden Markov models. This syllable based system did well in recognizing the alphabet but lagged in digit recognition.

IV. SPEECH RECOGNITION

Speech recognition system have three approaches namely:

1. The acoustic-phonetic approach
2. The pattern recognition approach
3. The artificial intelligence approach

The acoustic-phonetic approach:

The acoustic phonetic approach is based on the theory of acoustic phonetics that postulates that there exits finite, distinctive phonetic (phonemes) units in spoken language and that the phonetic units are broadly characterized by a set of properties that are manifest in the speech signal, or its spectrum, over time. Even though the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (the so-called articulation effect), it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned (by a machine).The acoustic-phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling.

Pattern recognition approach:

The pattern recognition approach involves two essential steps – namely, pattern training, and pattern comparison. The essential feature of this approach is that it uses a well-formulated mathematical framework and establishes consistent speech-pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (e.g. a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of the match of the patterns. The pattern matching approach has become the predominant method of speech recognition in the last decade.

Artificial Intelligence approach:

The basic idea of artificial intelligence approach is to compile and incorporate knowledge from variety of sources to realize the different stages of speech recognition system. This approach is a hybrid of the acoustic-phonetic approach and the pattern recognition approach. It exploits the ideas and concepts of both methods and attempts to mechanize the recognition procedure according to the way a person applies his intelligence.

V. DISCRETE WAVELET TRANSFORM

The discrete wavelet transform is a more useful tool for the analysis of non-stationary signals like speech when compared to Fourier Transform. Fourier Transform does not provide any temporal information whereas the wavelet transform, with its flexible time-frequency window, is an appropriate tool for the analysis of signals having both short high frequency bursts and long quasi-stationary components. Discrete Wavelet Transform also provides the added advantage of having a computationally fast algorithm. DWT represents higher frequency components with better time resolution but poor frequency resolution. The lower frequency components are represented with better frequency resolution but poor time resolution. This representation is a good model for the human auditory system that has decreasing frequency resolution for increasing frequencies.

The DWT analysis might be transformed to a fast, pyramidal algorithm for computing the wavelet coefficients which are given by the expressions:

$$\delta_1(k) = \sum_n x(n)g(2k-n)$$

$$\alpha_1(\mathrm{k}) = \sum_n x(n)h(2k-n) \qquad (4.5)$$

Where δ_1(k), and α_1(k) represents the detail and approximation coefficients respectively at level 1 and translation k. The filters g (LPF) and h (HPF) are of length L. The detail coefficients of speech signals are extracted using the above algorithm. The speech signal is subject to the wavelet analysis wherein the DWT detail coefficients are extracted up to eight levels. The energy at each level is computed. Eight features are extracted per frame of the spoken word. A set of these features is called a feature vector. The set of all feature vectors corresponding to all the frames of a word is called a pattern. A database of patterns of all the reference words in the vocabulary is created. In the recognition phase, this pattern is passed to the Dynamic Time Warping for comparison with reference pattern database.

VI. DYNAMIC TIME WRAPING

There is need for time alignment of test and reference frames before pattern matching because different utterances of the same word may be of different durations. So dynamic programming is used to warp the feature vectors of reference speech over the test speech such that there is maximum match. This process is called Dynamic Time Warping.

DTW matches an incoming test word represented by a string of features. The degree of dissimilarity between the two speech frames is the distance between two speech frames is the distance between the two feature vectors extracted from the DWT coefficients corresponding to those frames. This is given by

$$d\ (i,j)=d(a(m),b(m'))$$

Where a(m) and b(m') are the feature vectors extracted from the i^{th} and the j^{th} frame respectively.

VII. METHODOLOGY

For feature extraction, firstly the input analog signal is converted into digital form by A/D converter. Next the function of the pre-emphasizer is to boost the signal spectrum approximately 20 dB per decade. The Digitized speech signal is put through a low-order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The next step in the processing is to windowing each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame and framing. At last step in feature extraction technique, Discrete Wavelet Transform is used to compute the details and approximation coefficients and form the feature vector codebook or pattern database to whom, test pattern is compared. For pattern matching dynamic wrap programming is used to warp the feature vectors of reference speech over the test speech such that there is maximum match. DWT mainly computes the minimum distance between test pattern and stored pattern. Fig 2 explains the above methodology.

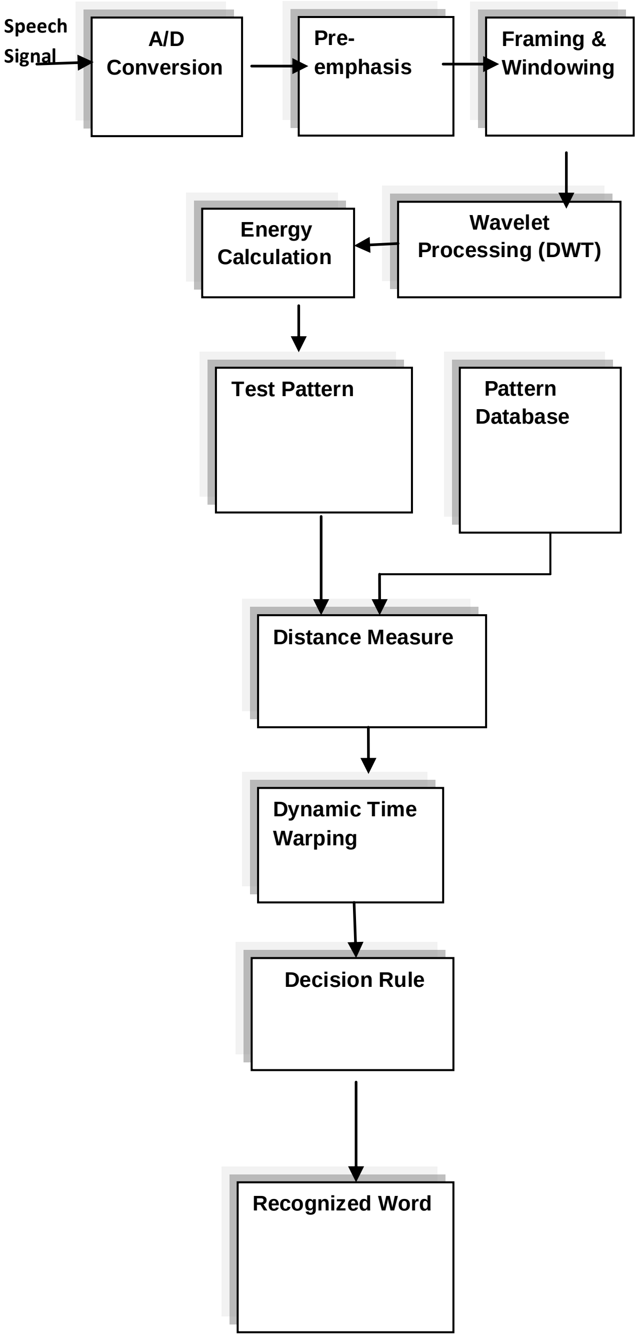


Fig.2 Block diagram of DWT based speech recognition

VIII. RESULTS AND ANALYSIS

The experiments were conducted to extract features by wavelet transform in order to classify spoken words. A template was made which contained the words: These are the words for offline project.

1. Dheeraj
2. Amit
3. Satyender
4. Naveen

5. Vikas

6 Akash

7. Sachin

8. Rakesh

9. Atul

10. Rajesh

Other template was made which contained the words: These are the words for online project.

1. Dheeraj

2. Amit

Dynamic time warping is used to account for signals that have different durations. This is useful for trying to match speech signals because of the stretching and compression of the different phonetic portions of the speech signals. A linear time alignment comparison is not sufficient.

Dynamic time warping finds the minimum distortion between frames of the input signal and the template signal. The Euclidean distance measure is more sensitive to distortions between signals on the time axis. Dynamic time warping tries to take care of this problem by shifting the time axis in order to detect signals that are out of the phase with each other.

Wavelet transform was used for feature extraction on these words. The overall system was divided in to two parts: training phase and recognition phase. The type of wavelet used for both phases is same. Further the frame duration ranging from 18 ms to 32 ms was used for experimentation for both the stages. Then it was tested for speaker dependent and speaker independent. In speaker dependent system, the template included for reference for the same speaker was used. Experiment was done on the speech patterns from the five speakers and in speaker independent system; the patterns of the speaker were not included. Finally the percentage word error rate was calculated for both types of systems. The % word error rate is calculated using the following method.

%word error rate= (no. of mismatch words /total no. of words uttered)*100

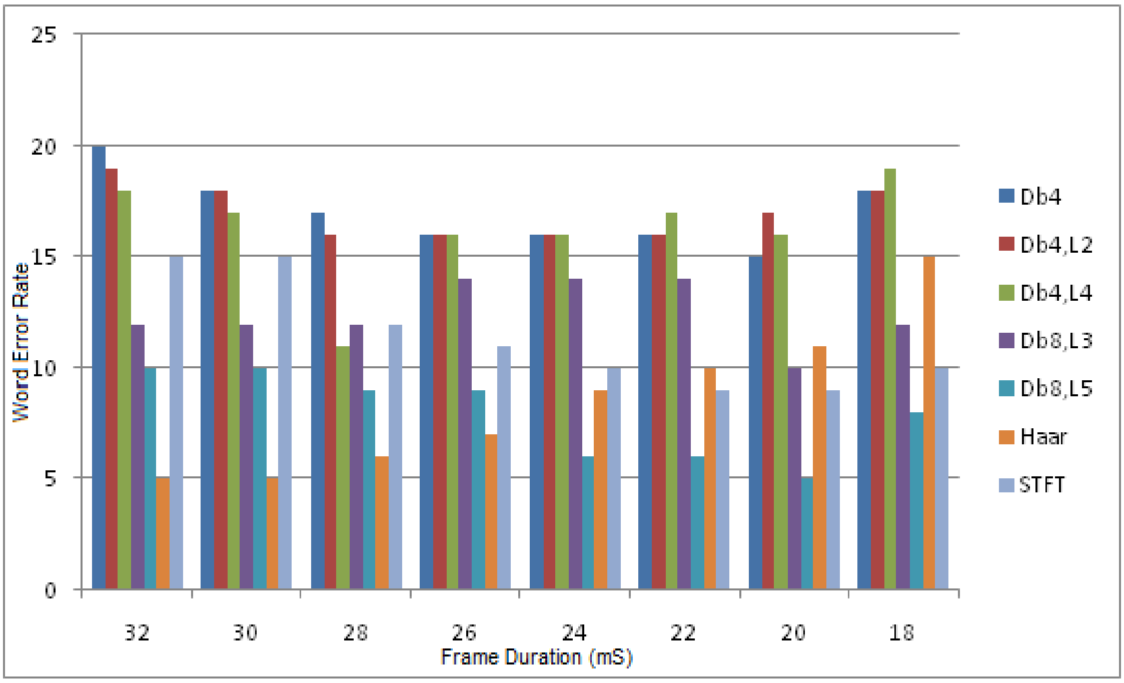
Percentage word error rate is also calculated using short time Fourier Transform for feature extraction.

Frame duration (ms)	32	30	28	26	24	22	20	18
Wavelet Name								
Db4	20	18	17	16	16	16	15	18
Db4,L2	19	18	16	16	16	16	17	18
Db4,L4	18	17	11	16	16	17	16	19
Db8,L3	12	12	12	14	14	14	10	12

Db8,L5	10	10	9	9	6	6	5	8
Haar	5	5	6	7	9	10	11	15
STFT	15	15	12	11	10	9	9	10

Fig.3 Percentage word error rate for speaker dependent system

The above table shows that the % word error rate (5%) is lowest at 20 ms of frame duration for Db8 wavelet and for Haar wavelet has lowest word error rate (5%) is at 32 ms and at 30 ms. In Short Time Fourier Transform Percentage word error rate is minimum at 20 ms (frame duration) which is (9%).



Percentage word error rate for speaker dependent system

IX. CONCLUSION

Speech recognition is the task of extracting features from a speech signal and using a classifying algorithm on these features. The goal is to accurately distinguish any speech signal from other speech signals. The speech recognition process is divided into two phases: the feature extraction stage and classifying stage. During feature extraction stage, features of the speech signal that help in differentiating the signal from others are extracted from the signal and saved. Classifying processes use these features to try to determine what user utters. Wavelets express signals as sum of wavelets and their translation and dilations. They act in much the same way as Fourier analysis but can approximate signal which contain both large and small features, as well as sharp spikes and discontinuities. This is due to the fact that wavelets do not use a fixed time-frequency window. The underlying principle of wavelets is to analyze according to scale. The approach taken in this is to use wavelet transform to extract coefficients from the spoken words and to use dynamic time warping for classifying them as a part of pattern recognition approach. Pre-emphasis is done to boost up the voiced section of the speech signals. Experiments are carried out by using different wavelets at different Frame durations. A template of all words is made to carry out experiments on both the speaker dependent as well as speaker independent systems. The experiments also carried out using short time Fourier transform at feature extraction stage and dynamic time warping for comparison. The results show that using wavelet transform, the lower percentage of word error rate is achieved than using short time Fourier transform. Haar wavelet has shown lower percentage word error rate at 32 ms and at 30 ms of frame duration than STFT for speaker dependent as well as speaker independent systems. Db8 wavelet has shown lower % word error rate at 20 ms of frame durations than STFT for speaker dependent and for speaker independent systems.

X. FUTURE SCOPE

Further improvements can be made by using more complex classification technique such as Artificial Neural Network or Hidden Markov Model and integrating the language model with it.

Frequency components of the signal can be investigated more to look for the possibility of using that content along with the coefficients in classifying the signals. The frequency information could also help distinguish male and female speaker.

REFERENCES

[1] Evandro B. Gouva, Pedro J. Moreno, Bhiksha Raj, Thomas M. Sullivan, and Richard M. Stern, “Adaptation and Compensation: Approaches to Microphone And Speaker Independence in Automatic Speech Recognition,”*Proc. DARPA Speech Recognition Workshop*, February 1996, pages 87-92.

[2] Neal Leavitt, “Let’s Hear It for Audio Mining,” *Computer*, October 2002, pages 23-25.

[3] P.J. Jang and A. G. Hauptmann, “Learning to Recognize Speech by Watching Television,” *IEEE Intelligent Systems*, Volume 14, No. 5, 1999, pp. 51-58.

[4] Amara Graps, “An Introduction to Wavelets,” *IEEE Computational Science and Engineering*, Vol. 2, Num. 2, 1995.

[5] C.J. Long, and S.Dutta, “Wavelet based feature extraction for phoneme recognition” Proceedings of the International conference on spoken language processing, volume 1, October 1996, Pages 264-267.

[6] Evandro B. Gouvea, Petro J. Moreno, Bhiksha Raj, Thomas M. Sullivan and Richard M.Stern, “Adaptation & Compensation: Approaches to microphone and speaker independence in Automatic Speech Recognition”, Proceedings of the Defense Advanced Research Projects Agency Speech Recognition Workshop, Harriman, NY, February 1996, Pages 87-92.

[7] P.J. Jang, and A.G. Hauptmann, “Learning to recognize speech by watching television”, IEEE Intelligent systems, volume 14, No. 5, 1999, pages 51-58.

[8] A. Ganapathiraju, J. Hemaker, M. Ordowski, G.Doddington and J. Picone, “Syllable Based Large Vocabulary Continuous Speech Recognition”, IEEE Trans. on Speech and Audio Processing, Volume 9, No. 4, May 2001, Pages 358-366.