Comparative Study on Preprocessing Techniques on Automatic Speech Recognition for Tamil Language

S.Pannirselvam, Ph.D
Research Supervisor & Head
Department of Computer Science,
Erode Arts & Science College, Erode, Tamil Nadu,
India

G.Balakrishnan
Assistant Professor& Head
Department of Computer Science
Navarasam Arts & Science College for Women,
Erode, Tamil Nadu, India

ABSTRACT

Automatic Speech Recognition (ASR) is a flourishing and swift area for the conversion of acoustic signals acquired from human speech into various other forms such as text, actions, etc., Conversion of Speech To Text (STT) is an incredible and challenging Task. In this paper, we present the study on comparing various digital representations for recording the speech, various pre-emphasis methods for removing the unwanted background noises from the recorded acoustics using suitable filtering techniques. The Filters also help to identify the formant waves for the betterment of syllable and phonetic identification in the subsequent operations for the detection of corresponding alphabetical text on STT Process. This study focuses only on the human speech source as in Tamil which is one among the various Dravidian Languages in India. The connection between oral and written form in Tamil is that individual phonetic segment of the speech denotes individual Tamil alphabets. This feature makes the recognition process as easier and accurate. The detection of location of each phoneme in the speech samples are based on accurate preprocessing outputs of the given speech signal. The last section of this paper shows the experimental results that compare the performance of some of the powerful pre-emphasis methods which are suitable for the Tamil utterance. Finally, we give the suggestions to prefer to use a particular method for the good segmentation.

Keywords

ASR, Tamil Phonemes, Digital Representation, Pre-emphasis.

1. INTRODUCTION

Speech is not only the most natural way of information exchange among human beings, but also it involves in the conversion of the speech to text (STT) process by machine in various applications. Good STT conversion system performs the accurate conversion of word by word speech as well as Continuous speech into text messages. The sources of speech can be in any one or few regional languages. The difficulty in the Automatic Speech Recognition System, in general, is the identification of the boundaries dividing the segments of speech corresponding to each underlying sub-word (syllable) in the sequences which is not known clearly.

The basic Idea for STT conversion process is dividing the overall work into two parts namely Training part and Recognition part. The Training part involves sequences of steps such as recording the sample speech of individual phonemes and individual syllables from a set of people of various age groups. This training is repeated many more times from each individual, removes the noises from each of them, extracts the unique features and store them in the codebook database. The Recognition part makes use of the system as a useful one to generate the text messages for the given speech uttered by any people in an acoustic mode. It includes the works such as digitization of continuous speech in analog wave into the discrete samples, removal of background noises by applying the Filters, identifying the syllable boundaries by applying segmentation, extracting the relevant features and matches the extracted feature sets with the existing codebook database by applying various classification strategies and generates the set of text alphabets for each of the matched coefficient sets.

Hence the good boundary based syllable segmentation relies on noise free samples of the speech signal, it is essential to apply the suitable noise removal method (which is also known as pre-emphasis). Also, the good recognition relies on suitable features which are extracted by applying relevant feature extraction method.

2. CHARACTERISTICS AND LINGUISTIC RULES OF TAMIL

Tamil Language has the feature that is one of the wide spoken (more than 77 million people speakers) languages of the world and also the very ancient and official languages in India. In Tamil Nadu, it is the major and prime language. In Tamil script, there are 12 vowels, 18 consonants and one special character namely, the “Aytam”. In Tamil, a total of 247 characters which are formed from the vowels (12), consonants (18), special character (1) and combined form of vowels and consonants (216). Additionally, six characters from the Grantha script which are adopted from Sanskrit and accepted officially by the Government.

Tamil phonology is attributed by the presence of retroflex consonants and multiple rhotics [5]. In Tamil, there are no distinguished phonological items between voiced and unvoiced consonants; phonetically, voice is assigned based on the position of the consonants in a word. Tamil phonology permits few consonant groups, which cannot be the word initial. Another consonant group known as Nasal Consonants which are uttered differently depends upon the occurrence on the particular environment.

Tamil has very well-built linguistic base with well defined set of morphological syntactic regulations. Generally suffixes are used to mark class, numerals and cases attached to noun or verb root.

The List of ARPABET for the Tamil Vowels and Consonants are given in [11]. Research work on STT in Tamil, although lagging than other languages, is becoming more intensive than before and several researches have been published in the last few years.

3. DIGITAL REPRESENTATION METHODS OF SPEECH

The digitization involves various familiar representations such as 1) Linear Pulse Code Modulation (PCM), 2) Linear Delta Modulation (LDM), 3) Direct Stream Digital (DSD) and hence the selection of optimal representation for the chosen voice signal is essential. Initially the speech signal is digitized as in Fig 1.1

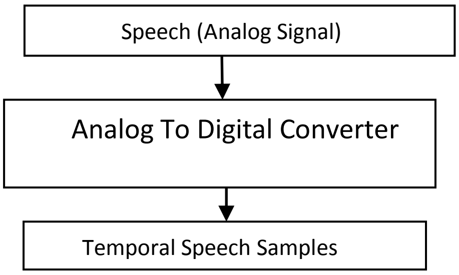

Fig.1.1 Digital Representation-Process

3.1 Pulse Code Modulation

The most basic form of digitized sound is called Pulse Code Modulation, or PCM. It stores an analog signal in discrete time steps with discrete amplitudes. Even though, an analog sound wave is continuous representation, a computer can only store discrete values. The digitization process involves the initial work of recording sound source by a microphone, which induces a current in a wire. Then, an analog to digital converter (ADC) takes the incoming signal and samples it at a given frequency [1]. Each sample is stored as a particular numerical value, and then the next sample is considered in temporal manner. The number of values available can be thought of the quantization of the ADC, which depends on the ADC itself.

For example, an 8-bit ADC can store 2^8=256 different values, and a 16-bit converter can distinguish between 2^{16}= 65536 values. The sound quality of a 16-bit ADC to be far superior to an 8-bit when both are operated at the same sampling frequency.

3.2 Linear Delta Modulation

Linear Delta modulation is a simplified form of pulse code modulation (PCM), which requires a difficult-to-implement analog-to-digital (A/D) converter. The output of a delta modulator is a bit stream of samples, at a relatively high rate (e.g., 100 Kbit/s or more for a speech bandwidth of 4 kHz) the 2 state value of each bit being determined accordingly whether the input message sample amplitude has increased or decreased relative to the previous sample[1]. The operation of a delta modulator is to periodically sample the input message, to make a comparison of the current sample with that preceding it, and to output a single bit which indicates the sign of the difference between the two samples.

It uses a constant step size for all the signal levels. The designing of a linear delta modulator mainly concerns the step size and sampling rate. The signal-to-granular noise ratio must be minimized so that the low level signal can be encoded. The signal-to-slope-overload distortion ratio must be minimized to encode the highest level signals. Minimizing these optimizes the performance of a linear delta modulator. The LDM suffers from a very limited dynamic range due to two types of errors, namely, the slope overload noise and the granular noise.

The experimental analysis showed that this method reduces the quantization noise and can work quite successfully in endpoints detection and pitch extraction for Tamil Language.

3.3 Direct Stream Digital

Digital audio is often stored during production in a single bit format. In addition, the high-end audio distribution format, the single bit recording format known as Direct Stream Digital, or DSD are used as one bit signal throughout the audio recording, editing and playback process. Most analog to digital and digital to analog converters employ a sigma delta modulator that converts a signal to a bit stream [4]. The benefits of the DSD format are numerous. Improvements in the traditional pulse code modulation (PCM) format from higher bit rates and higher sampling rates have experienced diminishing returns. This is partly due to the difficulties in implementing accurate high bit quantisers, but primarily due to the losses incurred from filtering. PCM systems require steep filters at the input to block any signal at or above half the sampling frequency. Ideally, a brick wall filter should be used; passing all frequencies below the Nyquist frequency, and rejecting all above. Yet an ideal brick wall filter does not exist.

There are several features of DSD which distinguish it from PCM. At its heart, DSD is specified as being a 1-bit format, with a sampling rate of 64 * 44.1 KHz, or 2.8224 MHz's.

Unvoiced (silence) signal patterns do not make sense in 44.1 kHz PCM since any repeating pattern would be≤22.05 kHz and hence potentially audible. A constant DC level represents silence in PCM. But for a DSD signal, constant levels (i.e ., all zeroes or all ones) are not allowed. A repeating pattern of 8 bits or less, on the other hand, only has frequency components above 176 kHz, i.e., far outside the range of human hearing. Thus whenever inaudible output is required, a silence pattern should be used. This is important in the construction of many audio effects, such as noise-gating.

4. NOISE AND UNVOISED SIGNALS REDUCTION

Most of the recorded audio samples contain noise due to the disturbances in the recording environment. The noise and the silence (unvoiced signals) in the audio samples should be removed before the features are extracted. Hence, the part of the signals with very high or very low amplitude is removed. The general measure of quality of a noise reduction system is its improvement in signal-to-noise ratio (SNR improvement). SNR can be computed in several ways for a nominally quiet speech signal (known as 'clean' speech) which is combined with noise, depending on whether or not periods of silence are included, and based on the energy balance within the speech samples. The principle of the noise reduction methods is to characterize the in-line noise in the silence intervals and deduct it from the subsequent speech. Essentially, various filter methods can be applied on the spectrum of the speech.

4.1 FIR Filter

FIR filter has been more popularly used due to providing more stable frequency response in terms of sharpness of cut off frequency edge and low ripple than IIR filter. It is of course, the discrete convolution process of set of input signals with specified coefficients and filter order that meet certain specifications. Window Design method is one of the Filter

Design methods which involves the design of ideal FIR filter and then truncates the infinite impulse response by multiplying it with the finite length window function. FIR filters are particularly useful for applications where exact linear phase response is required. One kind of FIR filter known as low-pass FIR Filter can generally be implemented in a sequential way which can guarantee it as a stable filter. The experiments are shown that a low-pass FIR filter is designed using the form

$$y(n)=x(n) - \gamma\ x(n-1) \qquad (1)$$

With γ, constant taking values around 0.95 for speech with a required cut-off frequency of 5 KHz and 300 Hz transition bandwidth and sampling frequency set equal to 10 KHz which relates to the frequency response of human speech signal. Most concentrated energy can be found within frequency range of 0-5KHz or less than that in some speaker.

4.2 Adaptive Filter

IIR filter is a digital filter that provides infinite impulse response. Generally, FIR filter gives the feedback (a recursive part of a filter) and are known as recursive digital filter therefore. This is the main reason that IIR filters have much better frequency response than FIR filters of the same order. Adaptive noise cancellation is an alternative technique of estimating signals corrupted by additive noise or interference. It has an advantage that, with no prior knowledge of signal or noise, levels of noise rejection are obtainable that would be difficult to achieve by other signal processing methods of noise removing.

The principle of adaptive noise cancellation is to obtain an estimate of the noise signal and subtract it from the corrupted signal. The adaptive noise cancellation technique uses adaptive filters for signal processing.

This filter is a popular one in many signal enhancement methods. The basic idea of this filter is to obtain estimate of speech signal from the corrupted signal particularly by additive noise. This estimate is calculated by minimizing the Mean Square Error (MSE) between the desired signal s(n) and the estimated signal ^s(n). It is based on a statistical approach. The Wiener filter weights noisy signal spectrum according to SNR at different frequencies. The main aim of wiener filter is to find out the signal estimate. This signal estimate is calculated by multiplying spectral gain with noisy speech spectrum. This spectral gain depends upon the priori SNR.

4.3 Wavelet Based Filter

In recent years, wavelet transform [7] has become a powerful tool for the multi scale representation of speech and speech signals analysis. Particularly, it localizes the information in the time-frequency plane. Also, it is capable of trading one type of resolution for another that makes them especially suitable for speech signal analysis. A number of researchers have applied this new technique for the reduction of noise in speech signal [8].

For the speech recognition, the mother wavelet is selected by constructing the “Hanning” window. The accuracy of recognition relies on the coverage of the frequency domain. The goal for good speech recognition is to increase the bandwidth of a wavelet without significantly affecting the time resolution. The wavelet transformation can be represented as a tree. The root of the wavelet tree refers the coefficients of wavelet series of the original speech signal. The next level of the tree is the result of one step of the Discrete Wavelet Transform (DWT). Subsequent levels in the tree are formulated by applying the wavelet transform step recursively to split the signal into the low (approximation) and high (detail) parts.

Noise Removal of speech signals using wavelet transform is usually based on thresholding and shrinking wavelet coefficients of noisy signals. However, there are serious problems in choosing the right wavelet, by determining appropriate threshold value and level of decomposition [6].

DWT provides sufficient information both for analysis and synthesis and reduces the computation time sufficiently. It analyzes synthesis and reduces the computation time sufficiently. It analyzes the signal at different frequency bands with different resolutions, and partitions the signal into a coarse approximation and detail information. Our ear has good frequency resolution at low frequencies and lower frequency resolution at high frequencies. Thus the Decomposition of the signal is made by passing time domain signal through low pass and high pass filters.

Thresholding is used in wavelet domain to smooth out or to remove some coefficients of wavelet transform sub signals of the measured signal. This reduces the noises of the speech signal under the non-stationary environment particularly in the speech signals. Thus, Removal of noise components using thresholding of the wavelet coefficients is mainly relying on the observation. In most of the speech signals, the energy is particularly concentrated in a small number of wavelet dimensions. The coefficients of these dimensions are relatively large compared to other dimensions or to any other signal such as noise that has its energy spread over a large number of coefficients. By setting smaller coefficients to zero, we can almost eliminate the noise while preserving the important information of the original speech signal.

For Tamil Utterance, One of the popular wavelet referred as discrete Meyer wavelet or symlets can be chosen as a basis for the DWT because of their time domain symmetrical nature and compact support in the frequency domain.

5. EXPERIMENTAL RESULTS

For selecting the particular pre-emphasis method, the Objective Quality measurement is made based on a mathematical comparison of the original and processed speech signals. It is experimented and reliably reproduced in the MATLAB Environment.

The above 3 types of Filter methods are applied for the 10 Speech Samples which are recorded from 5 age groups [Age Range:7-15,16-25,26-35,36-45 and 46-55] (2 people of each group). For this purpose, the sample utterance "Amma Vanakkam" is sampled and applied to all the 3 Filters.

The following figures (Figure Numbers 2, 3,4 and 5) show the example for comparing the wave forms for one sample utterance (from age group 36-55) applied for various filters.

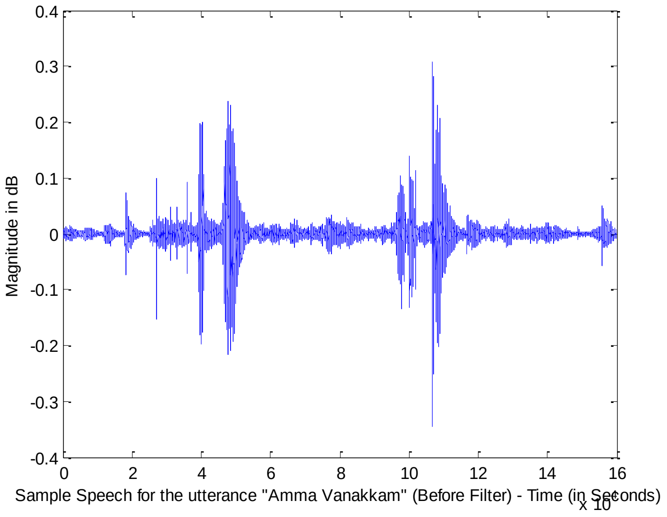

Fig.2 Sample Speech before Applying Filter

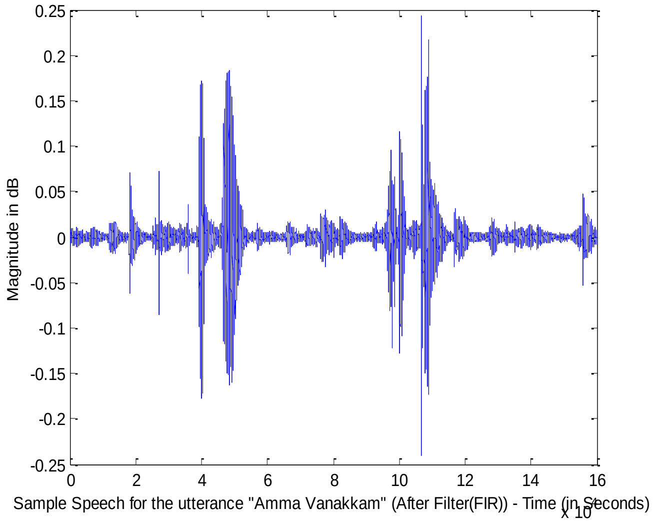

Fig.3 Sample Speech after Applying FIR Filter

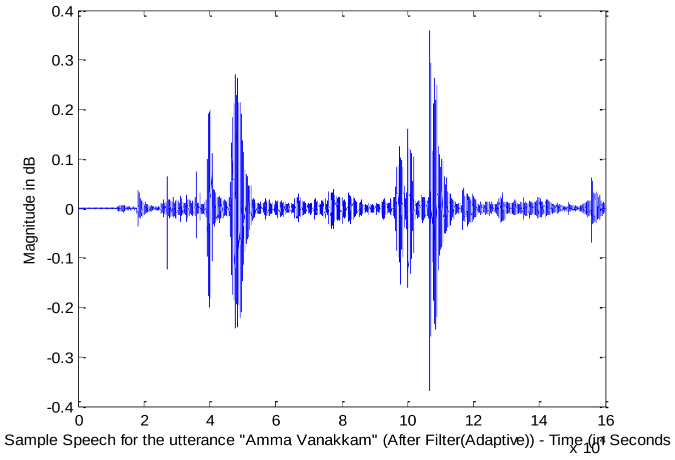

Fig.4 Sample Speech after Applying Adaptive Filter

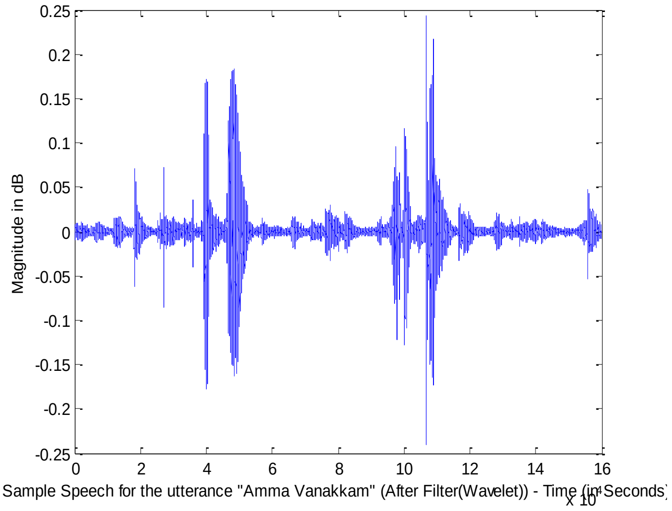

Fig.5 Sample Speech after Applying Wavelet Based Filter

The Peak Amplitude in the speech spectrum corresponds to the high energy spent in the individual phonetics. The higher intensity shows the accumulated energy when the phonetics are uttered continuously without time gap. The long minimal energy path in the speech spectrum indicates the occurrence of the word segmentation. These features are utilized further phases like windowing and Framing in the Entire Speech Recognition Task.

5.1 Comparison of Error Rate

For comparing the performance among the 3 Filters, the Mean square error which is calculated by

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=0}^{N}[\mathrm{x}(\mathrm{k}) - \mathrm{y}(\mathrm{k})]^2 \qquad (2)$$

Where N is the total number of samples in the Speech Signal, x denotes the original speech and y denotes the filtered speech. The Percentage of MSE for the 3 Filters for all the set of 10 Speech samples is tabulated in Table 1.

Thus the MSE assesses the quality of a filtered speech signal in terms of its variation and degree of bias. When the MSE is very less, it is treated as better quality.

5.2 Peak Signal to Noise Ratio (PSNR)

The performance is evaluated not only using the Mean Square Error (MSE) but also using Peak Signal to Noise Ratio (PSNR) in order to evaluate the corresponding speech signal quality.

The Peak Signal to Noise Ratio is calculated by:

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \qquad (3)$$

For the signal quality measures, if the value of the PSNR is very high for the sample speech of a particular noise type then it is considered as best quality signal.

Table 2 shows the experimented values obtained for those methods using PSNR. By the analysis of the values in the table, the wavelet based filter is better with less MSE and high PSNR values.

Table 1. MSE Comparison

Speech Sample No	FIR Filter	Adaptive Filter	Wavelet Based Filter
1	53.55	50.27	40.23
2	42.25	40.59	30.05
3	34.78	33.38	25.82
4	30.85	30.60	20.24
5	42.45	41.25	36.44
6	44.23	42.33	32.69
7	47.29	46.62	40.36
8	43.75	40.01	34.54
9	44.55	39.60	31.17
10	48.05	44.42	37.25

Table 2. PSNR Comparison

Speech Sample No	FIR Filter	Adaptive Filter	Wavelet Based Filter
1	30.84	31.11	32.08
2	31.87	32.04	33.35
3	32.71	32.89	34.01
4	33.23	33.27	35.06
5	31.85	31.97	32.51
6	31.67	31.86	32.98
7	31.38	31.44	32.07
8	31.72	32.10	32.74
9	31.64	32.15	33.19
10	31.31	31.65	32.41

6. CONCLUSION

In this paper, the study is made on various preprocessing methods on STT and it is observed that for Tamil speech recognition system, the syllables or sub-words can be segmented from voiced part of the signal and not on the frame of the signal. For choosing the Digital Recording Method, we suggest to use LDM method because of the noise distortion is minimal one comparing with other two techniques. Moreover, based on the Experimental Result Analysis, it is concluded that the wavelet based noise removal technique is preferable for the STT work on chosen language.

6.1 Acknowledgements

The authors would like to thank the anonymous reviewers for their comments which were very helpful in improving the quality and presentation of this paper.

7. REFERENCES

[1] Ronald.W.Schafer, Senior Member, IEEE, & Lawrence.R.Wbiner, Member, IEEE, “Digital Representations of Speech Signals”, PROCEE-DINGS OF THE IEEE, vol. 63, no. 4, 1975.

[2] G.Lakshmi Sarada, A Lakshmi, Hema A Murthy and T Nagarajan, "Automatic transcription of continuous speech into syllable-like units for Indian languages",Sadhana Vol. 34, Part 2, pp. 221–233,2009.

[3] Abhishek Nandy," Pitch Detection of Speech Synthesis by Using Matlab",IOSR-JECE Volume 8, Issue 1, 2013.

[4] Josh Reiss and Mark Sandler “Digital audio effects applied directly on a dsd bitstream”, Proc. of the 7th Int. Conference on Digital Audio Effects, 2004.

[5] Sarada, G. L., Nagarajan, T., Hema A. Murthy., ”MultipleFrame Size And Multiple Frame Rate Feature Extraction For Speech Recognition”, SPCOM-2004, 2004.

[6] Milind Kansara and Prof. Neeta Chapatwala “Noise Reduction from the Speech Signal using Wavelet Packet Transform” IJECSE,Volume2, Number 2, 2013.

[7] Md. Mijanur Rahman,& Md. Al-Amin Bhuiyan "Dynamic Thresholding On Speech Segmentation",IJRET, Volume: 02 Issue: 09, 2013.

[8] A.Montazeri,M.H.Kahaei &J.Postan,"A New Stable Adaptive IIR Filter For Active Noise Control Systems",2011

[9] S.Jothilakshmi,S.Sindhuja &V.Ramilingam "Dravidian–Tamil Tts For Interactive Voice Response System",IJIRD, Vol 2,Issue 4, 2013.

[10] Hanitha Gnanathesigar,"Tamil Speech Recognition using Semi Continuous Models",IJSRP,Vol 2, Issue 6, 2012.

[11] Hanitha Gnanathesigar,"Tamil Speech Recognition using Semi Continuous Models",IJSRP,Vol 2 Issue 6 June 2012.