PREPROCESSING TECHNIQUE IN AUTOMATIC SPEECH RECOGNITION FOR HUMAN COMPUTER INTERACTION: AN OVERVIEW

Yakubu A. Ibrahim [1], Juliet C. Odiketa [2], Tunji S. Ibiyemi [3]

[1] Department of Computer Science, Bingham University, Karu, Nigeria
[2] Department of Computer Science, The Federal Polytechnic Idah, Idah, Nigeria
[3] Department of Electrical Engineering, University of Ilorin, Ilorin, Nigeria

Corresponding Author: Yakubu A. Ibrahim, talktoibro80@gmail.com

ABSTRACT: Automatic Speech Recognition has found its application on various aspects of our daily lives such as automatic phone answering service, dictating text and issuing voice commands to computers. Speech recognition is one of the fastest developing fields in the framework of speech science and engineering. Also, in computing technology, it comes as the next major innovation in human computer interaction. However, in speech signal processing, Pre-processing of speech plays a vital role in development of an efficient automatic speech recognition system. Nowadays, Humans are able to interact with computer hardware and other machines through human language. In view of the above, researchers are putting efforts to develop a perfect and efficient speech recognition system but machines are unable to match the performance of human utterances in terms of accuracy of matching and speed of response. Therefore, preprocessing of signal is based on number of applications and drawback of the available techniques of ASR systems. Hence, the process of preprocessing in speech recognition discussed in the study includes: Noise removal, Voice Activity Detection, Pre-emphasis, Framing and Windowing.

KEYWORDS: Automatic Speech Recognition (ASR), Human Computer Interaction (HCI), Pre-processing.

I INTRODUCTION

Speech is the most natural form of human-to-human communications and is related to human physiological capability. It is the most important, effective and convenient form of information exchange. Speech processing is a complete subject and a popular research field, which involves a wide range of content ([ZB15]). In Automatic Speech Recognition system the first phase is pre-processing phase. Moreover, Pre-Processing of Speech is very important in the applications where silence or ambient noise is completely undesirable. Voice activity detection is a well known technique adopted for many years in preprocessing of speech signal, Noise canceling, pre-emphasis and dimensionality reduction of speech facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced ([D+00]) sounds finds other applications mainly in Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, Stop Consonant Identification and End Point Detection for isolated utterances. There are several ways of classifying (labeling) events in speech. It is accepted convention to use a three-state representation in which states are (i) silence (S), where no speech is produced; (ii) unvoiced (U), in which the vocal cords ([AR76]) are not vibrating, so the resulting speech waveform is a periodic or random in nature and (iii) voiced (V), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic ([CHL89]).

II PREPROCESSING

In development of an ASR system, preprocessing is considered the first phase of other phases in speech recognition to differentiate the voiced or unvoiced signal and create feature vectors. Preprocessing adjusts or modifies the speech signal, x(n), so that it will be more acceptable for feature extraction analysis. The major factor to consider when it comes to speech signal processing is to check the speech, x(n) if is corrupted by some background or ambient noise, d(n), for example as additive disturbance

$$x(n) = s(n) + d(n) \quad (1)$$

Where s(n) is the clean speech signal. In noise reduction, there are different methods that can be adopted to perform the task on a noisy speech signal. However, to develop perfect speech recognition system, the two frequently used methods of noise reduction algorithms in speech recognition system is spectral subtraction and adaptive noise cancellation ([D+00]).

(a) BACKGROUND/AMBIENT NOISE REMOVAL

The ability to detect the useful parts of a speech signal from stream of signals can be of high importance during the initial processing stages of an audio analysis system process. Ambient noise is any signal other than the signal being monitored. It is a form of noise pollution or interference. As a matter of fact, background noise is an important concept in setting noise levels in ASR systems. The performance measure of speech recognition systems degrades drastically when training and testing data are carried out with different noise levels. Signal-to-Noise Ratio (SNR) is the ratio of the power of the correct signal to the noise ([ZM04]). SNR is usually measured in *decibels* (dB).

$$SNR = 20\log_{10}\frac{V_{signal}}{V_{noise}} \quad (2)$$

Where V_{signal} is the voltage of correct signal, V_{noise} is the voltage of the noise. Background or ambient noise is normally produced by sounds of air conditioning system, fans, fluorescent lamps, type writers, computer systems, back conversation, footsteps, traffic noise, alarms, bird's noise, opening and closing of doors. The developers of ASR system usually have little control over these noises in the real life environments. Every noise is additive in nature and usually steady state except for impulse noise sources like type writers ([HC14]). In training and testing stage, the frequently used method to reduce the effect the ambient noise on speech recognition is to use a close-talk microphone. When a speaker is generating speech utterance at normal communication level, the average signal to noise ratio (speech level) increase by about 3dB any time the microphone is filtering the speech utterance. The filter adopted to remove the background or ambient noise is as follows ([JMR94]):

$$E_s = 10 * \log_{10}\left[\epsilon + \frac{1}{N}\sum_{n=1}^{N} S^2(n)\right] \quad (3)$$

Where, the E_s is log energy of block of N samples and ϵ is a small positive constant added to prevent the computing of log zero. *S(n)* be the n[th] speech sample in the block of N samples.

(b) VOICE ACTIVITY DETECTION /SPEECH WORD DETECTION

The major issue of getting or locating the endpoints of a signal in a speech is a main problem for the speech recognizer. Inaccurate endpoint detection will decrease the performance of the speech recognizer. However, in detecting endpoints of a speech utterance, it seems to be relatively trivial, and has been found to be very difficult in practice in speech recognition systems. When a proper SNR is given, the work of developing ASR system is made easier. V*oice activity detectors* (VAD) are devices used to divide the speech signal into voiced or unvoiced, *speech segments* and *non-speech segments*. Non-speech or unvoiced parts of speech utterance are pre-utterance, post-utterance and between words silences. Although, methods or algorithms to detect automatically non-speech parts of utterance are necessary for a wide range of applications like speech coding, speech recognition, speech enhancement, etc. In the case of estimation of noise characteristics during nonspeech segments, VADs have to adapt to the changes of the noise characteristics ([MJR92]). Robustness against noise variations is difficult to obtain. Unvoiced segments of the speech signal are more difficult to detect than voiced segments, because they are more similar to the noise and the SNR is generally lower in unvoiced than in voiced segments. Speech recognition adopts the following commonly used techniques for finding VAD in speech recognition are as follows:

1. THE ZERO-CROSSING RATE

ZCR of a speech signal frame is the rate of sign-changes of the signal during the frame. In other words, it is the number of times the signal changes value, from positive to negative and vice versa, divided by the length of the frame. The ZCR is defined according to the following equation:

$$Z(i) = \frac{1}{2N}\sum_{n=1}^{N}|\, sgn[x_i(n)] - sgn[x_i(n-1)]| \quad (4)$$

Where *sgn()* is the sign function, that is

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n)\geq 0, \\ -1, & x_i(n)<0, \end{cases} \quad (5)$$

ZCR is used to discern unvoiced speech. Usually unvoiced speech has a low short-term energy but a high ZCR.

2. ENERGY (ENTROPY OF ENERGY)

Let $x_i(n)$, $n = 1, \ldots, N$ be the sequence of audio samples of the i^{th} frame, where W_L is the length of the frame. The short-term energy is computed according to the equation ([TA14]):

$$E(i) = \sum_{n=1}^{W_L}[x_i(n)]^2 \quad (6)$$

Usually, energy is normalized by dividing it with W_L to remove the dependency on the frame length. Therefore, Equation (5) which provides the so called power of signal becomes:

$$E(i) = \frac{1}{W_L}\sum_{n=1}^{W_L}[x_i(n)]^2 \quad (7)$$

It is observed that short-term energy is the most effective energy parameter for VAD. Speech signal has most of its energy collected in the lower frequencies, whereas most energy of the unvoiced speech exists in the higher frequencies ([L+81]). The short-term entropy of energy can be defined as a calculation of unexpected changes in the energy level of a speech signal. To compute it, divide every short-term frame in *K* sub-frames of fixed duration by short-term frame in K short-frames. Then, for each $E_{sub\text{-}frame}$, j, its energy is computed as in Equation (6) and divides it by the total energy, $E_{shortFrame,}$ i, of the short-term frame. The division operation is a standard procedure and serves as the means to treat the resulting sequence of sub-frame energy values, e_j, j = 1, . . . ,*K*, as a sequence of probabilities, as in Equation (7) ([TA14]):

$$e_j = \frac{E_{subFrame}}{E_{shortFrame}} \quad (8)$$

Where

$$E_{shortFrame} = \sum_{k=1}^{K} E_{subFrame} \quad (9)$$

At a final step, the entropy, *H(i)* of the sequence e_j is computed according to the equation:

$$H(i) = -\sum_{j=1}^{K} e_j * \log_2(e_j) \quad (10)$$

The resulting value is lower if abrupt changes in the energy envelope of the frame exist. This is because, if a sub-frame yields a high energy value, then one of the resulting probabilities will be high, which in turn reduces the entropy of sequence e_j .

3. THE AUTOCORRELATION FUNCTION

It allows computing the correlation of a signal with itself as a function of time.
Normalization auto-correlation coefficient at unit sample delay C_1 is defined ([BVN12]):

$$C_1 = \frac{\sum_{n=1}^{N} S(n)s(n-1)}{\sqrt{[\sum_{n=1}^{N} S^2(n)][\sum_{n=0}^{N-1} S^2(n)]}} \quad (11)$$

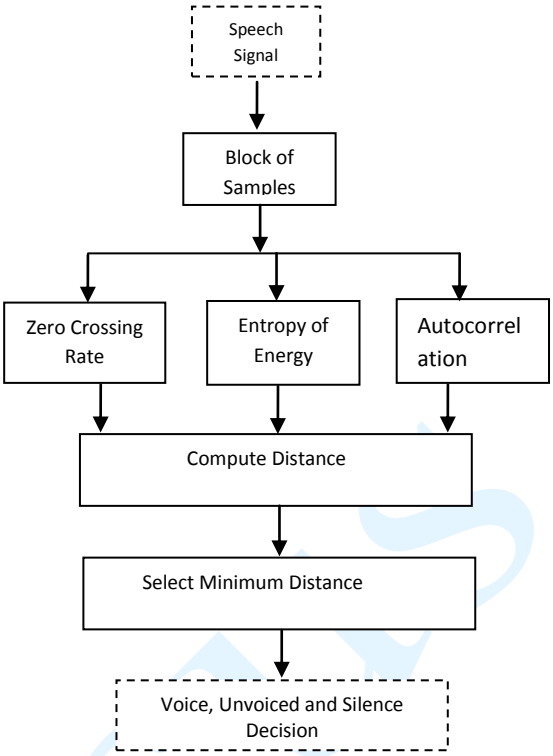

Figure 1: Block diagram of end point detection ([BVN12])

(ii) FILTER FOR END POINT DETECTION

Filters are widely employed in signal processing and communication systems in applications such as channel equalization, noise reduction, radar, audio processing, video processing, biomedical signal processing, and analysis of economic and financial data.
The essence of the filter can also be defined as a process of flattening, where the spectrum is whitened. It is believed that a speech may have diverse components separated by some pauses. Every component can be determined by detecting a two of endpoints named component beginning and ending points. In the energy contours of speech, there is always a higher edge following a beginning point and a lowering edge preceding an ending point ([BR04]). These points are known as beginning and ending edges of the speech signal. However, to be certain that the low-complexity, short-term energy is adopted in the cepstral feature to be the feature for endpoint detection. The energy filter is given as:

$$E(L) = 10\log 10\sum_{j=n(L)}^{n(L)+l-1} o(j)^2 \quad (12)$$

Where, *o(j)* is data sample, *L* is frame number, *'l'* is window length, *E(L)* is frame energy in decibel, *n(L)* is number of first data sample in the window.
Thus, the detected endpoints can be aligned to the ASR feature vector automatically and the computation can be reduced from the speech-

sampling rate to the frame rate. For correct and effective endpoint detection, we need a good detector that can detect all available endpoints from the energy feature. Since the output of the detector may contain false acceptances, a decision module is then required to make final decisions based on the detection output. However, since endpoints detection always comes with the edges, the intention is to detect the edges first and thereafter to find the corresponding endpoints ([RS78]).

(iii) ENERGY NORMALIZATION

At this stage, the aim of normalization of energy is to normalize the speech energy E(*l)*. The normalization of energy is performed by finding the maximum energy value E_{max} over the spoken words as:

$$E_{max} = max(E_l),\ 1 \leq l \leq L \quad (13)$$

By subtracting E_{max} from E*l* to give

$$\hat{E}(l) = E(l) - E_{max} \quad (14)$$

In this way the peak energy value of each word is zero decibels and the recognition system is relatively insensitive to the difference in gain between different recordings. In performing the above calculations, there is constraints that word energy contour normalization cannot take place until the end of the word is located ([Kul84]).

(c) PRE-EMPHASIS

A spoken audio signal may have frequency components that fall off at high frequencies. As a matter of fact, in some systems such as speech coding, to avoid overlooking the high frequencies, the high-frequency components are compensated using pre-emphasis filtering ([Pic93]). Pre-emphasis is therefore, aimed at compensating for lip radiation and necessary attenuation of high frequencies in the sampling process. High frequency components are emphasized and low frequency components are attenuated. This is quite a standard preprocessing step. The digitized speech waveform has a high dynamic range and suffers from additive noise to reduce this range pre-emphasis is applied. By pre-emphasis, we imply the application of a high pass filter, which is usually a first-order FIR of the form ([Q+07]):

$$H(z) = \sum_{k=0}^{N} \propto (k) z^{-k} \quad (15)$$

Normally, a single coefficient filter digital filter known as preemphasis filter is used:

$$H(z) = 1 - \propto z^{-1} \quad (16)$$

Where the preemphasis factor α is computed as:

$$\propto = exp(-2\pi F \Delta t) \quad (17)$$

Where F is the spectral slope will increase by 6dB/octave and is the sampling period of the sound. The pre-emphasis factor is chosen as a trade-off between vowel and consonants discrimination capability ([SS11]).

The usual form for the pre-emphasis filter is a high-pass finite impulse response (FIR) filter with a single zero near the origin. It intends to whiten the speech signal spectrum as well as emphasizing those frequencies at which the human auditory system is most sensitive. However, for human ear, this is only suitable at 3 to 4 kHz. Above this range, the sensitivity of human hearing falls off, and there is relatively little linguistic information. Therefore, it is appropriate to adopt a second order pre-emphasis filter. This causes the frequency response to roll off at higher frequencies. This becomes very important in the presence of noise. The pre-emphasizer is used to spectrally flatten the speech signal. This is usually done by a high pass filter. The most frequently adopted filter for this phase is the FIR filter. Typically, the speech signal produced by human being has a spectral slope of approximately-6dB for voiced sounds. The slope is because of two major reasons namely: *(a)* the shape of the glottal pulse introduces a slope of - 12dB and *(b)* The lip radiation introduces a slope of +dB. Therefore, the resultant slope of approximately -6dB exists in the recorded voiced speech sounds. Pre-emphasis is performed to remove this slope of -6 dB.

To accomplish the task, the speech signal is passed through a high-pass finite impulse response (FIR) filter of order 1. The pre-emphasis is defined by ([Kul84]):

$$y[n] = s[n] - Pxs[n-1] \quad (18)$$

Where, *s[n]* is the nth speech sample, *y[n]* is the corresponding pre-emphasized sample and P is the pre-emphasis factor typically having a value between 0:9 and 1. Pre-emphasis ensures that in the frequency domain all the formats of the speech signal have similar amplitude so that they get equal importance in subsequent processing stages ([D+00]). In the frequency domain, it looks like:

$$H(z) = 1 - \propto z^{-1} \quad (19)$$

(d) FRAMING OR FRAME BLOCKING

Framing is the process of breaking the continuous stream of speech samples into components of constant length to facilitate block-wise processing of the signal. In the same vein, speech can be thought

of been a quasi-stationary signal and is stationary only for a short period of time ([BVN12]). As a result, the speech signal is slowly varying over time (quasi-stationary) that is when the signal is examined over a short period of time (5-100msec), the signal is fairly stationary. Therefore, speech signals are often analyzed in short time components, which are sometimes referred to as short-time spectral analysis in speech processing.
This simply means that the signal is divided or blocked in to frames of typically 20-30 msec. In this aspect, adjacent frames normally overlap each other with 30-50%, this is done in order not to lose any vital information of the speech signal due to the windowing.

(e) WINDOWING

At this stage the signal has been framed into segments, each frame is multiplied with a window function w(n) with length N, where N is the length of the frame. Windowing is the process of multiplying a waveform of speech signal segment by a time window of given shape, to stress pre-defined characteristics of the signal. To reduce the discontinuity of speech signal at the beginning and end of each frame, the signal should be tapered to zero or close to zero, and hence minimize the mismatch. Moreover, this can be arrived at by windowing each frame of the signal to increase the correlation of the Mel Frequency Cepstrum Coefficients (MFCC) and spectral estimates between consecutive frames ([BVN12]). ASR system designers have always had to solve for an issue of a compromise in their selection of analysis window. To obtain good frequency resolution, a long window is desirable but the linguistic importance of some short transients makes a short window desirable and effective. The normal compromise that is always available to settle for is the frame lengths of about 20 or 30 ms, with a frame spacing of 5 to 10 ms. On the other hand, a shorter window is always adequate to capture the salient spectral features, given that the frame spacing is also sufficiently short enough. An eight (8) ms window, with two (2) ms frame spacing is always adopted. However, when the feature curves are represented as described in the following subsections, the frequency resolution appears to be very similar to that obtained with the longer window. The windowing is always performed to a speech signal to avoid problems due to truncation of the signal as windowing helps in the smoothing of the signal ([ZM04]).
The proper selection in the choice of window *w(n)* is a grade-off between different factors: (i) The shape of the window may reduce differences, but it may increase signal shape alteration. The length is proportional to the frequency resolution and inversely proportional to the time resolution. (ii) The signal overlap is proportional to the frame rate, but it is also proportional to the correlation of subsequent frames.
Where w(n) designates the window function. Types of common window functions used in FIR filter design for speech are given below:

(i) Rectangular window:
w(n) $w(n) = 1,\ 0 \le n \le M-1$ (20)

(ii) Triangular window:
$w(n) = 1 - \left[1 - \frac{2n}{M-1}\right],\ 0 \le n \le M-1$ (21)

(iii) Hanning window:
$w(n) = 0.5 - 0.5\cos\left(\frac{2n\pi}{M-1}\right),\ 0 \le n \le M-1$ (22)

(iv) Hamming window:
$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{M-1}\right),\ 0 \le n \le M-1$ (23)

(v) Bartlett window:
$w(n) = 0.42 - 0.5\cos\left(\frac{2n\pi}{M-1}\right) + 0.08\cos\left(\frac{4n\pi}{M-1}\right),\ 0 \le n \le M-1$ (24)

III CONCLUSION

The study of preprocessing has been carried out to develop a speech recognition based for human computer interaction system. This system can be used in various applications related with disable persons those are unable to operate computer through keyboard and mouse, these type of persons can use computer with the use of automatic speech recognition system, with this system user can operate computer with speech commands so extra advantages of human computer interaction will be that if any disable person is using this system he/she feels that he/she is working in real environment as what they want to do. Also, the application is available for those computer users which are not comfortable with English language or any of the available international language but feel good to work with their native language such as Hausa language.

REFERENCES

[AR76] **B. Atal, L. Rabiner** - *A pattern recognition approach to voiced unvoiced- silence classification with applications to speech recognition* Acoustics, Speech, and Signal Processing [see also IEEE transactions on Signal Processing], Vol. 24, pp. 201-212, 1976.

[BR04] **C. Becchetti, L. Ricotti** - *Speech Recognition Theory and C++ Implementation*, John Wiley & Sons, Wiley Student Edition, Singapure, pp. 121-188, 2004.

[BVN12] **S. Bhupinder, R. Vanita, M. Namisha** - *Preprocessing in ASR for Computer Machine Interaction with Humans: A Review*, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2 pp 396-399, 2012.

[CHL89] **D. G. Childers, M. Hand, M. J. Larar** - *Silent and Voiced/Unvoiced/Mixed Excitation (Four Way), Classification of Speech*, IEEE Trans. On ASSP, Vol. 37, 11, Nov 1989, pp1771-74, 1989.

[D+00] **J. R. Deller, J. L. Hanse, J. G. Proakis** - *Discrete-Time Processing of speech signals*. IEEE Press, ISBN 0-7803-5386-2, 2000.

[HC04] **T. Hwang, S. Chang** - *Energy Contour enhancement for noisy speech recognition*, International Symposium on Chinese Spoken Language Processing, Vol. 1, pp. 249-252, 2004.

[JMR94] **J.-C. Junqua, B. Mak, B. Reaves** - *A robust algorithm for word boundary detection in presence of noise*, IEEE Trans. on Speech and Audio Processing, Vol. 2, pp. 406– 412, 1994.

[Kul84] **K. P. Kuldip** - *Effect of Pre-emphasis on Vowel Recognition Performance*, speech communication *3* (pg.101-106), North-Holland, 1984.

[L+81] **L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon** - *An improved endpoint detector for isolated word recognition*, IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 29, pp. 777–785, 1981.

[MJR92] **B. Mak, J.-C. Junqua, B. Reaves** - *A robust speech/non speech detection algorithm using time and frequency-based features*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. I, pp. 269–272, 1992.

[Pic93] **L. Picone** - *Signal modeling technique in Speech Recognition*, IEEE ASSP Magazine, Vol. 81, Issue 9, pp. 1215-1247, 1993.

[RS78] **L. R. Rabiner, R. W. Schafer** - *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey, Prentice Hall, 512-ISBN-13:9780132136037, 1978.

[Q+07] **L. Qi, Z. Jinsong, T. Augustine, Z. Qiru** - *Robust Endpoint Detection and Energy Normalization for real-Time Speech Recognition and Speaker Recognition.* IEEE Transactions. On Speech and audio processing. Vol. 10 no 3 pp. 146–157, 2007.

[SS11] **B. Singh, P. Singh** - *Voice Based user Machine Interface for Punjabi using Hidden Markov Model*, in the proceeding of IJCST Vol. 2, Issue 3, pp. 222-224, 2011.

[TA14] **G. Theodoros, P. Aggelos** - *Introduction to Audio Analysis: A MATLAB Approach*, Elsevier Academic Press USA, pp 77-110, 2014.

[ZB15] **G. Zhuo, W. D. Bian-Ba** - *A Study of Tibetan Speech Pitch Detection Algorithm Based on Matlab.* Modern Electronics Technique, 10, 20-22, 2015.

[ZM04] **L. Ze-nian, S. D. Mark** - *Fundamentals of multimedia*, Pearson Pretence Hall Press, USA, pp 130-140, 2004.