International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Word Detection and Global SNR Estimation of Speech Signals

Sreelakshmi M.S.[1], Ann Nita Netto[2]

[1]M .Tech student, Sree Buddha College of Engineering, Elavumthitta, Kerala, India

[2]Assistant Professor, Department of Electronics and Communication Engineering, Sree Buddha College of Engineering, Elavumthitta, Kerala, India

Abstract: ***Speech recognition systems consists of machines or devices which are capable of receving speech as input and detecting the words or phrases of the input speech, converting the speech in to machine readable format (speech to text format) etc. But the introduction of noise to these systems may affect the performance of the system. The estimation of signal-to-noise ratio (SNR) gives an idea about the amount of noise present in the original signal. SNR compares the level of the data signal to the level of background noise. A novel method based on certain features of the speech signal is used for estimating the SNR. Combining the word detection scheme and estimation of global SNR improves the performance of such systems.***

Keywords: Speech processing, MFCC, Signal-to-noise ratio, DNN, Word detection

1. Introduction

A word detection scheme is introduced along with estimation of global SNR. Word detection is used in several applications like mobile phones, security related applications etc. Combining the global SNR estimation from the signal features and word detection scheme can improve the performance of the system in such applications. Speech processing is defined as a technique used for the study of speech signals and the processing methods of these signals. But due to several factors like environmental conditions and channel properties noise is added to the speech signals, which in turn reduces the signal performances. Signal to noise ratio (SNR) gives an idea about the amount of noise present in the original signal. SNR compares the level of the desired input signal to the background noise level. SNR estimation algorithms is of two types. Instantaneous SNR and global SNR. Here we are estimating the global SNR of our speech signal. Instantaneous SNR focus on the frame of the original signal while global SNR focus on the entire signal. In[12] M.Vondraseketal. presented the algorithms for speech SNR estimation and the tool SNR where these methods are implemented. The definitions of SNR optimized for speech application are summarized and implemented in above mentioned tool. The described tool can estimate the SNR of speech signal containing noise with or without reference signal. The tool can be used to create a speech and noise mixture with required SNR level. Mel frequency cepstrum coefficients (MFCC) feature extraction is included to obtain better performance. MFCCs are one of the most popular feature extraction techniques used in speech recognition. Mel-frequency Cepstral Coefficients (MFCC) is used for feature extraction[2]. A speech waveform is used as an input to the feature extraction module. The efficiency of this phase is important for the next phase since it affects the modeling process. The most dominant method used to extract spectral features is the calculation of Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition. MFCCs being considered as frequency domain features are more accurate than time domain features [4].

A deep neural networking is used to match the noise type. Our proposed method is based on certain signal features like long-term signal energy, signal variability, pitch and voicing probability. A regression model is build based on these features. Regression analysis is a process of estimating the relationship among variables.

In section 2 we describe the related works and in section 3 we present the MFCC feature extraction technique. In section 4 the different signal features are explained. In section 5 we describe our experimental results. Finally in section 5 we present our conclusions and future work .

2. Method Overview

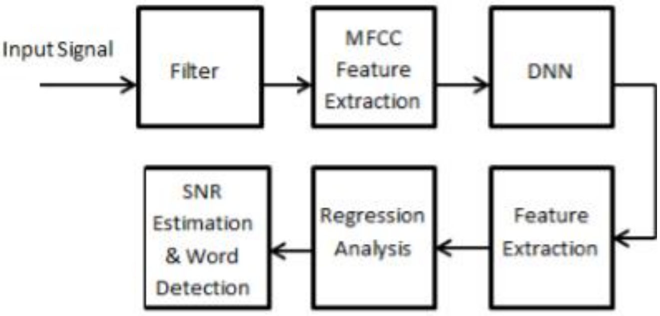


Figure 1: Method Overview

The proposed system consist of a word detection scheme along with SNR estimation. The input speech signal is first filtered to remove noise. Then Mel-frequency cepstrum coefficients (MFCC) extraction technique is performed on the filtered signal. MFCC feature extraction helps to effectively remove background noise from the speech signal. The output of the MFCC feature extracted system is given to a deep neural network. Neural networks helps to cluster and classify. Since the type of noise that corrupts the signal is unknown, deep neural network helps to find the closest noise type. Finding the noise type is important because regression analysis for different noise types. After finding the noise type different signal features like long-term signal energy, signal variability, pitch and voicing probability is extracted from the signal. In order to estimate the global SNR from the features a regression model is build. Regression model helps to relate the final SNR to the extracted signal features. Combining this technique with word detection scheme helps to estimate the SNR level in the input speech signal. Based on the SNR level, techniques can be added to the proposed system to improve the intelligibility of the speech signal and thereby the performance of a word detection system can be improved.

3. MFCC Feature Extraction

Feature extraction is an important technique in speech processing. The main objective of feature extraction is to find robust and discriminative features in the sound signal. Here, Mel Frequency Cepstral Coefficients (MFCC) feature extraction technique is used to extract useful features from the input speech signal. The main aim of feature extraction is to calculate a sequence of feature vectors providing a compact representation of the input signal. The input signals are passed through the feature extraction, feature training & feature testing stages. Feature extraction transforms the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it.

Mel Frequency Cepstral Coefficients (MFCC) is the widely used feature extraction technique in speech processing applications. MFCC are also increasingly finding applications in music information such as gener classification, audio similarity measure voice recognition etc.

4. Features

4.1 Long – Term Energy

SNR is the ratio of energies, therefore the long-term energy in each frame from the spectrogram calculate. Long – term energy is defined as the average energy in each frame. Different smoothing windows are applied, using the moving average smoothing method to smooth the abrupt signal transitions. The average energy in each frame n of the signal y can be found by,

$$\varepsilon_y(n) = \frac{1}{|F|}\sum_{f_j \in F} S_y(n, f_j) \qquad (1)$$

where Sy(n,fj) is the spectrum at frame n and frequency bin fj, F is the the set of frequency bins, and |F| is the cardinality of F.

4.2 Signal Variability

The next feature we use to create our regressors is Long-Term Signal Variability (LTSV). Since speech is non-stationary, we can use LTSV to identify speech regions in a signal. It is a way of measuring the degree if non-stationarity in a signal. This is done by measuring the entropy of the normalized short time spectrum at every frequency over consecutive frames. LTSV is computed using the last R frames of the observed signal x with respect to the current frame of interest r.

4.3 Pitch

Pitch can be defined as the quality of sound governed by the rate of vibrations produced by it. Pitch gives the degree of highness or lowness in a signal. Pitch can be expressed as the number of cycles or number of Hertz per second. One cycle means, a complete vibration back and forth. The frequency of the tone is defined by the number of Hertz. For a higher frequency, the pitch will be higher. Pitch detection distinguishes the speech regions of the signal and then this information is exploited to create additional regressors for our models.

$$\text{Pitch} = \frac{Fs}{P}, \qquad (2)$$

Where Fs is the signal frequency and P is the frame period.

4.4 Voicing Probability

The final feature used for detecting speech regions is the voicing probability. Voicing probability assigns a value in every time frame that indicates the probability that a speech exists in that frame. Finally a regression model is created based on the voicing probability.

5. Deep Neural Networking

Neural networks is defined as a computing system made up of many simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. One of the important feature of a neural network is its ability to learn. A neural network usually consists of a large number of processors operating in parallel and arranged in tiers. The first tier receives the input information similar to optic nerves in human visual processing. Each successive tier receives the output from the preceding tier in the same way neurons further from the optic nerve receive signals from those closer to it. The last tier generates the output of the system. Each processing node has its own knowledge, including what it has seen and any rules it was originally programmed with or developed for itself. The tiers are highly interconnected, which means each node in tier n will be connected to many nodes in tier n-1 – its inputs – and in tier n+1, which provides input for those nodes. There may be one or multiple nodes in the output layer. Figure below illustrates a deep neural network.

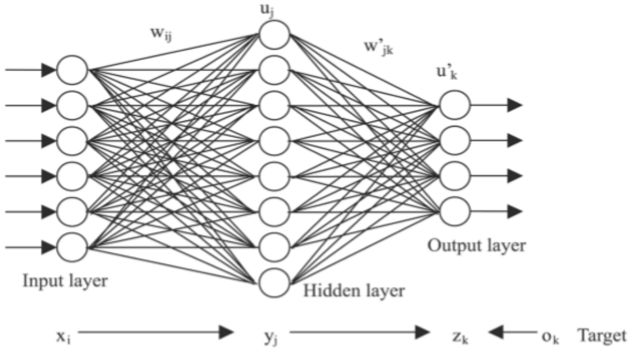


Figure 2: Deep neural networking

6. Word Detection

Speech recognition is the capability of a machine or program to recognize words and phrases in spoken language and convert them to a machine-readable format. Speech recognition software has a limited vocabulary of words and phrases, and it may only identify these if the words are spoken very clearly. Word recognition is commonly used now a days to operate a device, perform commands, or write without using a keyboard, mouse, or press any buttons. Word recognition involves several steps like error histogram, auto correlation, cross correlation etc.

Cross-correlation and autocorrelation are commonly used for calculating the similarity of signals especially for pattern recognition and for signal detection. Autocorrelation mathematically represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals while cross-correlation can be defined as a measure of similarity of two series as a function of the displacement of one relative to the other. Error histogram helps to show how to visualize errors between target values and predicted values after training a feed forward neural network.

7. Experimental Results

MATLAB R2015b is used as the implementation tool. The real-time speech signal is given as the input. For making the system more user friendly a GUI window is provided with different buttons and windows.

The input signal is filtered and MFCC feature extraction technique is applied. The output of MFCC feature extraction is a matrix having feature vectors extracted from all the frames. This output matrix consists rows which represent the corresponding frame numbers and columns which represent the corresponding feature vector coefficients.

Feature extraction is performed on the output of MFCC. Signal features like long-term signal energy, signal variability, pitch and voicing probability is found.

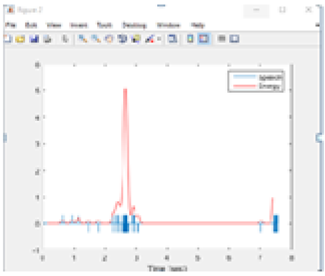


Figure 3: Long-Term Energy

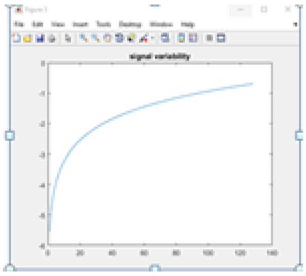


Figure 4: Signal Variability

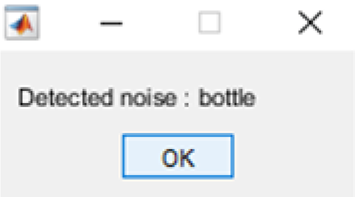
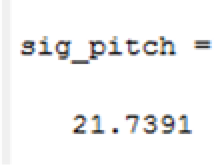

Detected noise : bottle

Figure 5: Detected Noise


```
sig_pitch =

   21.7391
```

Figure 6: Signal Pitch

Deep neural networking is used to match the closest noise type and to detect the words.

8. Conclusion

Using estimation of global SNR in speech signals along with word detection is used. Signal-to-noise ratio gives the information about the level of noise present in a signal. Using MFCC feature extraction improves the performance of the system since the data signals are extracted efficiently from the background noise signals. Regression analysis helps to study the dependence of SNR in various signal features like long-term signal energy, signal variability, pitch and voicing probability. A deep neural networking is used to find the noise type and to detect the input words. Word detection has several advantage in security related systems, speech to text conversion etc. The system is noise independent, therefore works efficiently in an unknown noise conditions. Future scope include methods to improve the signal quality in noisy environment. SNR gives an idea about the amount of noise present in the signal. Designing a system which can improve the signal intelligibility if the SNR is low can improve the performance of word detection applications.

9. Acknowledgment

I would like to express profound gratitude to our Head of the Department, Asst. Prof. Ms. Sangeeta T.R., for her encouragement and for providing all facilities for my work. I express my highest regard and sincere thanks to my guide, Asst. Prof. Ms. Ann Nita Netto, who provided the necessary guidance and serious advice for my work.

References

[1] Siddhant C. Joshi, Dr. A.N.Cheeran, "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition," International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 6,June2014

[2] Namrata Dav," Feature Extraction Methods LPC, PLP, And MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, Volume 1, Issue VI, July 2013

[3] Pavlos Papadopoulos,Andreas Tsiartas and Shrikanth Narayanan , "Long−term SNR estimation of speech Signals in known and unknown channel condition", IEEE / ACM trans. On audio, speech, and language process., vol. 24, no. 12, Dec. 2016

[4] H. G. Hirsch and C. Ehricher , "Noise Estimation techniques for robust speech recognition ", Proc. I EEE Int. Conf. Acoust., Speech, Signal Process. pp. 153 −156. 1995.

[5] J. Morales − Cordovilla, N. Ma, V. Sanchez, J. Carmona, A. Peinado , and J. Barker , "A pitch based noise Estimation technique for robust speech recognition with missing data ", IEEE Int. Conf. Acoust., Speech, Signal Process., 2011, pp.4808−4811.

[6] Ephraim and D. Malah , "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator ", IEEE Trans. Acoust., Speech ,Signal Process., vol. ASSP−32, no. 6, pp. 1109−1121, Dec. 1984.

Author Profile

Sreelakshmi M.S. received B-Tech degree in Electronics and Communication Engineering from M.G University, Kerala at Sree Buddha college of Engineering in 2014. And now she is pursuing her M-Tech degree in Communication Engineering under APJ Abdul Kalam Technological University in Sree Buddha college of Engineering.

Ann Nita Netto is working as Assistant Professor in department of Electronics and Communication, Sree Buddha college of Engineering, Elavumthitta, Pathanamthitta.