TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 2448 - 2453

Speech Signal Intelligibility Improvement by Statistical Signal Processing

[1]D. Sreekanth[*], [2]Kesana Mohana Laxmi, [3]Yalavarthi Priyanka, [4]V. Aswini
[1,2,3,4]Assistant Professor, ECE Department, CMR Technical Campus, Hyderabad
[1]d.sreekanth05@gmail.com, [2]mohana.kesana@gmail.com, [3]yalavarthipriyanka@gmail.com, [4]aswinivunnava@gmail.com

Article Info
Volume 83
Page Number: 2448 - 2453
Publication Issue:
May - June 2020

Article History
Article Received: 11August 2019
Revised: 18November 2019
Accepted: 23January 2020
Publication:10May2020

Abstract:
The speech signal is an effective carrier of information between communicating persons or in public addressing systems. A clean and clear speech signal efficiently conveys the information to the intended destinations. But the background reverberation and noise addition eventually lead to degradation of signal quality and some perceptual aspects such as intelligibility and degree of listener fatigue are affected to some level. It is required to decrease the noise content in the speech signal and improve its intelligibility. This paper articulates the statistical methods and techniques used in suppressing the background noise in a speech signal and improve its intelligibility by removing the unwanted portion of the signal. The statistical signal processing methods are based on observed data analysis and inferences rather than pure mathematical postulations and calculations.

Keywords: Segmentation, Feature Extraction, Binary Mask, Gauusian, Spectogram, Sub-bands, Framing, Training the data.

I. Introduction

Communication is a process of exchanging information. There have been many methods and techniques developed for reliably conveying information from a person to another person or among communicating entities. Voice communication has been the dominant way of exchanging information for centuries. A speech signal is used as a carrier for this information transfer. A noise corrupted speech signal lags in efficiently communicating the information. The understandability of a speech signal masked by noise signals is its perceptual aspect and may be termed as speech intelligibility. Fortunately, numerous techniques and methods are available in the literature to reduce the noise while improving speech quality and intelligibility. A typical noise reduction algorithm may segregate the signal into its unique frequency components each having a specific signal power and uses mathematical or logical models to either suppress or decrease the noise components. It will have to retain signal frequency components and only suppress noise frequency components. The optimal methods of any signal enhancement algorithm will focus on either decreasing the noise level in the signal or increasing the desired signal power. Noise signals usually will be at low frequencies. Since the typical speech frequencies are in the range of 100 Hz to 300 Hz, the removal of noise from speech signal is a cumbersome task. Noise suppression will improve the signal quality. Background reverberation in public addressing system interferes with original speech signal thus limits its intelligibility.

II. Basics of Speech Signal

Speech is defined as the expression of thoughts and feelings by articulating sounds. Speech is the most natural, intuitive and preferred means of communication by human beings. The perceptual variability of speech exists in the form of various languages, dialects, accents, while the vocabulary of speech is growing day by day. More intricate variability at the speech signal level exists in the form of varying amplitude, duration, pitch, timbre

May – June 2020
ISSN: 0193-4120 Page No. 2448 - 2453

and speaker variability. The intricate variabilities in speech makes it more complicated to analyse but provides additional information using the tone and amplitude variation.

III. Characteristics of Speech Signal

A speech signal has both temporal characteristics and spectral characteristics. These characteristics are essential in the analysis of speech components. In general, both time domain and frequency domain descriptions are used to study a signal. Some characteristics are easy to be drawn from time domain view of the signal while some characteristics are to be mandatorily drawn only form spectral view of the signal. A deterministic signal characteristics such as amplitude, frequency, phase, energy , power , and bandwidth are computed after observing the signal in time domain and frequency domain. Unlike a deterministic signal, a random signal characteristics are primarily derived from spectral representation of the signal while time domain analysis gives limited scope to understand and estimate its characteristics.The characteristics which are observed in the time domain representation of the speech signal are called temporal characteristics. Theses characteristics include first order stationarity, second order stationarity, Nth order stationarity, strict sens stationarity, correlation, zero crossing rate, maximum amplitude, minimum energy etc. The characteristics observed in the signal when it is transformed into frequency domain are called spectral characteristics. They include fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. These features can be used to identify the notes, pitch, rhythm, and melody. The most successful spectral features used in speech are (i) Mel frequency cepstral coefficients (MFCC) and (ii) Perceptive Linear Prediction (PLP) features. It is well known that the basilar membrane in the inner ear actually analyzes the frequency content of the speech we hear. In fact, the analysis of basilar membrane can be modeled by a bank of constant Q, band pass filters. There also exist the critical bands, which give rise to the phenomenon of masking - where one strong tone or burst can mask another weaker tone within the critical band. Actually, both MFCC and PLP capture these characteristics of our auditory system in some way; so, even though it looks strange, the same features give reasonably good performance for speech recognition, speaker recognition, language identification and even accent identification ! However, these spectral features are not very robust to noise. On the other hand, some of the time domain (temporal) features such as plosion index and maximum correlation coefficient are relatively more robust to noise.

IV. Statistical Speech Processing Model

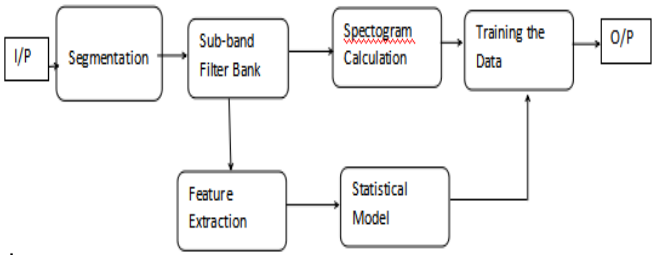

Figure 1: Statistical Speech Processing Model

V. The Procedural Steps of Statistical Signal Processing

1. A speech signal is a non-stationary signal whose statistical values such as Mean, Correlation etc, vary with time shift. A random signal process is said to be stationary if all its statistical parameters do not change with a shift in time origin. A stationary signal can be obtained by dividing the signal into short-time slots. A speech signal time slice of 10ms - 30ms is approximately a stationary signal. The division of the signal into shorter time slots is done to view the speech signal as a stationary function of time. These short-duration signals will also have stable frequency characteristics. The division of speech signal into frames is done using windows. In order to have continuity in signal spectral characteristics, the frames are overlapped to some

May – June 2020
ISSN: 0193-4120 Page No. 2448 - 2453

percentage. This process of speech signal division into short time slots is called speech Segmentation.

2. The signal then decomposed into sub-bands by using filter banks. The input signal is applied to a filter bank with various cut-off frequencies. Each filter has a unique cut-off frequency. A filter bank is a collection of band-pass filters which separates the input signal into multiple frequency components, each carrying a unique sub-band of the original signal. This process of decomposing the speech signal into sub-band frequencies is called Analysis process. The reverse process of combining back all these sub-bands to obtain original signal is called Synthesis. This process enables the ability to treat several frequency components of the signals as unique frequency components.

3. Every speech sub-band signal has some unique features that make it distinguished from other sub-bands of the signal. Those features are Formant Frequencies, Cepstral coefficients, Mel-frequency Cepstral Coefficients, Delta Mel-frequency Cepstral Coefficients, Linear Prediction Coefficients, Lienar Prediction Cepstral Coefficients, Amplitdue Modulated Spectrograms etc. These features are useful in many speech signal enhancement and processing applications such as compression, enhancement, speech coding, speech recognition, speech analysis etc. A feature value of a signal is used to explicitly distinguishing a signal from other speech signals. The features are derived both in time domain and frequency domain versions of the signal.

4. Power spectrum of the short-time signal frames is obtained by squaring the Fourier transform of every time slot. It is a graphical representation of the power in each frequency of the signal. In general, Magnitude of Fourier transform of a signal is called Signal Magnitude Spectrum and Square of Magnitude of Fourier transform is called Signal Power Spectrum. Spectogram is a two-dimensional representation of the speech signal. It is a Time versus Frequency Plot. It is a visual representation of frequencies of a sound signal as it varies with time. The range of fundamental frequency components of a signal is known as its considerable bandwidth. This bandwidth is useful and essential while selecting a filter bank or determining range of desired frequency components for speech signal applications.

5. Statistical models are developed for the purpose of training the data. A statistical model makes inferences on the sub regions of the speech signal. These inferences are about the properties of the signal segments. The un-deterministic nature of the speech signal limits prediction of signal parameters in advance. This drawback is overcome by statistical parameters calculations. These parameters are computed from the past observed values and probabilistically infering the future values of the speech signal. The Gaussian pdf is a state-dependent function in that there is assigned a different Gaussian pdf for each acoustic sound class. The states are like Quasi periodic, noise-like, and impulse like sounds or on a very fine level such as individual phonemes. Statistical models are believed to be more accurate and reasonably effective when random nature of data limits the mathematical analysis and parameter computing.

6. In the techniques used to enhance a speech segment in the presence of noise, a binary mask has been a primary choice. It may be called as a priori mask. It was explored in CASA analysis. Computational Auditory Scene Analysis retains the Time-Freequency regions of the destined signal if it is exceeding the threshold of noise. This noise threshold is defined by a Noise Masker. The signal segments weaker than noise masker will be removed. To improve the intelligibility gians, even at low SNR levels, the optimal method is to multiply the noise masked signal with binary mask. To estimate the binary mask in advance, an accurate knowledge of the desired signal spectral SNR is required. In practice the binary mask and true SNR are assumed prior.

May – June 2020
ISSN: 0193-4120 Page No. 2448 - 2453

7. Training the data using some statistical or probabilistic models to reduce noise levels or improve perceptual aspects. For this purpose many statistical model have been used such as Gaussian Mixture Models, Bayesian Classifiers, Statistical Inferences, Binary Masks etc. Usually this process retains the target dominated frames and suppressing the noise dominated frames. The efficiency of any model to train the data is utmost important in the processing method.

8. After the data training and enhancement, the signal speech parameters are improved and noise levels are decreased in each speech segment. These speech segments are combined back to obtain a continuous posterior enhanced signal. The intelligibility improvement is said to be attained when its perceptual aspects are raised to higher levels. A comparative analysis is done after the signal processing to account the level of improvement and efficiency in speech intelligibility. The binary mask or Gaussian mixture models are most used probabilistic models in the intelligibility improvement.

VI. Results and Discussions

There is an improvement in the gains of intelligibility by using the statistical signal processing methods over other methods. The intelligibility identified by human listeners is significantly better when compared with with un-processed speech signal corrupted by noise and background reverberation.

The classification of speech signal T-F units into target dominated units and masker dominated units makes the enhancement and improvement process optimal. There are wide demonstrations as how binary mask estimation is effective in substantial improvement in intelligibility gain.The accurate classification of T-F units into target- and masker-dominated T-F units was accomplished with the use of neurophysiologically-motivated features (AMS) and carefully designed Bayesian classifiers (GMMs). Unlike the mel-frequency cepstrum coefficients features commonly used in ASR, the AMS features capture information about amplitude and frequency modulations, known to be critically important for speech recognition. GMMs are known to accurately represent a large class of feature distributions, and as classifiers, GMMs have been used successfully in several applications and, in particular speaker recognition. Other classifiers (e.g., neural networks, and support vector machines) could alternatively be used.

	Babble noise		Factory noise		Speech shaped noise	
	-5dB	0 dB	-5dB	0 dB	-5dB	0 dB
Before processing	21%	78%	43%	82%	43%	84%
After processing	93%	96%	91%	94%	90%	91%

Table 1. Percentage of intelligibilities (Percentage of persons who correctly identified) for various noisy environments and at different SNR levels before and after processing

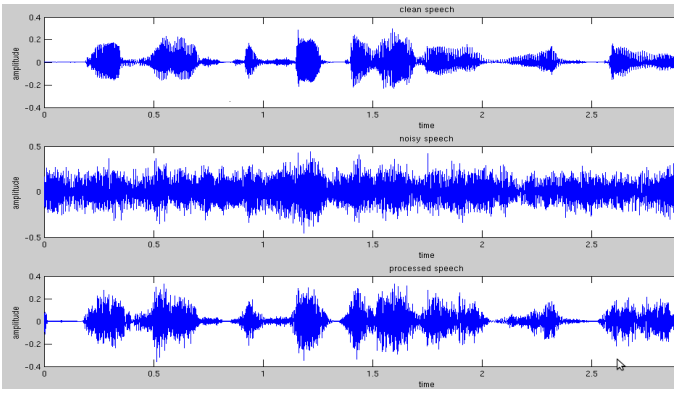

Figure 2. Clean, noisy and processed speech signal

May – June 2020
ISSN: 0193-4120 Page No. 2448 - 2453

VII. Conclusion

The speech signal is a very useful physical carrier of information between communicating parties. There have been limitations caused by noise and background reverberation in qualitatively conveying this information to the intended destinations. Numerous algorithms and techniques are developed to enhance and improve the signal parameters, of such methods one is Statistical signal processing which effectively improves the speech intelligibility. As we have understood, the speech intelligibility is the utterance of words or understandable level of content in the speech in the presence of noise or background reverberation. This method has a step by step procedure to enhance the speech signal and improve intelligibility. The speech signal is first divided into short duration time signals to observe stationarity in its statistical characteristics. This phenomena makes statistical models consistent with respect to time. Later these frames of the signal are sub-band filtered to separate them on frequency scale. The features of the sub-bands are extracted which are later used to build a statistical model. The statistical model is used to train each unit of data. Particularly the binary masking is used to suppress noise dominated signal units and retain mask dominated signal units. The statistical method is a reliable and optimal method of intelligibility improvement.

References:

1. Huan-Yu Dong, Chang-Myung Lee (2018), "Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering", Dong and Lee EURASIP Journal on Audio, Speech, and Music Processing.
2. Sunita Dixit, Dr. MD Yusuf Mulge (2014), "Review on Speech Enhancement Techniques", International Journal of Computer Science and Mobile Computing.
3. D. Sreekanth, D. Sunitha (2013) "Probabilistic Approach For Speech Intelligibility Improvement And Noise Reduction", International Journal of Advanced Trends in Computer Science and Engineering.
4. Tayseer M F Taha and Amir Hussain. A Survey on Techniques for Enhancing Speech. *International Journal of Computer Applications* 179(17):1-14, February 2018.
5. S.China Venkateswarlu, A.Karthik, K. Naveen Kumar, (2019)," Performance on Speech Enhancement Objective Quality Measures Using Hybrid Wavelet Thresholding", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6, August 2019.

Author Biographies

1. D. Sreekanth is currently working as an Assistant Professor in CMR Technical Campus, An Autonomous Institute, Kandlakoya, Mechal, Telangana. He completed M.Tech in Wireless and Mobile Communications from JNTU Hyderabad. He has 10 years of teaching experience. His area of research is Signal and Image Processing. He has published 7 International Journal Papers on signal processing and communication systems.

2. Kesana Mohana Laxmi is currently working as Associate Professor in CMR Technical Campus, An Autonomous Institute, Kandlakoya, Medchal, Telangana. She completed M.Tech in Systems and

Signal Processing, from JNTU Hyderabad. She is pursuing Ph.D in Image Processing from Acharya Nagarjuna University, Andhara Pradesh. Her area of research is Telugu Word Image recognition and retrieval. She has published 10 International Journal Papers on Signal Processing and Digital Image Processing.

3. Yalavarthi Priyanka is currently working as an Assistant Professor in CMR Technical Campus, An Autonomous Institute, Kandlyakoya, Medchal, Telangana. She complete M.Tech in Embedded Systems from CMR Technical Campus Hyderabad. She has 4 years of teaching experience. Her research area is Embedd systems and Signal Processing.

4. V. Aswini is currently working as an Assistant Professor in CMR Technical Campus, An Autonomous Institute, Kandlyakoya, Medchal, Telangana. She completed M.Tech in Digital Electronics and Communications form JNTU Hyderabad in. She has 8 years of teaching experience. Her research area is Signal processing and communication systems.