Open lab –Speech/Audio signal Processing using MATLAB

Lab Sheet 4

Time Domain Analysis of Speech Signals

Aim

- To understand need for feature extraction and short time processing of speech.
- To compute short time energy and study its significance.
- To compute short time zero crossing rate and study its significance.
- To compute short time autocorrelation and study its significance.
- To estimate pitch of speech using short time autocorrelation.

Feature Extraction

In speech based systems, feature extraction is needed because the raw speech signal contains information besides the linguistic message and has a high dimensionality. The characteristics of the raw speech signal would be unfeasible for the classification of sounds and result in a high word error rate. Therefore, the feature extraction algorithm derives a characteristic feature vector with a lower dimensionality, which is used for the classification of sounds. Also feature vector should emphasize the important information regarding the specific task and suppress all other information. As the goal of automatic speech recognition is to transcribe the linguistic message the information about this message needs to be emphasized. For eg. Speaker dependent characteristics, the characteristics of the environment and the recording equipment should be suppressed because these characteristics do not contain any information about the linguistic message. Furthermore, the feature extraction should reduce the dimensionality of the data to reduce the computational time.

Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

Need for Short Time Processing of Speech

Speech is produced from a time varying vocal tract system with time varying excitation. As a result the speech signal is non-stationary in nature. Most of the signal processing tools studied in signals and systems and signal processing assume time invariant system and time invariant excitation, i.e. stationary signal. Hence these tools are not directly applicable for speech processing. A solution proposed for processing speech was to make use of existing signal processing tools in a modified fashion. To be more specific, the tools can still assume the signal under processing to be stationary. Speech signal may be stationary when it is viewed in blocks of 10-30 msec. Hence to process speech by different signal processing tools, it is viewed in terms of 10-30 msec. Such a processing is termed as Short Time Processing (STP).

Short Time processing of speech can be performed either in time domain or in frequency domain. The particular domain of processing depends on the information from the speech that we are interested in. For instance, parameters like short time energy, short time zero crossing rate and short time autocorrelation can be computed from the time domain processing of speech. Alternatively, short time Fourier transform can be computed from the

frequency domain processing of speech. Each of these parameters give different information about speech that can be used for automatic processing.

Short time Energy Parameter

The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific, energy associated with short time region of speech. By the nature of production, the speech signal consist of voiced, unvoiced and silence regions. A voiced speech segment is also known as pitch of voiced speech. It has high energy content and is periodic in nature. The unvoiced part of the speech looks like a random noise with no periodicity. Some parts of the speech that are neither voiced nor unvoiced are called transition segments.

We can write the relation of short term energy as follows

$$e(n) = \sum_{m=-\infty}^{\infty} (s(m).w(n-m))^2$$

where "w(m)" represent the windowing function of finite duration. There are several windowing functions present in the signal processing literature. The mostly used ones include rectangular, hanning and hamming. For all time domain parameters estimation we use the rectangular window for its simplicity. "n" is the shift / rate in number of samples at which we are interested in knowing the short term energy. The shift can be as small as one sample or as large as frame size. The short term energy computed for every sample shift may not be required since the energy variation in case of speech is relatively slow. For this reason the shift is kept much larger than one sample. Usually it is about half the frame size.

Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy. Thus short time energy can be used for voiced, unvoiced and silence classification of speech.

Experiment1

I. From the segmented speech signal in the previous lab sheet, calculate the short time energy. The segments should be 30 msec duration with 50% overlap.
II. Record the alphabet ‘a’ and alphabet‘s’. Compare the short time energy in each case. Comment on your inference

Short Time Zero Crossing Rate (ZCR)

Zero Crossing Rate gives information about the number of zero-crossings present in a given signal. Intuitively, if the number of zero crossings are more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information. On the similar lines, if the number of zero crossing are less, hence the signal is changing slowly and accordingly the signal may contain low frequency information. Thus ZCR gives an indirect information about the frequency content of the signal. The ZCR in case of stationary signal is defined as,

$$z = \sum_{n=-\infty}^{\infty} |sgn(s(n)) - sgn(s(n-1))|$$

$$where\ sgn(s(n)) = 1\ if\ s(n) \geq 0$$

$$= -1\ if\ s(n) < 0$$

And for non stationay signals, it is given by

$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m)$$

The equation is divided by a factor "1/2N" . This comes in the denominator to take care of the fact that there will be two zero crossings per cycle of one signal

Experiment 2

III. Plot one cycle of the sine wave and verify the number of zero crossings using the above equation.

IV. From the segmented speech signal in the previous lab sheet, calculate the short time ZCR. The segments should be 30 msec duration with 50% overlap.

V. Record the alphabet 'a' and alphabet's'. Compare the short time energy in each case. Comment on your inference.

Short Time Autocorrelation:

Crosscorrelation tool from signal processing can be used for finding the similarity among the two sequences and refers to the case of having two different sequences for correlation. Autocorrelation refers to the case of having only one sequence for correlation. In autocorrelation, the interest is in observing how similar the signal characteristics with respect to time. This is achived by providing different time lag for the sequence and computing with the given sequence as reference.

The autocorrelation is a very useful tool in case of speech processing. However due to the non-stationary nature of speech, a short term version of the autocorrelation is needed. The autocorrelation of a stationary sequence rxx(k) is given by

$$r_{xx}(k) = \sum_{m=-\infty}^{\infty} x(m).x(m+k)$$

The corresponding short term autocorrelation of a non-stationary sequence s(n) is defined as

$$r_{ss} = \sum_{m=-\infty}^{\infty} s_w(m).s_w(k+m)$$

$$r_{ss}(n,k) = \sum_{m=-\infty}^{\infty} (s(m)w(n-m).s(k+m).w(n-k+m)))$$

where $s_w(n)=s(m).w(n-m)$ is the windowed version of s(n). Thus for a given windowed segment of speech,the short term autocorrelation is a sequence. The nature of short term autocorrelation sequence is primarily different for voiced and unvoiced segments of speech. Hence information from the autocorrelation sequence can be used for discriminating voiced and unvoiced segments. The typical frame size for computing short term autocorrelation should include at

least two cycles of speech signal in the voiced speech case. To ensure this the size is used in the range 30-50 msec. The nature of autocorrelation sequence in case of autocorrelation of voiced speech can be explained for finding the periodicity of voiced speech. Accordingly, the autocorrelation of voiced speech should give strong peak at the periodic value and no such peak in case of unvoiced speech. Therefore, the autocorrelation of speech has become a standard approach for enhancing pitch

Experiment 3

VI. From one segment of speech signal plot the Short time ACF

Lab Assignments

- Refer literatures and write a matlab routine using the above features STE and STZCR for finding the distinguishing speech from silence
- Find the Pitch contour from STACF