UDC 004.934.2

SPECTROGRAM OF SPEECH SIGNAL IN MATLAB

Asinenka A. M.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Alefirenko V.M. – Cand. of Sci., associate professor of the department of ETT

Klokova A.G. – Cand. of Sci., associate professor, head of the department of foreign languages.

Annotation. This article explains spectrogram of the speech signal (analysis and processing) with MATLAB to get its frequency-domain representation. In real life, we come across many signals that are variations of the form f(t), where «t» is independent variable «time» in most cases. Temperature, pressure, pulse rate, etc can be plotted along the time axis to see variations across time.

Keywords: spectrogram, MATLAB, speech signal, waveform, hanning window.

Introduction. In signal processing, signals can be classified broadly into deterministic signals and stochastic signals. Deterministic signals can be expressed in the form of a mathematical equation and there is no randomness associated with them. The value of the signal at any point of time can be obtained by evaluating the mathematical equation.

Many of the information-bearing signals may not be predictable in advance. There is a certain amount of randomness in the signal with respect to time. Such signals cannot be expressed in the form of simple mathematical equations. For example, in the noise signal inside a running automobile, we may hear many sounds, including the engine sound, sound of horns from other vehicles and passengers talking, in a combined form with no predictability. Such signals are examples of stochastic signals [1].

Main part. In the speech signal produced when you utter steady sounds like «a», «i» or «u», the waveform is a near-periodic repetition of some well-defined patterns. When you produce sounds like «s» and «sh», the waveform is noise-like. The periodicity in the speech signal is due to the vibration of vocal folds at a particular frequency, known as pitch or fundamental frequency of the speaker. Steady sounds (a, i or u) are examples of vowels and noise-like sounds (s and sh) are examples of consonants. Human speech signal is a chain of vowels and consonants grouped in different forms.

Most of the signals in real life are available continuously and may assume any amplitude value. These signals are called analogue signals and they are not in a form suitable for storing or processing using a digital computer. In digital signal processing, we process the signal as an array of numbers. We do sampling along the time axis to discretise the independent variable «t». In other words, we look at the signal at a number of time instances separated by a fixed interval «T». Signal values observed at these time instances are further discretised in the amplitude domain to make these suitable for storage in the form of binary digits. This process is called quantisation.

Record vowel sound «aa» using the computer’s microphone and save it as a wav file. Select sampling frequency as 10kHz.

The waveform of the signal, which is a plot of the amplitude of the speech signal for each sample instant, looks like Figure 1. The horizontal axis is time units in samples and the vertical axis is amplitude of the corresponding samples. If you record sound «as» in which consonant sound «s» follows the vowel sound «a» and plot the signal, the waveform may look like Figure 2.

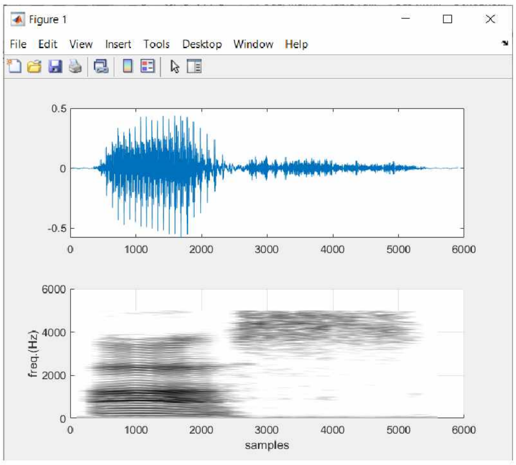

Figure 1 – Waveform of vowel sound «aa»

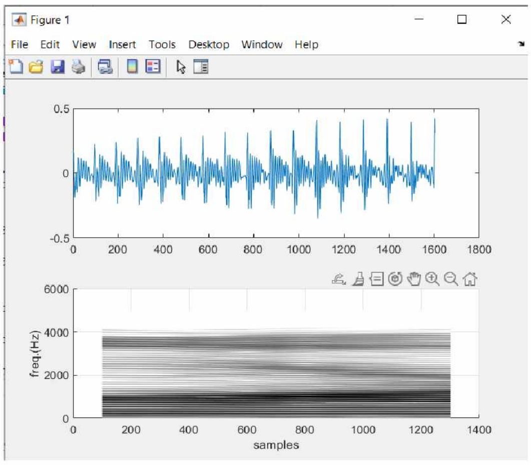

Figure 2 – Waveform of vowel-consonant sound «as»

Waveform is a representation of the speech signal. It is a visualisation of the signal in time domain. This representation is almost silent on the frequency contents and the frequency distribution of energy in the speech signal. To get a frequency-domain representation, you need to take Fourier transform of the speech signal. Since speech signal has time-varying properties, the transformation from time-domain to frequency-domain also needs to be done in a time-dependent manner. In other words, you need to take small frames at different points along the time axis, take Fourier transform of the short-duration frames, and then proceed along the time axis towards the end of the utterance. The process is called short-time Fourier transform (STFT). Steps involved in STFT computation are:1 Select a short-duration frame of the speech signal by windowing; 2 Compute Fourier transform of the selected duration; 3 Shift the window along the time axis to select the neighbouring frame; 4 Repeat step 2 until you reach the end of the speech signal.

To select a short-duration frame of speech, normally a window function with gradually rising and falling property is used. Commonly used window functions in speech processing are Hamming and Hanning windows.

A window function has non-zero values over a selected set of points and zero values outside this interval. When you multiply a signal with a window function, you get a set of «N» selected samples from the location where you place the window and zero-valued samples at all other points.

Once these parameters are finalised, framing operation is performed using the MATLAB user-defined function (needs to be copied to the same folder where the main program is stored):

```
frames = speech2frames (speech, Nw, Ns,‘cols’, hanning, false );
```

Generally, frame-duration parameter N_{wt} and frame-shift parameter N_{st} are selected such that consecutive frames have sufficient overlap. The condition $N_{st}<N_{wt}$ ensures an overlapping window placement. In speech processing applications, overlapping is generally kept above percent by proper selection of N_{st} and N_{wt}. The framing operation returns a number of short-duration frames selected using the window function with the specified frame length and frameshift parameters. Each frame is stored as a column vector in the returned array. Once the framing is performed, DFT operation is used to transform each frame to a frequency domain using the command:

```
MAG = abs ( fft(frames,nfft,1) );
```

Parameter «nfft» specifies the number of points in the DFT operation. It is kept as a power of 2 and must be greater than the frame length in samples. Assuming the wav file has sampling frequency fs of 10kHz, we have used 1024 points as «nfft» for a frame length of 400 samples (40ms). If the sampling frequency of the wav file is not 10kHz, the file needs to be resampled to 10kHz for proper working of the program. Frame shift parameter is set as 100 samples (10ms). MAG variable has the absolute value of Fourier transform of frames stored column wise. The magnitude of Fourier transform is also called spectrum of that frame of the signal. As the speech signal has time-varying properties, the spectrum also goes on varying with time as we move along the samples in the wav file [2].

Magnitude spectrum computed for individual frames can be represented in many forms. We have been following three parameters: Frame number (indicator of the time axis), DFT bin number (indicator of the frequency axis) and magnitude of DFT computed (indicator of the spectral energy).

These three parameters can be represented conveniently in a 2D format using spectrogram. Spectrogram can be considered as an image representing time and frequency parameters (along X and Y axes) and magnitude values as the intensity of pixels in the X-Y plane. Stronger magnitudes get represented by dark spots and silences (low- or zero-amplitude signals) get represented by white spots in the image [3].

Conclusion. Naturally, in such a challenge like spoken language recognition where performance is the most important measure, understand the phenomenon and analyse the different implications becomes secondary issue, or even. However, it is a fact that the best performance will be achieved only with deep knowledge and understanding of the underlying process, and that is an outcome of this work [4].

The presented spectrograms show 2 similar speech signals. We can observe a large difference in the spectrograms of these signals, but when listening to them, the difference is very difficult to hear. Therefore, the use of spectrograms plays an important role in protecting speech information.

References

1. Deterministic Signal – an overview. [Electronic resource]. – https://www.sciencedirect.com/topics/computer-science/deterministic-signal. – Date of access: 19.12.2023.

2. Understanding Spectrogram of Speech Signal Using MATLAB. [Electronic resource]. – Mode of access: https://www.electronicsforu.com/electronics-projects/software-projects-ideas/spectrogram-speech-signal-matlab. – Date of access: 19.12.2023.

3. Spectrogram using short-time fourier transform. [Electronic resource]. – Mode of access: https://www.mathworks.com/help/signal/ref/spectrogram.html. [Electronic resource]. – Date of access: 20.12.2023.

4. ResearchGate. [Electronic resource]. https://www.researchgate.net/publication/283014640_Language_Identification_Using_Spectrogram_Texture/link/. – Date of access: 21.12.2023.