IMPLEMENTATION OF FAST FOURIER TRANSFORM(FFT) FOR INFANT CRYING DETECTION

Latifah Listyalina [1, a)] Evrita Lusiana Utari [2, b)] Mario Warran Wizando [3, c)]

[1]*Rubber and Plastic Processing Technology, Politeknik ATK Yogyakarta*
[2,3]*Electrical Engineering, Yogyakarta Respati University*

[a)]Corresponding author:
latifah.listyalina@atk.ac.id
[b)]evrita_lusiana@yahoo.com
[c)]mariowarranw@gmail.com

Abstract. Babies cry based on the discomfort felt by the baby which is a reflex such as when a hungry baby will suck his hand and then he will start crying, hunger can be interpreted from the baby's crying. At each of the baby's cries, when each cryingpattern was responded to with the solution applied to the previous baby, each baby would stop crying. For this reason, to carry out this solution, a research was carried out to identify the pattern recognition of the sound of a baby's cry using the fast fourier transform (FFT) method with several different frequency ranges. The voice recording process is stored in digital form in the form of frequency-based sound spectrum waves, where signals that were previously in the time domain will be changed in the frequency domain. The sounds that will be distinguished in this study include the sounds of crying babies, adults, and colliding objects. This can be obtained through several stages, namely sound sample recording, sampling, signal cutting, frame blocking, final normalization, hamming window, and finally the FFT calculation process. From these series of stages, the results of the frequency range of baby crying are 101-1863 Hz, for adults the frequency range is 101-1376 Hz and for the sound of colliding objects 101-2233 Hz.

INTRODUCTION

The sound of a baby crying is one way of communicating to convey a situation that he experiences when he cries.Babies can make different crying sounds depending on their condition. Baby voice recognition is needed as a quick solution for babies when they cry. In addition, the increase in science and technology is currently developing rapidly.Advances in technology aim to facilitate human activities in all fields, intensive research in the field of signal processing causes technology to develop very rapidly. One of them is speech recognition. [1][2]

Sound is something unique and has a certain frequency range and sound intensity that can and cannot be heard byhumans. The unit for measuring sound intensity is the decibel (dB) taken from the name of its inventor, namely Alexander Graham Bell, who is known as the inventor of the telephone, while the unit for sound frequency is Hertz, taken from the name of a physicist, Heinrich Rudolf Hertz to appreciate services for his contribution in the electromagnetic field. [3][4]. The process of speech recognition by humans begins to form when toddlers are able to hear and are able to make sounds. In this study, the feature extraction of baby crying sound signals was carried out using the Fast Fourier Transform (FFT) method, which is a process of transforming sound signals in the time domaininto frequency signals.

An important process in Digital Signal Processing (DSP) is analyzing an input and output signal to determine thecharacteristics of a particular physical system. The process of analysis and synthesis in the time domain requires a long analysis involving the derivative of the function, which can lead to inaccuracies in the analysis results. Signal analysis and synthesis will be easier to do in the frequency domain, because the quantity that most determines a signalis frequency. In the time domain, signal analysis cannot be performed. Analysis can be done if the signal is in the formof a spectrum. So it is necessary to transform the signal from the time domain into a frequency domain signal. The function used to view the vibration spectrum of the time domain signal is the Fast Fourier Transform (FFT).

With the above background, in this final project research, a study will be conducted with the title "Implementationof Fast Fourier Transform (FFT) for Baby Cry Detection" because the author sees that babies need special handling in their care. The author wants to make pattern recognition using DFT to generate a frequency spectrum which will then be analyzed by finding the maximum values to determine the frequency range of each sample that has been recorded. If this research can produce data for detecting the sound of a baby crying, this research can be continued tobecome a special nursecall alarm for detecting baby crying, so that it can become a tool that makes it easier for nursesto monitor babies in special baby care rooms.

METHODOLOGY / RESEARCH METHODOLOGY

The research procedures carried out in this study consisted of sampling, sample classification, signal processing, and pattern recognition. The overall research procedure carried out can be seen in the flowchart below.

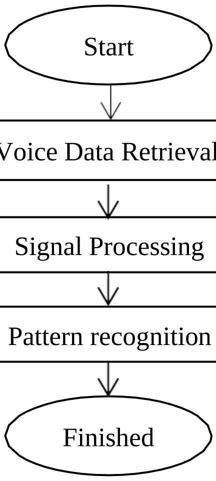

Figure 1. Research Flowchart

Voice Data Retrieval

The sample data is in the form of human voices and the sound of colliding objects. The sound will be used as datain this study. At this stage, the human voice is in the form of a baby crying and the voice of an adult, while the soundof colliding objects is the sound of metal, plastic, wood and glass objects.

The baby sound used is a baby crying from a maternity clinic with a total of ten babies. The adult voice samples used were six adults who were asked to say "My name (name)" and the other four adult voices were noise recordingsand adult chatter. Examples of object impacts are wood knocks, plastic knocks, metal knocks, and glass knocks. Withtwo samples of wood knock, three samples of plastic beat, two samples of metal knock and two samples of glass knock. Voice recording is done using a recording application and then trimming is done so that the audio recording is2 seconds. Then the sound sample will be converted and saved in .wav format. Saved in .wav format because .wav usually uses PCM (Pulse Code Modulation) coding. With this way,

Digital Signal Processing

Signal processing stages are the processes carried out for the process of generating spectrum and FFT. The purposeof this signal processing stage is to equalize the input sound signal so that it is easier to process for speech pattern recognition. The signal processing process has several stages that must be passed, namely normalization, signal cutting, frame blocking, windowing, FFT frequency spectrum, and finding the maximum value of the FFT frequency.The first process is sampling, which is the process of sampling sound waves into a continuous signal. In the process of sampling there is what is called the sampling rate. The sampling rate indicates how many analog waves are sampledin 1 second. The sampling rate is expressed in Hertz (Hz). The sampling rate used is 44100 Hz. The reason for usinga sampling rate of 44100 Hz is because the limit of the ability of the human ear to perceive sound frequencies is from20 Hz to 20000 Hz, so the most suitable sample rate to use is 44100 Hz. The time used is 2 seconds.

This normalization process is carried out after the output of the sampling process is in the form of wav. In the normalization process, the voice signal must have a maximum value. After searching for the maximum value, the nextstep is to divide the data by the maximum absolute value of the sound recorded.

$$X_{norm} = \frac{x_{in}}{\max(abs(x_{in}))} \qquad (1)$$

where $Xnorm$ is the result of normalized signal data, and Xin is the input data from sampling. The purpose of thisnormalization is to equalize the amplitude of the recorded sound to the maximum, so that the strong or weak effects of the recorded voice do not affect the speech recognition process.

The process of cutting the signal is done to remove the initial signal that is not used and is located on the left sideor the beginning of the signal, namely the silence area and the transition area. The purpose of cutting the silence section is to remove parts that are not part of the sound signal and the research object, and the purpose of cutting the transition area is to get a signal which is the sound signal of the research object. This cutting process is carried out after the normalization process. In the slicing process, the truncated portion is the initial portion of the signal. Signal truncation is done by determining the cutoff limit value, which is 0.3. Then look for the part of the signal that is greaterthan 0.3 and less than -0.3. The searched signal is initialized as b0. The process of cutting the transition area is done by removing ¼ part of the signal that is in the beginning (transition part) after cutting the silence part. Signals that donot include b0 will be omitted, this omitted signal is called the silence signal. Then cutting the transition signal is doneby multiplying the number of signal data by 0.25.

*Frame blocking*is the next process after going through the process of cutting the signal. The value of frame blocking aims to reduce the amount of signal data to be processed. Frame blocking functions to select data from the entire data recorded from the results of signal cutting. The first process carried out in frame blocking is to determine the midpoint value of the sampling data. From the midpoint of the data obtained, it is determined the amount of data to be retrieved for the next process. The value of frame blocking used in this study is 256. The input for the final normalization is the signal from frame blocking. In the frame blocking process, the resulting signal is not optimal, soa final normalization is needed to equalize the amplitude to the maximum. In the final normalization process, The input data resulting from frame blocking is divided by the maximum absolute value of the data resulting from the frame blocking. The result of the division is the output for the final normalization process

*windowing*has a function to eliminate the discontinuity effect caused by the previous process, namely frame blocking when the signal is transformed to the frequency domain. In this study, a hamming window is used where theuse of this window makes the results smoother in eliminating the effects of discontinuities. In this process, the result of frame blocking will be multiplied by the hamming window. This is done to get maximum results in the FFT process,so that samples that have been divided into several frames need to be made into continuous sound.

The process of changing from the time dimension to the frequency dimension begins with finding the computed value of the FFT followed by finding the absolute value of the computed FFT. The process is continued by finding the absolute value of the log value of the results of the FFT mathematical calculation. The computational results are cut by half of the predetermined signal size, then the part to be processed is selected from the results of the cut. The final process in feature extraction is changing the dimensions of the signal. The search for absolute value in each calculation is intended so that the value obtained is a real number so that the calculation process can be continued.

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-i2\pi ft}\, dt$$

$$= \int_{-\infty}^{\infty} x(t) \cos(2\pi ft)\, dt - i \int_{-\infty}^{\infty} x(t) \sin(2\pi ft)\, dt \qquad (2)$$

with:
x(t) = function or signal in time domain;
$e{-}i2\pi ft =$ kernel function;
X(f) = function in the frequency domain
and;f = frequency.
The function of the equation above is used to transform signals from the time domain into the frequency domain.

Pattern recognition

The search process for the maximum value aims to find the maximum values after getting the FFT signal. This process aims to analyze the results of the previous FFT process. These maximum values will be used to determine thepattern of the baby's crying sound. To find the maximum values of the FFT signal, first find the signal data length. Then look for the absolute values of the FFT, and look for the highest values in the FFT signal as an index of the maximum value.

RESULTS AND DISCUSSION

In this study, the human voice is the sound of a baby crying and the voice of an adult, while the sound of colliding objects is the sound of metal, plastic and wood objects colliding. Baby sounds are used with a duration of 2 seconds, but if the recording exceeds 2 seconds, it must be cut using the same application so that the recording becomes 2 seconds. These recordings are still in .mp3 format and must be converted to .wav files so that details are not lost whenthe analog audio is digitized and saved. There were 30 samples in this study, with 10 each for each sample of infants,adults and objects. The sound sampling program uses a sampling frequency of 44100 Hz which is initialized as fs anda duration of 2 seconds which is initialized as t. The following is the result of a plot of one of the recording signals.

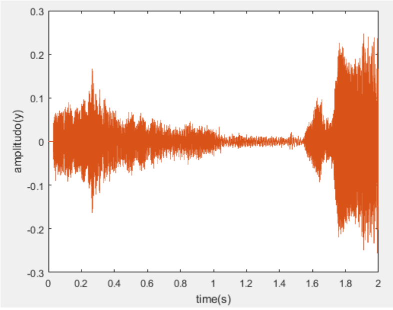

Figure 2. Plot of Sampling Signals (Baby Sounds 1)

In the sampling signal plot above, the x-axis is the recording time of the sound sample and the y-axis is the amplitude of the sound sample, the time is 2 seconds with a maximum amplitude of 0.2473 at 1.911 seconds and a minimum amplitude of -0.2552 at 1.994 seconds.

This normalization process is carried out after the output of the sampling process. In the normalization process, the voice signal must have a maximum value. The reason for using frame blocking 256 is because the higher the frameblocking, the more accurate speech signal recognition will be. Then x1 is the result of normalization, x is the recordedsound signal and max(x) is the maximum value of the recorded sound signal. Normalization is done by dividing the input data (recorded voice signal data) by the maximum value of the data. The following is the result of a plot of oneof the normalized signals

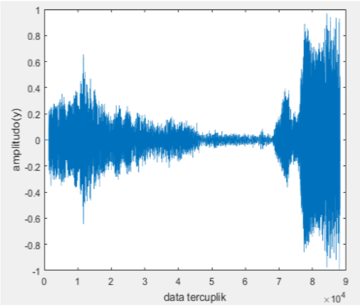

Figure 3. Plotting of Initial Normalized Results (Baby Sounds 1)

In the normalization result plot above, the signal normalization process from the sampling signal or recording signal is carried out in order to equalize the amplitude of the sampling signal data or recording signal so as to obtain the same scale. The x axis is the sampled data and the y axis is the amplitude, the length of the sampled data is 88160and the maximum amplitude is 0.9690 on the sampled data 84260 and the minimum amplitude is -1 on the sampled data 87930.

The next process is cutting the signal which is done twice for the signal from the normalization result. The first cut is made in the silence section or the empty signal section. The second cut is made at the transition. In the silence section signal truncation, data whose height is greater than 0.3 and less than (-0.3) is initialized as b0. Selected >0.3 and <0.3 because these signals are silence signals. Data that does not meet the requirements of b0 is part of the silence signal, so the signal is omitted ([ ]). The following is the result of the signal plot for the silence section.

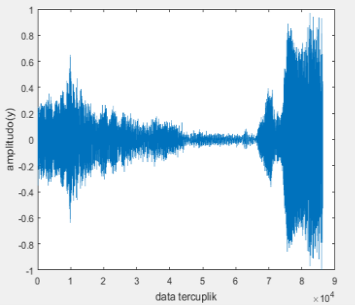

Figure 3. Plotting Results of Intersecting Signal 1 (Baby Sounds 1)

In plotting the results of clipping signal 1 above, the process of removing empty signals in the sense of signals that do not record sound from samples has been carried out. The x-axis is the sampled data and the y-axis is the amplitude, the length of the sampled data is 86300 before the length of the sampled data is 88160, meaning that the length of the sampled data is reduced after cutting this signal. The maximum amplitude is 0.969 in the sampled data 82350 and the minimum amplitude is -1 in the data sampled 86030. This proves that the amplitude remains the same as the amplitude in the previous normalization process, because the normalization process aims to equalize the amplitudes.

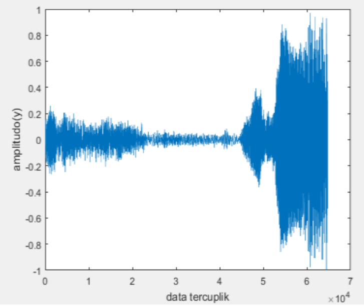

Figure 4. Result of Cutting Signal 2 (Baby Sound 1)

Furthermore, cutting the transition section is carried out by removing ¼ of the signal portion contained in the initial section. The signal is omitted ¼ part of the initial signal in order to obtain a sound signal that is actually recorded. The result of signal cutting 2 above is the process of cutting the transition signal part by cutting ¼ part of the initial signal after cutting the previous silence signal in order to obtain the desired frequency signal. The x axis is the sampled data and the y axis is the amplitude. The length of the sampled data becomes 64720 previously the length of the sampled data was 88160, because 2 transition parts have been cut. The maximum amplitude is 0.969 on the sampled data 60780 and the minimum amplitude is -1 on the sampled data 64490, the amplitude is still the same as the previous signal.

The next process is frame blocking which aims to retrieve some data according to the length of the frame blocking value. The value of frame blocking used is 256 because the higher the frame blocking the speech signal recognition will be more accurate. The retrieved data represents all recorded voice data. x2 is the result of frame blocking. The first process carried out in frame blocking is to determine the midpoint value of the signal clipped data (x1). Then determine the amount of data to be retrieved for the next process. In this process, data is taken starting from the leftmost signal and will be taken along the selected frame value, making it easier to calculate and analyze signals.

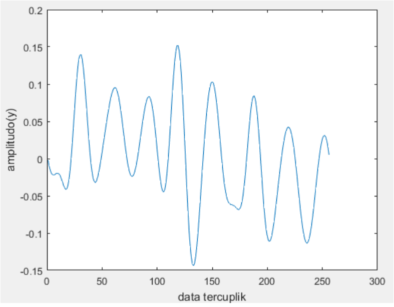

Figure 5. Plotting of Frame Blocking Results (Baby Sounds 1)

The x-axis is the sampled data and the y-axis is the amplitude, the length of the sampled data is 256 before the length of the sampled data is 64720, meaning that the length of the sampled data is reduced after frame blocking is done and this is deliberately done so that the calculation process, signal analysis and recognition of sound patterns become more easy. The maximum amplitude is 0.1517 in the sampled data 119 and the minimum amplitude is -0.144 in the data sampled 133, the frame blocking amplitude is reduced from the previous signal because the signal has been cut according to the length of frame 256 from the leftmost signal.

The final normalization is the normalization after going through the cutting and frame blocking processes. After going through frame blocking, the resulting signal is not optimal, so a final normalization process is needed. The normalization process will be carried out by dividing each data and input value, namely the result of frame blocking with the maximum absolute value of the input data. The purpose of the final normalization is to equalize the resulting frame blocking amplitude so that it is maximized.

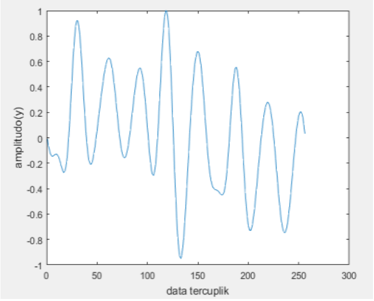

Figure 6. Final Normalized Signal (Baby Sounds 1)

The frame blocking results above are the final normalization process, which is to equalize the amplitude of the frame blocking results so that they are maximized. The x-axis is the sampled data and the y-axis is the amplitude, the length of the sampled data is 256 equal to the length of the previously sampled data in the frame blocking process. The length of the sampled data is the same because in the final normalization process it only aims to equalize the frame blocking amplitude so that it is maximized. The maximum amplitude is 1 in the data sampled 119 and the minimum amplitude is -0.9488 in the data sampled 133. The final normalized amplitude is higher than the previous amplitude because in this process the maximum amplitude value is sought.

hamming window has a function to eliminate the discontinuity effect caused by the previous process, namely frame blocking when the signal is transformed to the frequency domain. So that the sample that has been divided into several frames needs to be made into a continuous sound. This windowing process uses a hamming window. The following is the result of the hamming window.

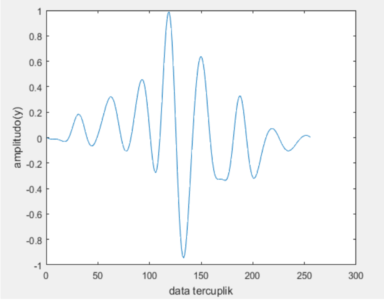

Figure 7. Hamming Window Results (Baby Sounds 1)

The result of the hamming window above is the process of eliminating the discontinuity effect. The x-axis is the sampled data and the y-axis is the amplitude, the length of the sampled data is 256. The maximum amplitude is 0.9875 on 119 sampled data and the minimum amplitude is -0.941 on 133 sampled data.

The next process is the FFT frequency spectrum. The FFT frequency spectrum is a process for obtaining a series of quantities in the recorded signal to determine a learning pattern or test pattern. The FFT calculation aims to find an absolute value which will then be analyzed to determine the pattern of the baby's crying sound.

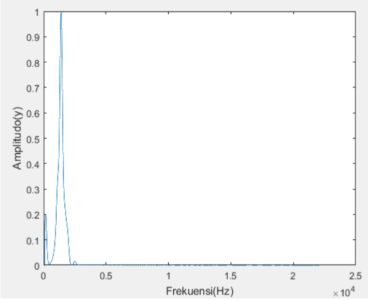

Figure 8. Frequency Spectrum (Baby Sounds 1)

The result of the frequency spectrum above is the final result in knowing the frequency of each sound sample. The x axis is the frequency and the y axis is the amplitude. The following is the frequency data for each sura that has been obtained using frame blocking 256.

Table 1. The Highest Frequency Value of Each Voice

Sample	Highest Frequency (Hz) of Sound		
	Baby	Mature	Object collision
1	1397	139	2233
2	1124	690	1010
3	145	666	226
4	1316	185	286
5	101	163	135
6	786	101	189
7	1863	669	182
8	589	154	101
9	132	1376	120
10	1104	720	258
Highest Frequency Average	**853.9**	**486.3**	**474**
Table Max Value	**1863**	**1376**	**2233**
Table Min Value	**101**	**101**	**101**
Range	**101-1863**	**101-1376**	**101-2233**

From the data obtained above, each frequency has been obtained for each sound sample with frame blocking 256. As the data above, the average infant frequency is 853.9 Hz, the average adult sound frequency is 486.3 Hz, and the average sound frequency object impact 474 Hz. The average frequency of a baby's cry is much higher and has a significant difference. With higher frame blocking it is easier to see the difference in the average frequency of each sample and finally to recognize patterns of crying babies based on frequency becomes easier.

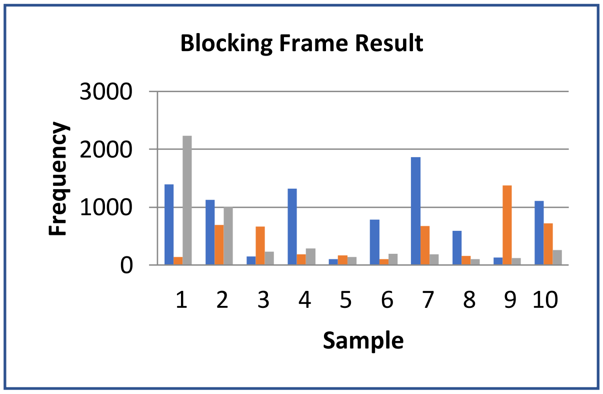

Figure 9. Voice Highest Frequency Data with Frame Blocking 256 (blue for baby, orange for adult, and grey for stuff)

From the data above, it can be concluded that the average frequency of a baby's voice is higher than the frequency of an adult's voice and the sound of colliding objects, so that from that frequency it can be recognized the pattern of the sound of a baby's cry in frequency using the FFT method which was carried out in this study.

CONCLUSION

This chapter contains the conclusions obtained from the identification stage of the recognition of baby crying sound patterns using the fast fourier transform (FFT) method and the conclusions from the frequency range results that have been obtained using the fast fourier transform (FFT) method. So the conclusion is as follows.

The steps for identifying sound signals in pattern recognition of a baby's crying sound are: sound sample recording, sampling, signal cutting, frame blocking, final normalization, hamming window, and finally the FFT calculation process. The frequency range for babies crying is 101-1863 Hz, for adults the frequency range is 101-1376 Hz and for the sound of colliding objects 101-2233 Hz.

REFERENCES

1. MP Brown and K. Austin, The New Physique (Publisher Name, Publisher City, 2005), pp. 25–30.
2. MP Brown and K. Austin, Appl. Phys. Letters 85, 2503–2504 (2004).
3. Fetra, Nicky. 2015. CHORD SEARCH APPLICATION TO ASSIST SONG CREATION USING THE FAST FOURIER TRANSFORM (FFT) ALGORITHM AND K-NEAREST NEIGHBOR (KNN) CLASSIFICATION METHOD. Riau. UIN Sultan Syarif Kasim.
4. Gunawan, D. 2011. DIGITAL SIGNAL PROCESSING WITH MATLAB PROGRAMMING. Jakarta. Science House.
5. Indriani, YH 2015. RECOGNITION OF BELIRA TONE USING AMPLITUDE ANALYSIS IN THE FREQUENCY DREAM. Yogyakarta. Sanata Dharma University.
6. Wahyudi,ST 2015. SPECTRUM ANALYZER APPLICATION TO ANALYZE AUDIO SIGNAL FREQUENCY USING MATLAB. Pekanbaru. Riau University.