Jurnal Mantik
Volume 4, Number 1, May 2020, pp. 709-717 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Implementation of Yin Algorithm to Detect Human Voice Emotions According to Gender

Fikri Aulia[1], Achmad Basuki[2], Bima Sena Bayu Dewantara[3]

[1,2,3,] Pascasarjana, Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya,

Kampus ITS, Jl. Raya ITS, Keputih, Kec. Sukolilo, Surabaya, 60111

Email: [1] micro18fsystem@gmail.com, [2]kisuki@pens.ac.id,[3]bima@pens.ac.id

ARTICLE INFO

Article history:
Received: 04/04/2020
Revised: 20/04/2020
Accepted: 30/05/2020

Keywords:
speech recognition,
Sound recording,
Yin algorithm,
Expression, Gender

ABSTRACT

Computer technology and artificial intelligence experience rapid development every year, one of which is in media speech recognition. Speech recognition is a virtual digital data assistant that exists in software applications, and is used as a tool to help human needs such as communication, but is often misused by users. This study conducted a voice recording to get the difference in the sound of each human gender data. The study uses the Yin algorithm to extract data, then the sound pitching process is performed using the histogram pitch feature of the standard deviation and the mean. From this study, it was found that the pitch of men is different from women. The shape of the pitch histogram contours is similar between men and women but the female pitch histogram shifts to a higher frequency than men. This pitch shift in women occurs in all expressions.

1. Introduction

*Virtual Digital Assistant (VDA)*is an automated software application that helps humans through the understanding of natural languages in written and oral. One technology that is often used in this application is speech recognition. Speech recognition is a communication media technology that is often used by humans to process every activity [1].

One speech recognition technology that is currently often used is Google's speech recognition is high and without the need to add noise cancelation on the client side with a database of more than 100 languages in the world.

Research that has been done in this field is from Team [2] by investigating the modeling of language and acoustic features for the classification of anger, the results obtained are an accuracy level of 78, 90%. Chi-Chun [3] used a hierarchical structure method to recognize the emotions of every human being by using the Support Vector Machine classification method (SVM). Elif [4] by proposing in his research is the use of the cepstral mel-frequency coefficient feature in weighting human emotional sounds. And produce an analysis that the features used have far better performance than standard spectral features. Mayank [5] with research increases the automatic emotion of speech by combining rhythm and temporal features. And the results of the research obtained are 66, 20% accuracy with angry emotional data.

Based on this research and problem, the speech recognition field is still very much needed. This research was conducted to distinguish human voices of each gender by using the Yin algorithm to extract data from human voices, and feature histograms with standard deviation values and the mean of each sound data. The results of this study obtained an analysis of each gender of human data.

2. Literature Review

2.1. Pitch

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Pitch or fundamental frequency which is defined as the lowest frequency of the periodic waveform. Most of the sounds of greeting music have a periodic structurewhen observed in a short time interval, and such sounds are perceived by the auditory system as having a quality known as pitch. Like loudness, pitch is the subjective attribute of sound that is related to the fundamental frequency of the sound, which is the physical attribute of the acoustic waveform [6]. To recognize the pitch method commonly used is cepstrum analysis and autocorrelation.

2.2. Gender Voice Character

The data states that there are differences in fundamental sound frequency (f0) in each speech with characters from age and gender. Data reported from each voice is an average of F0 of men at 120 Hz and women at 210 Hz. And values can change with age [7].

2.3. ADC (Analog Digital Converter)

Analog Digital converter is a system that converts analog signals. like sound taken by a microphone or light entering a digital camera, becomes a digital signal. The ADC converts analog signals with continuous amplitude and continuous time into digital signals of discrete time and discrete amplitude. Conversion involves input quantization [8].

2.4. YIN algorithm

The YIN algorithm is used for estimating the basic frequency (F0) of musical speech or sound. This is based on the well-known autocorrelation method with a number of modifications combined to prevent errors. This algorithm has several desired features. There is no upper limit on the frequency of the search range, so this algorithm is suitable for high-pitched sounds and music.

The YIN algorithm consists of several steps. Namely Autocorrelation to look for repetitive patterns in signal periods. Often used to find fundamental frequencies. The autocorrelation method compares the signal with itself shifting. In this case it is related to the Average Magnitude Difference Function (AMDF) method which makes comparisons using differences rather than products, and more generally for time domain methods that measure intervals between events in time.

*Differencefunction*is the stage to look for periods of autocorrelation. Difference function is useful for detecting wave crests.

Signal autocorellation results can be translated into formulas

$$x_t - x_{t+T} = 0,$$

Where xt is a periodic signal with T as a period.

To find out the peak of the wave formula is used.

$$\sum_{j=t+1}^{t+W} (x_j - x_{j+T})^2 = 0.$$

Where W is the window width.

*Cumulative mean normalized*The difference function is zero in zero lag and often non-zero in periods due to imperfect periodicity. Even if a limit is set, a strong resonance in the first formant might result in a series of secondary declines, one of which might be deeper than the period decline. Cumulative mean normalized is useful for removing the first Formant (F1).

*Absolute threshold*can easily occur if one decreases the level of higher functions deeper than the period decreases. If included in the search range, the result is a subharmonic error, sometimes called an octave error. The autocorrelation method also tends to choose high order peaks. Absolute threshold is useful for overcoming subharmonic errors.

Parabolic interpolation the previous steps function as described if the period is a multiple of the sampling period.

2.5. Statistical Visualization

A histogram is an accurate representation of numerical data distribution. A histogram is an estimate of the probability distribution of continuous variables. To create a histogram, the first step is to determine the "bin" ie the range of values, then divide the entire range of values into a series of intervals, then calculate how many values fall into each interval. The histogram gives a rough idea of the distribution density that underlies the data. to estimate the probability density function of the underlying variable.The total area of the histogram used for the probability density is always normalized to 1. If the length of the interval on the x-axis is all 1, then the histogram is identical to the relative frequency plot: Here is the formula of the histogram formula with the mean value and standard deviation

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3. Research Methodology

3.1. Design of Voice Data Retrieval Program

This system functions to retrieve voice data as well as the pitch process with the following flow.

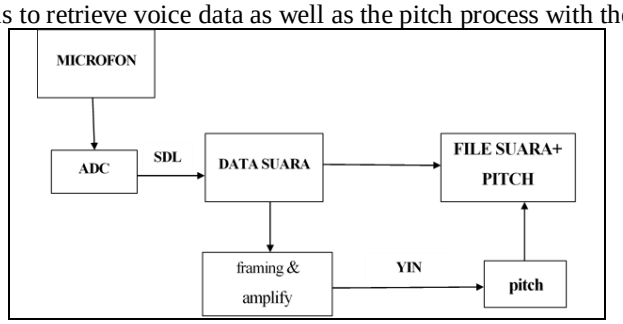

Fig 1. Retrieval of voice data

Figure 1 is a system starting from taking sound from a microphone and changing it in digital form. The SDL library is tasked with retrieving data from the buffer and then turning it into numeric type voice data. The voice data has a float data type. Before conducting the pitch recognition process using the YIN algorithm long voice data is divided into frames to facilitate data processing. Amplification is done so that the data can be processed in the YIN algorithm. The data from SDL is small. YIN algorithm cannot detect pitch from data whose value is too small. YIN algorithm outputs in the form of pitch with units of hertz (Hz).

3.2. Histogram Process Design

This process is to calculate the pitch that has been extracted. This system starts by reading the sound and pitch file and then taking the pitch part of the file. Furthermore, the pitch data taken is divided into bin with a range of 25Hz and then calculated into a histogram. Next is the output from the system that is saved as a histogram file.

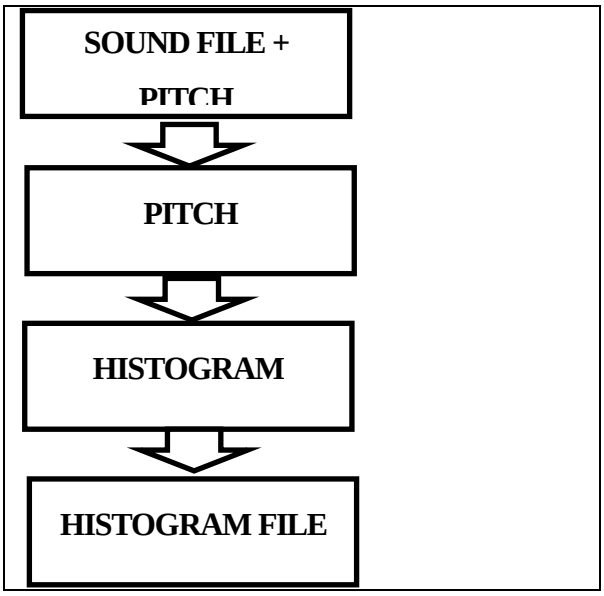

Fig 2. Histogram count process

The recorded file contains sound and pitch data. This file format is simple. The contents of the file start with voice data then followed by pitch data. Between voice data and data pitch there is the symbol "---" as a separator of voice data and data pitch. Voice data is 120000 lines and data pitch is 500 lines.

The next step to forming a pitch histogram is to extract the pitch data from the file. To retrieve data pitch, what needs to be done is to take data lines to 120002 to 120501.

The next step is to count the histogram. In the pitch data there will be a value of -1 which means there is no periodicity in the sound of the segment. A value of -1 indicates that the pitch cannot be detected and is not included in any bin. Histogram results are stored in a file and used as a dataset.

3.3. Data Characteristics

The data taken comes from the voice sampling of actors using a microphone. The voice of the actor is sampled for 10 seconds then the results of the sampling are stored in decimal form with a .txt file type. The contents of the data are the sampling tapes and the pitch extraction results. The data in lines 1 to 120000 are the recording data followed by the delimiter character "---" and the next 500 data are the extraction pitch.

119997	0.002259
119998	-0.00226
119999	0.002062
120000	-0.00191
120001	---
120002	-1
120003	158.554
120004	-1
120005	-1
120006	-1

Fig 3. Pitch value

Figure 3 is the data value of pitch -1 in the framte that cannot be recognized.

4. Results and Discussion

The research was carried out with several experimental stages with the following sections:

1. Retrieval of Sound Signals

This stage is the acquisition of voice data and recording, and the following is one of the results of the recorded data obtained.

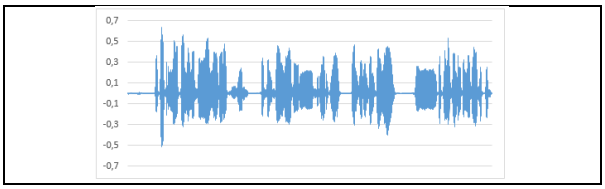

Fig 4. 10 second sound signal graph

Sound is recorded using a sampling frequency of 12000 Hz. **Error! Reference source not found.** is a sound recorded for 10 seconds so the number of n's is 120000. From **Error! Reference source not found.** it can be seen that the recorded sound data has a maximum amplitude of 0.6 and a minimum amplitude of -0.5.

2. Pitch Extraction

At this stage the sound signal is separated into several frames. Each frame has a size of 20ms. The sampling frequency is 12000Hz so one frame has 240 samples.

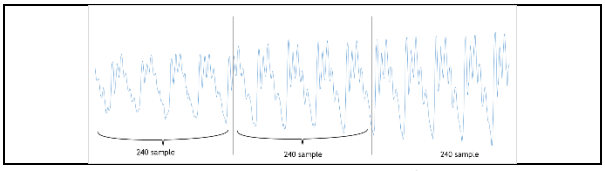

Fig 5. Sound signals are divided into frames every 240 samples

After that the signal is enlarged and processed using the YIN algorithm. The output of the YIN algorithm is the value -1 for the pitch not detected and the decimal value for the pitch detected.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

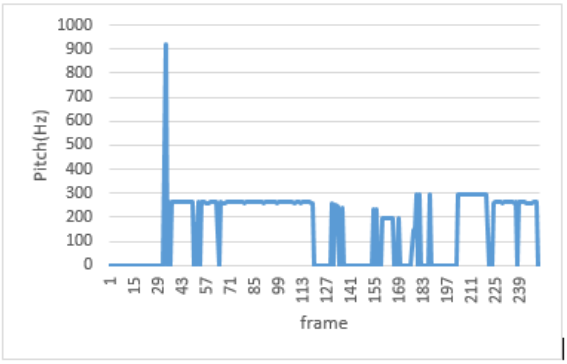

Fig 6. Virtual piano pitch

In Figure 6 is to prove that the YIN algorithm works as expected, then a small trial is performed by sounding a virtual piano with a C4 tone. The piano sound is recorded and the pitch is extracted. shows the virtual piano sound extraction pitch. The piano pitch mean calculation is 260.57 Hz. While the C4 tone frequency is 261.63 Hz. Frame 33 has a very high pitch frequency of 921 Hz. This is when the first virtual piano sounded the YIN algorithm was not able to recognize the transition pitch resulting in a high surge of frequency. Frames 38 through 118 on**Error! Reference source not found.**has a pitch between 261 Hz and 263 Hz. Exceptions are two parts, namely on frames 53 and 64. At that frame an error occurs and pitch is not detected.

3. Comparison of Human Voice Emotion Analysis

After getting a series of experimental processes, the following is the result of visualization of the emotional histogram of each gender in humans.

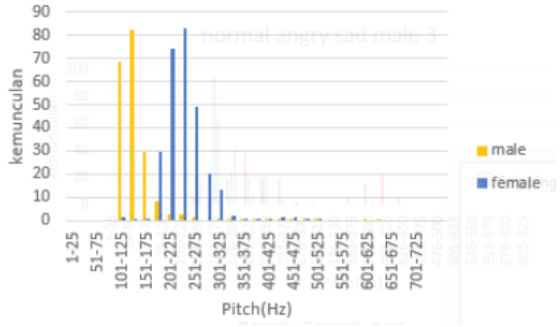

Fig 7. Graph of normal expression

Figure 7 is a comparison of normal emotions between men and women. The most common pitch in men is in the class 126-150 while the most frequent pitch in women is in the class 226-250. The frequency that most often appears in women is 100 Hz above the frequency that most often occurs in men.

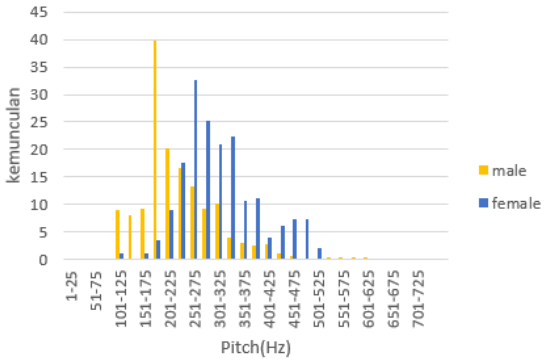

Fig 8. Graph of angry expressions

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Figure 8 is a comparison of normal emotions between men and women. The most common pitch in men is in the 176-200 class while the most frequent pitch in women is in grades 251-275. The frequency that most often appears in women is 75 Hz above the frequency that most often occurs in men.

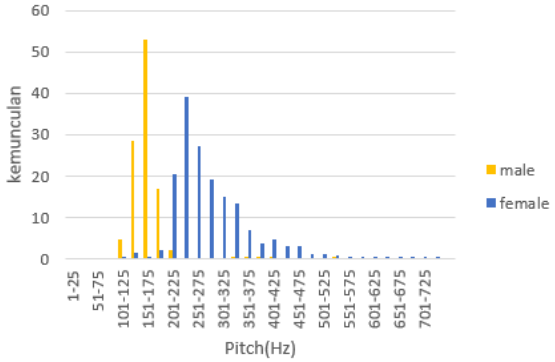

Fig 9. Sedi expression graph

Figure 9 is a comparison of normal emotions between men and women. The most common pitch in men is in the class 126-150 while the most frequent pitch in women is in the class 226-250. The frequency that most often appears in women is 100 Hz above the frequency that most often occurs in men.

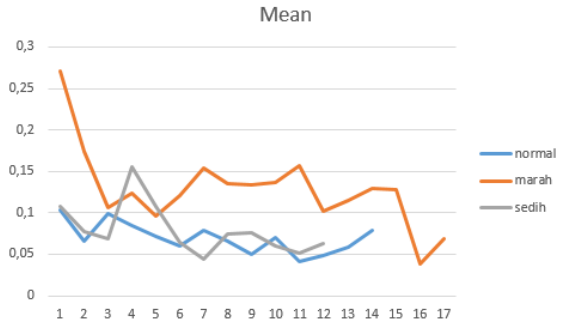

Fig 9. Average histogram

Figure 9 is a graph of the average histogram results of each individual sample. Y axis is the normalized average histogram. Whereas the X axis is an individual number. Not all individual samples express their emotions so there are individuals who do not have graphs.

Most of the samples show that the average angry histogram usually has a greater value than normal and sad. This shows that anger on the pitch frequency is higher than that on the individual 4 sad expressions have a higher value than anger and normal. This might be due to the sample expressing sadness with anger. Individuals 16 and 17 have a smaller average than normal expression. This may be because the emotion expressed is cold anger.

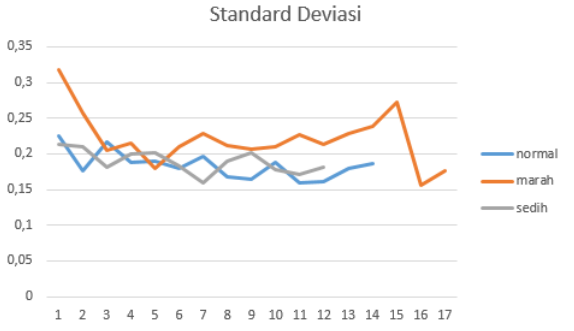

Fig 9. Standard deviation of the histogram

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Unlike the average, features with a standard deviation are more balanced / the same between one feature and the other. Angry features still look different from the others.

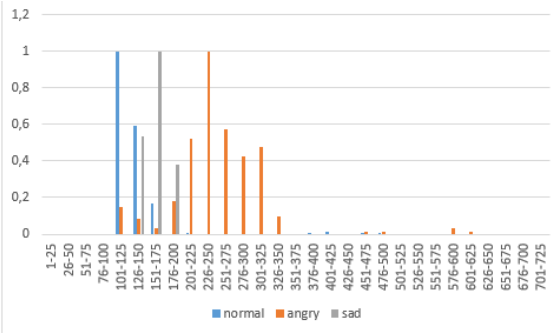

Fig 10. Normalization Results

Figure 10 isthe pitch that often appears in normal emotions is the 101-125 frequency class. The range of frequencies that can be observed from**Error! Reference source not found.**number 1 normal emotions are 101 to 175. Whereas the pitch that often appears on angry emotions is the frequency class 226-250. The range of frequencies that can be observed from angry emotions is 101 to 350. The pitch that often appears in sad emotions is the frequency class 151-175. The range of frequencies that can be observed from angry emotions is 126 to 200.

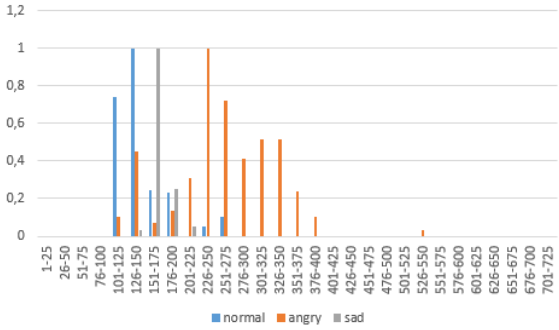

Fig 11. Normalization results

Figure 11 isa comparison between normal emotions, anger and sadness. The pitch that often appears in normal emotions is the 101-125 frequency class. The range of frequencies that can be observed from number 2 normal emotions is 101 to 175. While the pitch that often appears in angry emotions is the class frequency 226-250. The range of frequencies that can be observed from angry emotions is 101 to 350. The pitch that often appears in sad emotions is the frequency class 151-175. The range of frequencies that can be observed from angry emotions is 126 to 200.

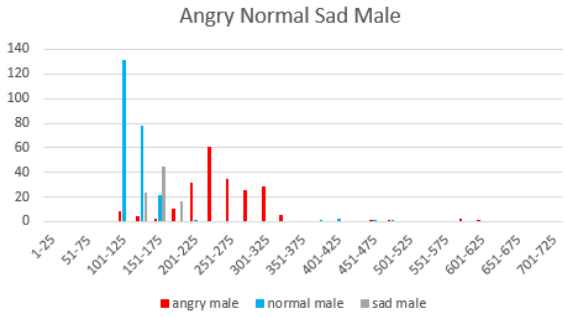

Fig 12. Comparison of normal emotions and angry male data 1

Figure 12 is a comparison between normal emotions, anger and sadness. The pitch that often appears in normal emotions is the frequency class 101-125 with the appearance value 131. The range of frequencies that can be

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

observed from the number 1 normal emotions is 101 to 175. While the pitch that often appears in angry emotions is the frequency class 226-250 with the appearance of 61 The range of frequencies that can be observed from angry emotions is 101 to 350. The pitch that often appears in sad emotions is the frequency class 151-175 with the appearance of 45. The range of frequencies that can be observed from angry emotions is 126 to 200.

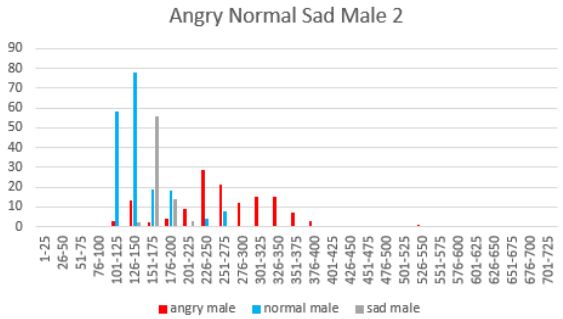

Fig 13. Comparison of normal and angry emotions in male data 2

Figure 13 is a comparison between normal emotions, anger and sadness. The pitch that often appears in normal emotions is the 101-125 frequency class with an emergence value of 131. The range of frequencies that can be observed from normal emotion images is 101 to 175. While the pitch that often appears in angry emotions is the 226-250 frequency class with the appearance of 61. The range of frequencies that can be observed from angry emotions is 101 to 350. The pitch that often appears in sad emotions is the frequency class 151-175 with the appearance of 45. The range of frequencies that can be observed from angry emotions is 126 to 200.

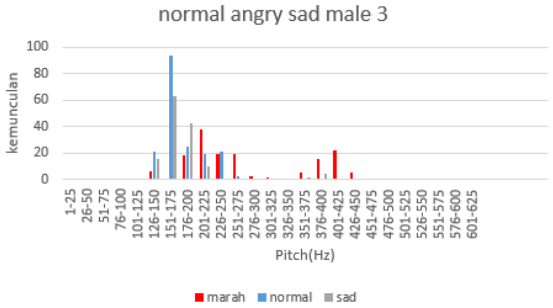

Fig 14. Comparison of normal emotions and angry male data 3

Figure 14 is a comparison between normal emotions, anger and sadness. The pitch that often appears in normal emotions is the 101-125 frequency class with an emergence value of 131. The range of frequencies that can be observed in normal emotion images is 101 to 175. While the pitch that often appears in angry emotions is the 226-250 frequency class with the appearance of 61. The range of frequencies that can be observed from angry emotions is 101 to 350. The pitch that often appears in sad emotions is the frequency class 151-175 with the appearance of 45. The range of frequencies that can be observed from angry emotions is 126 to 200.

5. Conclusion

From the results of this research are that the pitch of men is different from women. The shape of the pitch histogram contours is similar between men and women but the female pitch histogram shifts to a higher frequency than men. This pitch shift in women occurs in all expressions.

The advice that can be given in research is that further development is needed so that it can detect every human emotion given, so that it can be detected better.

6. Reference

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

[1] Bhargava, Mayank, and Tim Polzehl. 2012. "Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature." ICECIT.
[2] Banse, Rainer, and Klaus R Scherer. 1996. "Acoustic Profile in Vocal Emotion Expression." Journal of Personality and Social Psychology 70 (3): 614-636.
[3] Bozkurt, Elif, Engin Erzin, and Cigdem Eroglu Erdem. 2011. "Formant position based weighted spectral features." ScienceDirect.
[4] de Cheveigne, Alain, and Hideki Kawahara. 2002. "YIN, a fundamental frequency estimator for speech and music." Journal Acoustical Society of America 111.
[5] Deng, Jun, Xinzhou Xu, and Zixing Zhang. 2018. "Semi-Supervised Autoencoders for Speech Emotion Recognition." 26 (1): 31-43.
[6] Ghifary, M. T., & Faizah, N. (2019). PENGARUH KEPEMIMPINAN DAN KOMITMEN ORGANISASI TERHADAP KINERJA PEGAWAI DINAS PERINDUSTRIAN DAN PERDAGANGAN KABUPATEN PASURUAN. JURNAL EKBIS: ANALISIS, PREDIKSI DAN INFORMASI, 20(1), 1172-1180.
[7] n.d. Introduction to SDL 2.0. Accessed 7 12, 2019. https://wiki.libsdl.org/Introduction.
[8] McLoughlin, Ian. 2009. Applied Speech and Audio Processing. New York: Cambridge University Press.
[9 Polzehl, Tim, and Alexander Schmitt. 2011. "Anger Recognition in Speech Using Acoustic and Linguistic Cues." ScienceDirect.
[10] Potegal, Michael, Gerhard Stemmler, and Charles Spielberger. 2010. In International Handbook of Anger. New York: Springer.
[11] Rabiner, Lawrence R., and Ronald W. Schafer. 2007. Introduction to Digital Speech Processing. Foundations and Trends R in Signal Processing.
[12] 2016. The Virtual Digital Assistant Market Will Reach $15.8 Billion Worldwide by 2021. August 3. Accessed March 13, 2018. https://www.tractica.com/newsroom/press-releases/the-virtual-digital-assistant-market-will-reach-15-8-billion-worldwide-by-2021/.
[13] Syairozi, M. I., & Cahya, S. B. (2017). SUKUK AL INTIFAA: INTEGRASI SUKUK DAN WAKAF DALAM MENINGKATKAN PRODUKTIFITAS SEKTOR WAKAF PENDORONG INVESTASI PADA PASAR MODAL SYARIAH. JPIM (JURNAL PENELITIAN ILMU MANAJEMEN), 2(2), 12-Halaman.
[14] Syairozi, M., Rosyad, S., & Pambudy, A. P. (2019). PEMBERDAYAAN MASYARAKAT SEBAGAI PENGGUNA KOSMETIK ALAMI BERIBU KHASIAT HASIL PRODUK TANI UNTUK MEMINIMALKAN PENGELUARAN MASYARAKAT DESA WONOREJO KECAMATAN GLAGAH KAB. LAMONGAN. Empowering: Jurnal Pengabdian Masyarakat, 3, 88-98.
[15] Titze, Ingo R. 1989. "Physiologic and acoustic difference between male and female voices." Journal Acoustic Society of America 85.
[16] Watase, H. Nishizaki and K. 2017. "Emotion classification of spontaneous speech using spoken term detection." 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE). Nagoya.
[17] Zhang, Shiqing, and Shiliang Zhang. 2018. "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching." IEEE (IEEE) 20 (6): 1576-1590.