International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 2
ISSN: 2088-5334

Mel-frequencies Stochastics Model for Gender Classification based on Pitch and Formant

Syifaun Nafisah [#,*], Oyas Wahyunggoro[#], Lukito Edi Nugroho[#]

[#] *Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta 55281 Indonesia.*
E-mail:syifaun.nafisah@mail.ugm.ac.id

** Department of Library and Information Science, Islamic State University of Sunan Kalijaga, Yogyakarta 55281 Indonesia*
E-mail: syifaun@yahoo.com

***Abstract*—Speech recognition applications are becoming more and more useful nowadays. Before this technology is applied, the first step is test the system to measure the reliability of system. The reliability of system can be measured using accuracy to recognize the speaker such as speaker identity or gender. This paper introduces the stochastic model based on mel-frequencies to identify the gender of speaker in a noisy environment. The Euclidean minimum distance and back propagation neural networks were used to create a model to recognize the gender from his/her speech signal based on formant and pitch of Mel-frequencies. The system uses threshold technique as identification tool. By using this threshold value, the proposed method can identifies the gender of speaker up to 94.11% and the average of processing duration is 15.47 msec. The implementation result shows a good performance of the proposed technique in gender classification based on speech signal in a noisy environment.**

***Keywords*— Speech Recognition; Gender Identity; Mel-frequencies, Stochastic Model; Noisy Environment; Formants, Pitch.**

I. INTRODUCTION

Gender identity is defined as a personal conception of oneself as male, female or transgender [1]. This concept is intimately related to the concept of gender role, which is defined as the outward manifestations of personality that reflect the gender identity. Thus, gender role is often an outward expression of gender identity, but not necessarily so. It is important also to note that cultural differences abound in the expression of one's gender role, and, in certain societies, such nuances in accepted gender norms can also play some part in the definition of gender identity.

Gender identity can be a sensitive issue, so it is best to let other people tell about their gender rather than make assumptions, but it is impolite to take information about gender. Visual information plays an important role to process an information about him/her, such as gender, ethnicity and age. Various kinds of gender identity disorder such as transgender often make a visual information is deceptive. Voice is one of the most important biometric traits. By analyzing the voice we get a lot of information such as age, gender, ethnicity, identity, expression, etc. A gender classification system uses voice of a person from a given speech signal to tell the gender (male/female) of the person.

Voice (or vocalization) is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. Humans express thoughts, feelings, and ideas *orally* to one another through a series of complex movements that alter and mold the basic tone created by voice into specific, decodable sounds is called as speech. Speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration) to convey information about words, accent, expression, style of speech, emotion, the state of health of the speaker and speaker identity (name or gender identity). With the current concern, gender identity classification has received great deal of attention among of the speech researchers.

A gender classification system can be identify based on pitch, formants and combination of both. Pitch is a perceptual property of sounds that allows their ordering on a frequency-related scale[8] and formants derived from speech samples have been used for gender classification.

There are some studies that was related with gender classification. The first study is an autocorrelation, cepstrum and average magnitude difference (AMDF) methods that have been used for pitch determination from speech samples. The results of theses study was achieved the average of accuracy in the training phase is 74.2% and in the testing

phase is 74.4%. A vector quantization method also used as the method of gender identification. The average of accuracy using these method is 83.5% [2]. A Nearest neighbor is other method to identify the gender of speaker. In these method, euclidean distance was calculated from the mean value of males and females of the generated mean values of formant and pitch. The results of these method shows that the accuracy is up to 97.05% [3]. An automatic speech recognition system (ASR) also used to identify the gender of speaker. The combination between ASR and mel-frequencies cepstral coefficients (MFCC) produced the accuracy up to 98.8% [4] and the accuracy between ASR and fast fourier transform (FFT) algorithm is 80% on average [5]. Fuzzy logic and neural network also ever used to identify the gender of the speaker. To train fuzzy logic and neural network, training dataset is generated by using the above three features. Then mean value is calculated for the obtained result from fuzzy logic and neural network. By using this threshold, the performance is 57.5% [6].

Based on the previous research, this study was proposed a new method for gender classification using Mel frequencies stochastic model and back propagation neural network based on formant and pitch of speaker. In this paper, a general approach to identifying feature vectors that effectively distinguish gender of a speaker phoneme utterances is presented. Vowels and nasals are found to be effective in gender identification. They are relatively easy to identify in speech signal and their spectra contain features that reliably distinguish genders. The rest of the paper is structured as follows: The proposed technique with adequate mathematical models and illustrations are detailed in section Material and Research Method. The step of the experiments are explored in section Perceptual Experiment. The implementation results obtained are discussed in section Result and Discussion and conclusion of the paper in the last section.

II. Material And Research Method

A. Speech Data

In this research , the speech data was taken from 30 speakers, consisting of 19 male and 11 female speakers. The speaker ranging from 15-22 years of age. All of the speakers will be taken their voice through the recording process using Cool Edit 2.0. The recording was performed in a sound treated audiometric booth using vocal microphone PG48-LC, mini mixer EuroRack UB1002FX that was connected to the DAT recorder. The microphone had a flat frequency response ranging from 10 Hz to 20 KHz. The recoded speech was loaded into the computer through an M-Audio 32-bit DIO 2448 input/ouput card. Throughout the experiment, the mouth to microphone distance was carefully maintained at 1 inch from the left hand corner of the mouth. The data was segregated and individually stored as *.wav files.

The Speakers were asked to utter a set of 50 words in a normal manner which the utterance was repeated 12 times in a low-noise environment to reduce acoustic interference. From this process, there are 4394 wave files of data speech that will be stored in the system in the database such as described in Table 1.

TABLE I
THE FINAL FILES OF DATA SPEECH

Speech Data	Utterances
Female	2807
Male	1587
Total	4394

Then, the speech data were devided into two datasets as presented in Table 2.

TABLE II
GROUP OF DATA SET

Speech Data	Training (Dataset I)	Testing (Dataset II)
Female	1502	1305
Male	818	769
Total	2320	2074

The evaluation of the proposed model using the scenario in Table 3.

TABLE III.
THE SCENARIO OF EVALUATION

Scenario	Training (Dataset I)	Testing (Dataset II)
1	Dataset I	Dataset I
2	Dataset I	Dataset II
3	Dataset II	Dataset I
4	Dataset II	Dataset I

In this study, MATLAB 2010b was used to generate the spectogram of speech signal such as shown in Figure 1.

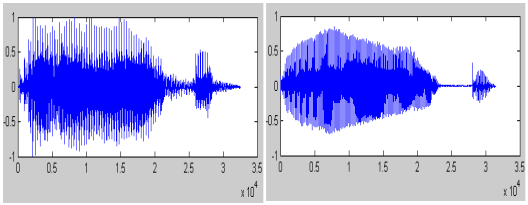

Fig 1. The difference of spectogram of word 'Ibu' between male and female speaker

B. Proposed Method

The main goal of this study is developing the new model to classify the gender of speaker. The block diagram in the Figure 2 shows the proposed model.

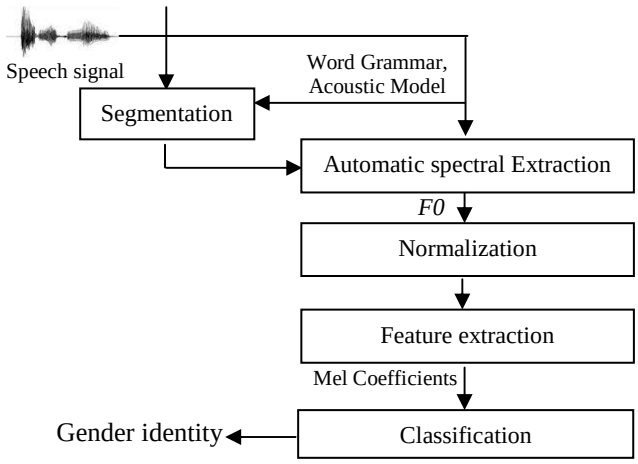

Fig 2. Block diagram of the proposed model

C. Segmentation

After the recording process, the data speech was segmented to get specific parts of the utterances by listen the wave file of each data to look for the boundary of the specific parts, then cut the wave file to extract the specific parts manually. For example, for the word "adik", the word will cut into "a" and "dik" waves files. After the segmentation step, from 4394 final wave files, there are 8728 wave files as data reference such as presented in Table 4 and which will stored in the database.

TABLE IV
THE FINAL WORDS OF DATA SPEECH

Speech Data	Segmented Files
Female	5878
Male	2850
Total	8728

D. Automatic Spectral Extractor

An automatic Spectral extraction was done to get the feature of pitch of the speech. In this step, to get the feature, the entropic wave software package was used to extract *f0* data from the word in each utterance. The *f0* implements a fundamental frequency estimation algorithm using a normalized cross correlation function and a dynamic programming function. This experiments used the default values of the parameters of *f0*, i.e., rectangle window with the length of 15 msec, and the shift time of 7.5 msec. The *f0* data are converted into logarithmic scale and passed through a normalizer.

E. Normalizer

In the normalizer, the normalization process was done to achieve robustness system againist the variation pitch contour from speakers. In this step, the log-energy output from filterbank values are normalized into frequency and time domains. Figure 3 shows the spectogram after normalization process.

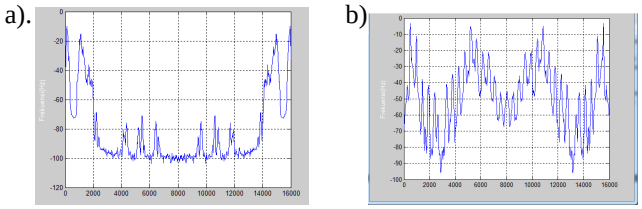

Fig 3. The result of normalization: (a) frequency domain and (b) time domain

If the sequences of the log *f0* is p_i = *(i=1,2,...,L)* and t_i = *(i=1,2,...,L)* is the time of the utterance with length L then the normalized pitch frequency of p_i, $\tilde{p}_i$ = *(i=1,2,...,L)* is calculated using Equation (1).

$$\tilde{p}_i = \frac{p_i - p_{min}}{p_{max} - p_{min}} \quad (1)$$

Where p_{max} and p_{min} are the maximum and the minimum log-energy output of the *f0* values of the utterance. The normalized time of t_i, $\tilde{t}_i$ = *(i=1,2,...,L)* is calculated using Equation (2).

$$\tilde{t}_i = \frac{t_1 - t_i}{t_L - t_1} \quad (2)$$

The normalized log of *f0* values are the input of the feature extractor.

F. Feature Extractor

In this process, MFCC was used to extract an acoustic features because ittakes human perception sensitivity with respect to frequencies into consideration, and therefore it is the best technique for speech recognition. The proposed model was designed based on MFCC which the known variation of the human ear's critical bandwidths with frequency filters spaced linearly at frequencies below 1 kHz and logarithmically at higher frequencies. Figure 4 shall explain the step-by-step computation of MFCC in this investigation [7].

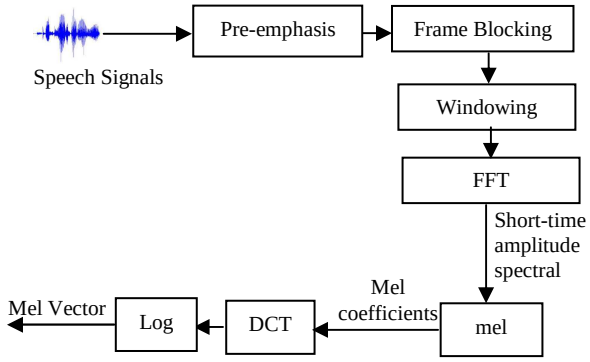

Fig 4. Block diagram of speech analysis procedure

The 1st step in MFCC algorithm is send the speech signal *s(n)* to a high-pass filter using the Equations (3).

$$s_2(n) = s(n) \quad a * S(n \quad 1) \quad (3)$$

Where $s_2(n)$ is the output signal and the value of *a*, which used in this study is between 0.9 and 1.0. The z-transform of the filter is calculated using Equation (4).

$$H(z) = 1 - a * z^{-1} \quad (4)$$

The goal of pre-emphasis is compensate the high-frequency part that was suppressed during the sound production mechanism of humans. The result of this process will be used as the input in the frame blocking process.

After the input signal was filtered in the pre-emphasis process, the signals were segmented into frames. The sample rate which used in this investigation is 44.1 kHz and the frame size is 1024 sample points, so the duration of each frames is:

1024/44100 = 0.02 sec = 20 msec

Based on this calculation, the speech data will segmented during 20 msec with overlap 50% of each frame. Additional, if the overlap is 512 points, then the frame rate is 44100/(1024-512) = 86.12 frames per second. In this process, the signals needed a zero padding process into the length value = 50000 as the nearest length of power of two frames.

The next process is windowing all of the frames. In this step, each frame has to be multiplied with a function of window to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by *s(n)*, *n=0,...,N-1*, then the signal after windowing is defined in the Equation (5).

$$s(n,a) = s(n) * w(n) \quad (5)$$

Where *w(n)* is the function of window. In this study, a rectangle window was chose as a function of window because it produce the highest accuracy than the other function such as presented in Figure 5. In practice, the value of a is set to 0.97.

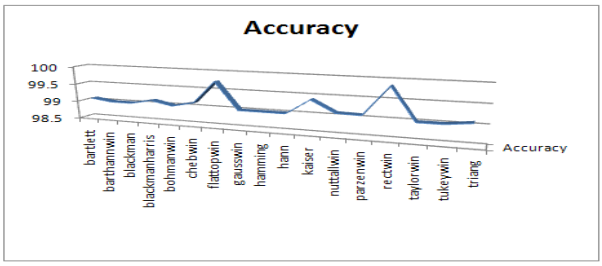

Fig 5. The accuracy based on the various function of windows

w(n) defined by using Equation (6).

$$w(n) = \text{RECT}\left[\frac{n}{0.97N}\right], 0 \leq n \leq N-1 \quad (6)$$

MATLAB provides the command RECT for generating the curve of a Rectangle window.

After the windowing process, spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. To obtain the magnitude frequency response of each frame, FFT was perform. In this process, the assumption is the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, the signals can still perform using FFT but the in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, the signals will be multiplied to each frame using rectangle window to increase its continuity at the first and last points.

If the input frame consists of three identical fundamental periods, then the magnitude frequency response will be inserted two zeros between every two neighboring points of the frequency response of a single fundamental periods. In other words, the harmonics of the frequency response is generally caused by the repeating fundamental periods in the frame. To extract an envelope-like features, the study use the triangular band pass filters, as explained in the next step.

After the coefficients were kept, this investigation will compute the DCT of the log filter bank energies. There are 2 main reasons this is performed. The first reason is the filter banks in this study are all overlapping and the second reason is the filter bank energies are quite correlated with each other. To compute the DCT of the log filter bank energies, the frequency of the signal should convert into Mel scale using the following equations:

$$M(f) = 2595 log_{10}(1 + f/700) \quad (7)$$

The result of this step is the diagonal covariance matrices can be used to model the features in the classifier. From this step, the matrices composed by 96 cepstral coefficients per feature which were consist of 47 MFCC coefficients, 47 MFCC delta features that indicate the degree of spectral change, one energy feature, and one delta-energy feature. For recognizer, cepstral-mean-subtraction (CMS) of the MFCC coefficients were done to remove some of the effects of noise.the result of this process are the input of the classifier.

G. Classifier

In this study, the backpropagation neural networks (BPNNs) was used as the classifier.The number of nodes in the input layer is equal to the number of features were extracted. The architecture of this classifier is 96 input nodes, 10 hidden nodes, and 2 output nodes which was trained using the reference model such asillustrated in Figure 6. The ouput values is [0:1]. The value [0] will set into male and the value [1] is female.

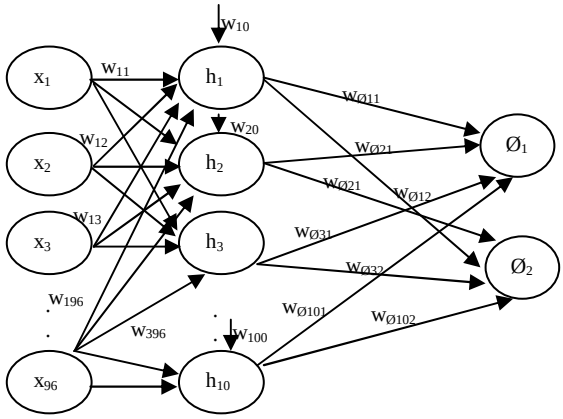

Fig 6. The architecture of BPNNs

The number of hidden layer in the ANNs classifier is compared using various numbers of hidden nodes. Based on the experiment, the number of hidden layer fixed to nine such as shown in Figure 7.

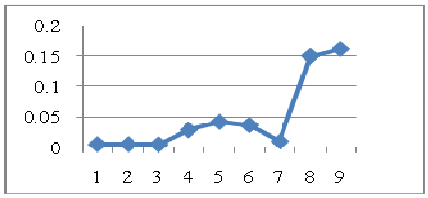

Fig 7. MSE using the various numbers of hidden nodes.

It concluded that the more number of hidden nodes in the hidden layer, the ANNs will be more accurately.

III. EXPERIMENTS

In this study, an acoustics stochastic model was constructed to evaluate the proposed model. This model was designed to classify the gender of the speaker based on the voiced speech. The voiced speech is different between male and female. The voiced speech of a male will have a fundamental frequency from 85 to 180 Hz and female from 165 to 255 Hz.A child's voice typically ranges from 250 Hz to 300 Hz and higher.

The general problem of fundamental frequency estimation is to take a portion of signal and to find the dominant frequency of repetition. The difficulties caused by all signals are periodic, the periodic may be changing in fundamental frequency over the time of interest, signals may be contaminated with noise, even with periodic signals of other fundamental frequencies, signals that are periodic with interval T are also periodic with interval 2T, 3T etc. Based on the difficulties, stochastic model was proposed to estimate the fundamental of frequency then the gender of speaker can be classified.

The essential steps in building stochastic models in this experiment consist of six step : (1) building the database of speech sample, (2) Identifying the sample of speech, (3) counting the spectral energy of the speech, (4) assigning probabilities to the sample speech, (5) identifying the events of interest and (6) computing the desired probabilities.

In this study, a model contains 2320 samples of data speech. First, the samples of data speech were extracted using MFCC, then all of the extracted files were stored in the database. The experiment, all of the data was trained using the feed forward back propagation algorithm. All of the extracted files were set as an input in the classifier and it trained to get the weights. The number of iterations use an epoch that was set as a variable. The variable was needed because the system iterated until all the errors were below the threshold of 0.5, or until the number of iterations reached 1,000,000. Every epoch comprised a variable number of back propagation iterations. The result of the test shows in Table 5-6.

TABLE V.
THE PERFORMANCE IN TRAINING PHASE

No	Data Training	Epoch	MSE	Time	Class
1	Abdul	69	9.86E-02	22	M
2	Alim	54	6.22E-03	21	M
3	Anwar	17	6.21E-03	8	M
4	Apri	20	3.80E-02	10	F
5	Ardi	28	3.25E-02	11	M
6	Arny	48	6.24E-03	19	F
7	Ekhsan	74	6.23E-03	27	M
8	Fenty	39	1.24E-02	15	F
9	Furqon	59	6.21E-03	20	M
10	Haqza	55	6.22E-03	21	M
11	Heri	57	1.24E-02	21	M
12	Lis	47	6.20E-03	16	F
13	Lutfy	51	7.36E-03	19	F
14	Rahma	34	2.58E-02	14	M
15	Ratih	21	6.21E-03	10	F
16	Ridwan	61	6.21E-03	22	M
Average		45.87	1.77E-02	17.25	93.75

TABLE VI.
THE PERFORMANCE OF TESTING PHASE

No	Data Testing	Epoch	MSE	Time	Class
1	Fitri	18	1.34E-02	8	F
2	Isti	49	2.07E-02	18	F
3	Hafidz	12	3.26E-02	7	M
4	Mamat	25	6.23E-03	11	M
5	Marisa	24	2.17E-02	9	M
6	Miftah	87	6.21E-03	29	M
7	Mukhlis	38	6.52E-03	15	M
8	Muyaz	36	8.21E-03	14	M
9	Nur	28	3.25E-02	11	F
10	Nuzul	49	6.28E-03	18	M
11	Rahmat	51	2.51E-02	18	M
12	Rofiq	45	1.19E-02	17	M
13	Sunarmi	66	2.09E-02	20	F
14	Susanto	52	1.24E-02	18	M
15	Tifano	49	2.48E-02	18	M
16	Tri	47	2.58E-02	18	F
17	Yasir	38	2.14E-02	14	M

IV. RESULT AND DISCUSSION

The evaluation of the proposed model was done using several steps. The first step is evaluating this model using the data training. After the evaluation using data training get the results as expected, the evaluation is continued using the testing data. All of the evaluation was done to ensure the dependability of this model in real use. Based on the experiments that has been done such as presented in Table 5 and Table 6, the result can be described as follows:
The evaluation of the proposed model using the data training is generate the accuracy of the model up to 93.75% and the processing time is 17.25 msec. The performance of the model are presented in Figure 8.

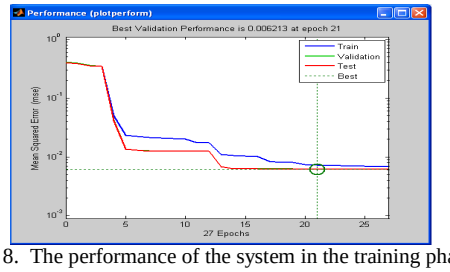

Fig 8. The performance of the system in the training phase

After the model was trained using the back propagation algorithm and get the evaluation results that the error can be tolerated, the evaluation is continued using the data testing. The result of the evaluation such as presented in Table 6 shows that the accuracy of the proposed model is 82.35 % with the processing time during 15.47 msec. Figure 9 shows the performance of the proposed model.

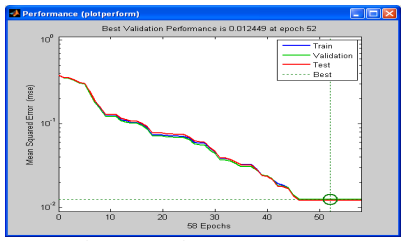

Fig 9. The performance of the system in the testing phase

In Figure 9 looks that the performance of the model is increased. It can be seen from the pattern of the data in training process, reference data and data testing which is located on almost in one graph.

V. CONCLUSIONS

The mel-frequencies stochastic model for gender identity classification is a system that was designed to evaluate the accuracy of the system to classify the gender of speaker

based on their speech. The difficulties that was encountered in this process caused by an irregular pattern of frequencies of human voice. The pitch was used to overcome the difficulties in this research. The pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale. It is compare the higher and lower frequencies. The pitch is require a clear and stable of frequencies to distinguish from noise.

Based on the experiment which has been done, the capability of the proposed model to classify the gender identity in a noisy environment is 94.11% with the processing time during 15.47 msec The back propagation algorithm that is combined in this model can cut down the processing time up to 1.78 msec. The increase of performance can also be seen from the average of the particular period of time marked by distinctive features (epoch) that are becoming shortly. The epoch of process is decreased from 45.875 being 42. It means that the iteration of process is shorter. It is influence the processing time so it can faster. If the result was compared, it is clearly visible that the efforts of the proposed method can improve the performance such as describe in Table 7.

TABLE VII.
COMPARATION PERFORMANCE BETWEEN PROPOSED METHOD AND OTHER METHOD

Method	Accuracy
AMDF [2]	74,44%
Vector quantization [2]	83.5%
Euclidean distance [3]	97.05%
MFCC [4]	98.8%
FFT [5]	80%
Neurofuzzy [6]	57.5%
Stochastic model	94.11%

In table 7 above, It is clearly visible that the accuracy of stochastic models is still lower than using MFCC, but in these experiments, the proposed model was tested under a noisy environment. It is hoped, if our proposed model was used in a conducive environment, it is expected that these model will produce a higher accuracy. It is expected, if the proposed model can be improve in training phase will produce the higher accuracy and shorter of the processing time. So, the proposed model can be used as one of the model for authentication system of the speaker not only limited to classify the gender identity.

ACKNOWLEDGMENT

The authors would like to thank for the support given to Ministry of Religious Affairs of the Republic of Indonesia for the scholarship of doctoral degree program to the first author.

REFERENCES

[1] Anonim. *The Guidelines for Psychological Practice with Lesbian, Gay, and Bisexual Clients* (American Physcological Association, 2011, 18-20 February).

[2] Kumar, P., Jakhanwal, N., Bhowmick, A., & Chandra, M. *Gender classification using pitch and formants.* International Conference on Communication, Computing & Security (ICCCS) India. 319-324. 2011.

[3] Rakesh, K., Dutta, S., & Shama, K. *Gender Recognition Using Speech Processing Techniques in LABVIEW*. International Journal of Advances in Engineering & Technology (IJAET) , 1 (2), 51-63. 2011.

[4] Deiv, D. S., Gaurav, & Bhattacharya, M. *Automatic Gender Identification for Hindi Speech Recognition.* International Journal of Computer Applications, 31 (5), 1-8. 2011.

[5] Ali, M. S., Islam, M. S., & Hossain, M. A. *Gender Recognition System Using Speech Signal*. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2 (1), 1-9. 2012.

[6] Meena, K., Subramaniam, K., & Gomathy, M. *Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network.* The International Arab Journal of Information Technology, 10 (5), 477-485. 2013.

[7] Huang, X., Acero, A., & Hon, H. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall. 2001.

[8] ^ Anssi Klapuri. *Introduction to Music Transcription in Signal Processing Methods for Music Transcription*, edited by Anssi Klapuri and Manuel Davy, 1–20 (New York: Springer, 2006): p. 8. ISBN 978-0-387-30667-4.

[9] Rahman, S. A., Omar, N. B., Mohamed, H., & Aziz, M. J. (2011). A Synonym Contextual-based Process for Handling Word Similarity in Malay Sentence. *Proceeding of the International Conference on Advanced Science, Engineering and Information Technology (ICASEIT 2011)*, (pp. 248-252). MAlaysia.

[10] Noor, A. O., Samad, S. A., Hussain, A., & Fauthan, A. (2011). A New Voice Controlled Noise Cancellation Approach. *Proceeding of the International Conference on Advanced Science, Engineering and Information Technology (ICASEIT 2011)*, (pp. 380-390). Malaysia.