Voice Identification Using Neural Network And Mel Frequency Cepstrum Coefficients

Corianti GMS[1], Fahmi[2], Maksum Pinem[3], Sihar P. Panjaitan[4], Suherman[5]
{coriantisimamora@gmail.com[1], fahmimn@usu.ac.id[2]}

[1,2,3,4,5]Electrical Engineering Department, Universitas Sumatera Utara, Indonesia

Abstract: Voice as a communication media for human and computer communication is developed by using voice recognition/identification technology. One of its applications is differentiate between male and female voices. Some speech recognition research have been proposed in disseminating male and female voices, this paper performs male and female voice extraction by using Mel Frequency Cestrum Coefficients (MFCC) as the characteristic vector in back propagation artificial neural network (ANN). The weight change cycle or EPOCH (Exponential Decay) is used to initialize male and female voice identification by using multiple recorded voices. As results, the female voice identification is better than of male with error less than 1.

Keywords: Voice recognition, Mel Frequency Cestrum Coefficients, Artificial Neural Network.

1. Introduction

Voice is one of the most popular and effective communication for human being. Voice can also be utilized for human-computer communication, by using speech to text conversion, for instant. This technology has been widely practiced in various languages and has been claimed and commercialized(Mahboob *et al.*, 2015).

The science of speech recognition is known as digital voice processing and natural language processing. Signal processing has been the popular method in solving voice recognition problems and mobile application (A IM Dunia, 2018).

This paperfocuses on the voice identification and differentition between male and female by using Mel Frequency Cestrum Coefficients(MFCC) from some recorded .wav voices of male and female.The recording was performed in a quiet environment to reduce noises.

Some existingworks, for instant, Agus Harjokoexamined voices by using MFCC (Abriyono and Harjoko, 2012) to extract voice features. The same research have also been performed by Prasad (Borde, 2015) andChan (Chan *et al.*, 2016) by using neural network.

2. Material and Method

In order to differentiate male and female voice, this paper employs four main processes, namely: the process of recording voice data, pre-processing process, feature extraction process, and identification process.

The recording process is the process of recording a human voice saying the word in Bahasa Indonesia. The recording process is carried out under quiet conditions by using Sennheiser PC110 micophone.

The pre-processing process is a process for initial processing of record data by reducing the effects of recording and amplifying the digital recording signals, and preparing the data in

ICASI 2019, July 18-19, Banda Aceh, Indonesia

DOI 10.4108/eai.18-7-2019.2288519

the proper form for feature extraction processes by using (BenZeghiba, 2005). The re-processing involves the normalization process of voice amplitude; pre-emphasis; framing process; windowing process and fourier transform process.

The feature extraction process is the process of taking certain values from a digital signal that can represent a digital signal(Lf *et al.*, no date). Proper character picking enablesclassification process between onevoice to another, and ultimately improvesthe accuracy of recognition.

The characteristic extraction process carried out by using MFCC characteristic extraction. The last process is the process of recognizing a sound signal. The identification process usesartificial neural network withbackpropagation model.The identification process requires the training process, followed by test process. The process sequences are shown in Figures 1 and 2(Mansour, Salh and Mohammed, 2015).

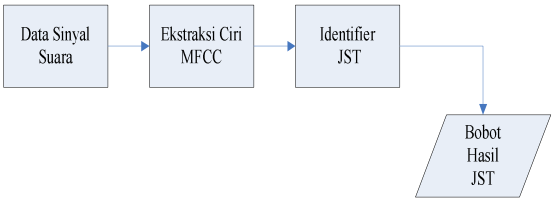

Fig. .1. Recognition System Flowchart Training Scheme

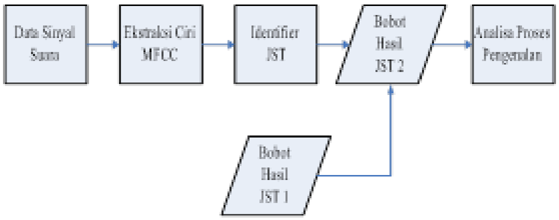

Fig.2. Recognition System Flowchart Testing Scheme

The results of the training scheme are in the form of ANN weights and reused in the test scheme(Mulyawan, Samsono and Setiawardhana, 2011).

2.1 Digital Signal Data

The recorded data contains speech words of Indonesian words. The oral spoken word is an Indonesian word fragment based on the stroke on the pronunciation of the words. The use of Indonesian spoken language syllables is performed to reduce the amount of data to be recognized without reducing the number of Indonesian words available.

Data recording was done at asampling frequency (Fs) of 11000 Hz, mono, and 8 bits encoded. The sound sourceswere taken from two people. Each person pronounced every syllable of spoken language.

2.2 Amplitude Normalization

Amplitude normalizationis the process to normalize the degradation of digital signal sample value due to the differences in distance between the mouth and the recording microphone. The amplitude normalization process is obtained by dividing all samples with the maximum absolute value of the digital signal samples.

2.3 Pre-emphasis

The pre-emphasis process is a process designed to reduce the adverse effects of transmission and background noise. The pre-emphasis process is excellent in reducing the effects of distortion, attenuation, and saturation of the recording medium.

2.4 Framing and Windowing

Framing is the process of disconnecting digital signals into a shorter time group. The framing process begins by specifying the time zone of each frame (Nw) and the frameshift area (Ns). The framing illustration can be seen in Figure 3. Windowing is a formulation to weaken the value at both ends of a digital signal.

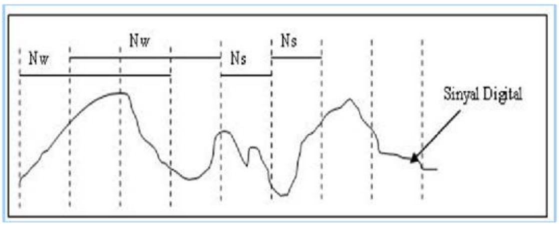

Fig. 3. Framing ProcessIllustration

2.5 MFCC

Generally, the perception of the frequency of human hearing is basically not linear as in general frequency modeling. This perceptual modeling named is Mel-scale,and its frequency is called the mel frequency. The final step of characteristic extraction process uses the formulation of discrete cosine transform to obtain the value of its MFCC coefficient(Chakraborty Asmita Talele Savitha Upadhya, 2014).

In this last stage, the coefficients are used as the features of classification. The minimum number of coefficients commonly adopted for the speech recognition process is 8 and the maximum number of filters defined. The zero coefficient of DCT, in general, will be omitted, even if it actually indicates the energy of the frame signal.

2.6 Backpropagation

Artificial neural network of backpropagation is a method to classify and identify a particular input pattern by improving the weights of aninterlayer between layers. Simple neural networks were first introduced by McCulloch and Pitts in 1943. McCulloch and Pitts concluded the combining of several simple neurons into a neural system that would improve their computational ability. Network weights proposed by McCulloch and Pitts are set to perform simple logic functions. The activation function used is the threshold function.

Artificial neural networks are information processing systems that have characteristics similar to biological neural networks(Dongare, Kharde and Kachare, 2012).

The training on the backpropagation method involves 3 stages: the feedforward training pattern, error counting,and weight adjustment. After training the network application using only the first stage of the computation of feedforward. The network can produce output very quickly, although the training phase is very slow. Backpropagation methods have been varied and developed to improve the speed of the training process.

Speech recognition application of the recorded voice is manifested into software using Matlab 6.5 which is shown using the flowchartdiagram in Figure 4.

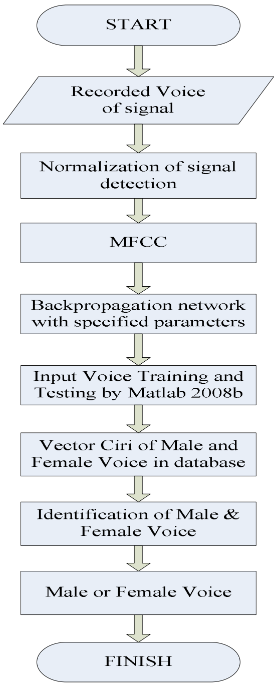
START
Recorded Voice of signal
Normalization of signal detection
MFCC
Backpropagation network with specified parameters
Input Voice Training and Testing by Matlab 2008b
Vector Ciri of Male and Female Voice in database
Identification of Male & Female Voice
Male or Female Voice
FINISH

Fig. 4. Flow Chart ofmale and female Voice Identification

3. Results And Discussion

In the framework of training and testing to know the ability of the syllables recognition of Indonesian language, test was performed following the stepson Figures 1 and 2.

Some parts of the system require the exact parameter value that should be determined at first place. These parameters are the value of the pre-emphasis coefficient: 0.95, the number of filterbank on MFCC characteristic extraction: 40, the number of training: 8, and FFT: 512.

These specified parametersexert better recognition performance. The voice recognition runs ona laptop equipped by a voice recording device (Sennheisser PC 110 mic), MATLAB 6.5 and visual basic 6 software. The amount of training data explains the ability of recognition the Indonesian spoken word. Vector characteristic of MFCC was obtained by anumber of coefficients 8 with the number of voice samples: 3 female voices and 3 male voices.

The next step is to cut the voice signal in certain frame size. The length of the tested frame is 256. Each frame that has been obtained was multiplied by the hamming window. The FFT process wasperformed to change each frame fromtime domain to frequency domain that produces afrequency spectrum. The MFCC output coefficient is the end result of the feature extraction process which is shown in Table 1 and Table 2.

Table 1. Characteristics of Voice Female Vectors by the number of Coefficients 8

Voice	C_n=8							
Apa Kabar	3.0476 3.0476 3.0476	2.0393 2.0393 2.0393	-4.4242 -4.4242 -4.4242	-4.4484 -4.4484 -4.4484	-2.1844 -2.1844 -2.1844	0.5254 0.5254 0.5254	-0.1185 -0.1185 -0.1185	0.5810 0.5810 0.5810
Selamat Malam	1.3952 1.3952 -0.2282	1.2975 1.2975 0.8108	1.1795 1.1795 -0.7328	0.6141 0.6141 -0.3643	-0.1084 -0.1084 0.3281	-0.6147 -0.6147 -0.5679	-0.6728 -0.6728 -07891	-0.3217 -0.3217 -0.0575
Selamat Pagi	1.4196 3.2675 2.3056	0.8408 -0.7399 1.7859	1.1024 4.5034 1.2685	-0.1453 -0.6672 -1.2152	-0.8463 -2.2294 -3.2675	-0.6549 0.1377 -14435	-0.8738 -2.1410 1.1225	-0.1221 0.0366 0.0553
Hallo	4.0513 5.5189 1.9123	5.1681 5.0799 0	-1.0440 -4.4324 0	-6.6598 -14.1351 0	-2.4112 -2.4108 0	0.9011 3.9784 0	-1.3025 -0.2347 0	-0.6497 -0.0591 0

Table 2. Characteristics of Voice Male Vectors by the number of Coefficients 8

Voice	C_n=8							
Apa Kabar	-0.4856 2.8441 3.6138	-0.2580 -0.4699 -2.7616	-6.0429 -10.4210 -11.8552	-0.2892 -2.0463 -2.5799	-1.5972 -3.9650 -2.7117	0.1343 2.3927 2.4599	0.6174 -0.7388 0.0637	0.254 0.5470 0.4953
Selamat Malam	4.1740 2.2334 1.0559	-2.5970 -4.0627 -3.8161	-9.5392 -3.2830 -2.6404	-2.9577 1.3270 1.2761	-5.7386 -0.6440 -1.5407	5.4921 3.8757 4.3728	-2.1221 -3.8071 -2.7898	0.1443 0.4591 0.5396
Selamat Pagi	3.2764 3.6810 1.7933	-4.4997 -6.1433 -7.4391	4.8008 7.8260 6.3208	0.3011 -0.2127 3.0253	-9.8272 -9.5365 -8.2103	-1.4573 -0.8229 1.1666	-0.6246 -1.7773 -2.3823	0.0687 0.1168 0.0432
Hallo	1.7933 1.7867 1.4641	-7.4391 3.9700 0	6.3208 -5.9165 0	3.0253 -8.5696 0	-8.2103 -2.3563 0	1.1666 2.3290 0	-2.3823 0.4173 0	0.0432 0.1315 0

In general, more MFCC coefficients involvedresult better recognition performances as more feature information generated. The result ofneural network output coefficient which is the error max of the signal percentage is shown in Table 3.

Table 3. The Result of Weight by The number of Coefficients 8

Gender	Voice	EPOCH	Max Error	EPOCH	Max Error
Pria	Apa Kabar	1-71	< 1	72-600	1
	Selamat Malam	1-66	< 1	67.600	1
Wanita	Selamat Pagi	1-67	< 1	68.600	1
	Hallo	1-102	< 1	103-600	1

The differences of weight cycle resultin voice identification between male and female by usingArtificial Neural Network (ANN) with the word “Selamat Pagi” and “Apa Kabar” are shown in Figure 5 and Figure 6.

In JST Backpropagation, there are several parameters that used as an input to determine the performance of ANN in processing. Some parameters are the frequency of training, training rate constant, input layer value, and hidden layer.

Figure 5 and 6 presents the influence of the weightchange to voice identification between male voice and female voice.

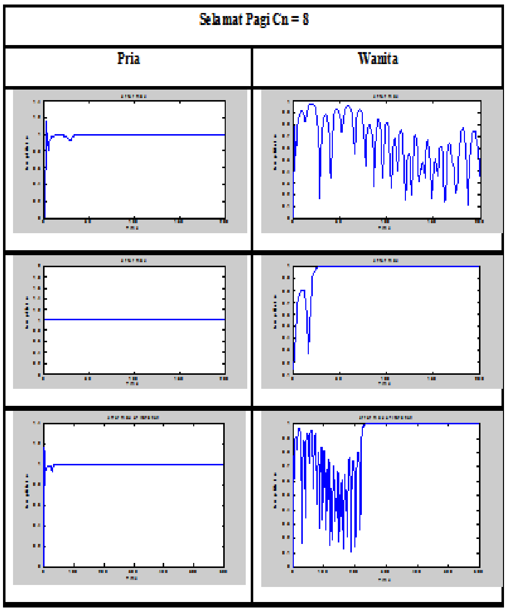

Fig. 5. Male and FemaleDifferences for"Selamat Pagi."

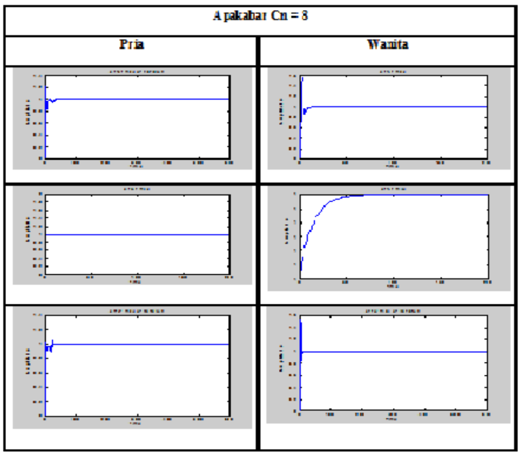

Fig. 6. Male and FemaleDifferencesfor"Apakabar."

The weight change cycle is performed by changing the number of feature coefficients (Cn) as the input of artificial neural networks. Based on the training data in the Figure, the error rate exceeds the error limit and the error value is less than one(0.05 <error <1), with the amount of training data on male and femalevoices are 3, the frame length 256, the number of Mel Bank Filter 40, the number of coefficients 8, FFT 512 and the value of α is 0.2.

The artificial neural network iterationfrom the available voice samples in thedatabaseis ableto identify female voices better than male. The error of female voices detection is smallerthan male voices.

4. Conclusion

Based on the experiment, the MFCC characteristic extraction is not suitable in recognizing voices with a large number of targets. However,it suits small number of voice data. The recognition abilityis better when the amount of training data from the number of feature coefficients is larger.If target error ($0.05 \leq$ target error ≤ 1) is higher, the cycle of weight (EPOCH) is lower.

The level of speech recognition is influenced by some external and internal factors, including microphone, recording device, voice signal process and equipment conditions. The internal factors are physical and psychological conditions of voiceintonation, gender,and age of respondents.

In order to improve the ability of theIndonesian syllablesrecognition, an initial classification should be performed at first place.

References

[1] A IM Dunia, Suherman, A. H. R. and R. F. (2018) ‘Measuring the power consumption of social media applications on a mobile device’, *J. Phys.: Conf. Ser.*, 978(1), p. 012104.

[2] Abriyono and Harjoko, A. (2012) ‘Pengenalan Ucapan Suku Kata Bahasa Lisan Menggunakan Ciri LPC, MFCC, dan JST’, *Indonesian Journal of Computing and Cybernetics Systems*, 6(2), pp. 23–34.

[3] BenZeghiba, M. F. (2005) ‘Joint speech and speaker recognition’. doi: 10.5075/epfl-thesis-3193.

[4] Borde, P. (2015) ‘Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition’, *International Journal of Speech Technology*, 18, pp. 167–175.

[5] Chakraborty Asmita Talele Savitha Upadhya, K. (2014) ‘Voice Recognition Using MFCC Algorithm’, *International Journal of Innovative Research in Advanced Engineering*, 1(10), pp. 2349–2163.

[6] Chan, W. *et al.* (2016) ‘Listen, attend and spell: A neural network for large vocabulary conversational speech recognition’, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964. doi: 10.1109/ICASSP.2016.7472621.

[7] Dongare, A. D., Kharde, R. R. and Kachare, A. D. (2012) ‘Introduction to Artificial Neural Network’, *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(1), pp. 189–194.

[8] Lf, M. I. *et al.* (no date) ‘Aplikasi Pengenalan Ucapan Dengan Jaringan Syaraf Tiruan Propagasi Balik Untuk Pengendalian Robot Bergerak’, pp. 1–7.

[9] Mahboob, T. *et al.* (2015) ‘Speaker Identification Using GMM with MFCC’, *IJCSI International Journal of Computer Science Issues ISSN ISSN*, 12(2), pp. 1694–814.

[10] Mansour, A. H., Salh, G. Z. A. and Mohammed, K. A. (2015) ‘Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms’, *International Journal of Computer Applications*, 116(2), pp. 34–41.

[11] Mulyawan, H., Samsono, M. Z. H. and Setiawardhana (2011) ‘Identifikasi Dan Tracking Objek Berbasis Image’, pp. 1–5.