MIXED TYPE AUDIO CLASSIFCATION USING BACKPROPAGATION WITH FIXED SIZE MOMENTUM

Dicky Teguh Satrio[1], Nelly Indriani W., S.Si., M.T.[2]

[1] Teknik Informatika – Universitas Komputer Indonesia
Jalan Dipatiukur No.112-116, Coblong, Bandung, Jawa Barat 40132
E-mail : dicky.teguh48@gmail.com[1], nelly.indriani@email.unikom.ac.id[2]

ABSTRACT

Mixed type of audio is an audio file that has more than one type of sound sources, such as speech or human voice combined with music, or speech combined with environment. Mixed type audio previous research concludes that Support Vector Machines and Backpropagation produce the best accuracy that Support Vector Machines Backpropagation 90.84% and 90.07% for the class speech and music, but for a speech class and environment of accuracy obtained is a Support Vector machines Backpropagation 64.76% and 63.81%.

This study show how to classify mixed type audio using backpropagation with fixed size momentum. Use of fixed size momentum is intended to improve the accuracy of the method backrpopagation. For this research, testing is done by testing the algorithm parameters with some variation of condition parameters. Parameters that produce highest accuracy is five neuron in the hidden layer, learning rate 0.3, windows size 3. The results obtained from the testing of several parameters that have been optimized to produce 95% accuracy using a fixed size backpropagation with momentum.

Based on the results of testing that has been done, it is concluded that the classification of mixed types of audio using backpropagation with momentum fixed size can improve accuracy by 95% when compared to backpropagation are getting an accuracy of 80% for the classes of speech with the environment and speech with music.

Keywords **:** *Mixed Type Audio, Backpropagation, Fixed Size Momentum, jAudio, Windows size*

1. INTRODUCTION

In general audio is a representation of the sound or the sound, while the sound is a longitudinal wave propagated through intermediaries. As for the digital audio easily is an alternative way in the present wave of audio in digital form that is in binary. Mixed Type Audio is an audio file that has more than one type of sound source, such as a speech or sound a humans combined with music, or combined with the sounds of the surrounding environment. Various types of audio ever gave rise to the need for an approach to be able to classify the type of audio based on the content of the audio. To do so required audio files which are processed so that the conclusion to be drawn what type of audio.

The research of Lei Chen and coauthors [1] do classification on mixed type i.e. audio speech with music and music with the environment using SVM classifier and ZCR as as feature extraction. The study also compared with some other classifier, that Naive Bayes, Backpropagation, and k-NN and gives the conclusion that the SVM and Backpropagation produce the best accuracy, namely SVM 90.84% and BP 90.07% for speech and music, but for the speech and the accuracy of the obtained environment is SVM 64.76% and Backpropagation 63.81%. While on research classify music genres using Backpropagation with Fixed Size can add to the Momentum of speed training and improve the accuracy of [2]. Fixed Size Momentum is basically designed to overcome some of the limitations associated with Backpropagation. Fixed Size of Momentum using the update from the previous weights to determine how big an update can be made.

In this research the methods that will be used for classifying is a method of Backpropagation Neural Network classification. With the added method Fixed Size can add Momentum to speed training and improve accuracy. There are some feature to get the hallmark of audio, among others, ZCR (zero crossing rate), MFCC (mel frequency cepstrum's), RMS (root mean square). And in this study for extracting voice signal using Mel Frequency Cepstrum Coeficcients (MFCC), because it has several advantages which were able to capture important information in a voice signal, resulting in minimal data is possible without removing the existing information and mereplikasikan of human hearing organs in a craft made perception against the sound signal [3]. and research [2] using ANN backpropagation with fixed size momentum to classify music genres also use MFCC as featured extraction.

It is expected to add a Fixed Size Momentum on Backpropagation and using the Mel Frequency Cepstrum's (MFCC) on a feature extraction can increase accuracy. The background of the authors are interested in taking on the subject of thesis entitled

"classification of Mixed Type Audio using Backpropagation with Fixed Size Momentum".

2. Content Of Research

2.1 *Backpropagation*

Backpropagation algorithm was one of the supervised learning used in artificial neural networks. Backpropagation [4] looking for a combination of weights to minimize the error output to the correct solution is considered to be. Backpropagation is simply carried out in two stages, namely:

1) Feed-forward

The pattern will be trained on each unit is set to input layer, then the resulting output is transmitted to the next layer until the output layer.

2) Backpropagation

Based on the expected output, each weighting adjusted to produce minimum error ranging from a weighting which is connected to the output neurons, and then continue to retreat up to the input layer.

One of the activation function popular for backpropagation network are binary values range with sigmoid (0.1) which is defined as (2.1). to the stage of backpropagation using functions of its derivatives (2.2).

$$f(x) = \frac{1}{1+e^{-x}} \quad (2.1)$$

$$f'(x) = f(x)(1-f(x)) \quad (2.2)$$

Alternatives in addition to the binary function is a sigmoid sigmoid bipolar with a range of values (-1,1) which is defined as (2.3). to the stage of backpropagation using functions of its derivatives (2.4).

$$f(x) = \frac{2}{1+e^{-x}} - 1 \quad (2.3)$$

$$f'(x) = \frac{1}{2}(1 + f(x))(1-f(x)) \quad (2.4)$$

The following stages in the process of the training algorithm backpropagation.

Step 1: Initialize the weights and parameters

Step 2: calculate the value of the activation on hidden layer

$$Y_net_j = \sum_{j=1}^{p} Xi\, W1_{i,j} \quad (2.5)$$

Where:

Y_net j: number of total inputs to node to-j

P: number of nodes in the hidden layer, earlier for hidden layer node is the sum of the first n input

XI: value node to-I secreted by the hidden layer to hidden layer, earlier the first Xi is the input received by the system

W1i, a: weighting connecting node to-i and j-node to

Calculate the output of each node in the hidden layer with the activation function

Yj = f (Y_net j) (2.6)

Where:

Yj: output from node to-j

$$Y_j = \frac{1}{1+e^{-Y_net_j}} \quad (2.7)$$

Step 3: calculate the input of each node in the output layer.

$$Z_net_k = \sum_{k=1}^{m} Y_j\, W2_{j,k} \quad (2.8)$$

Where:

Z_net k: the total number of input to the node

m: the number of nodes in the hidden layer to output layer before

W2j, k.: weighting connecting node to the node and to-j-k

Calculate the output on each node in the output layer

Zk = f(Z_net $_k$) (2.9)

$$Z_k = \frac{1}{1+e^{-Z_net_k}} \quad (2.10)$$

Step 4: calculate the error of each node in the output layer.

$$\delta_k = (t_k - Z_k)\, f'(Z_net_k) \quad (2.11)$$

where:

k: the number of nodes in the output layer

t: the expected output

Step 5: calculate the weighting change calculation ΔW2 (2.18)

Step 6: calculate the error of each node in the hidden layer

$$\delta_net_j = \sum_{k=1}^{m} \delta_k w2_{j,k} \quad (2.12)$$

where:

m: the number of nodes in the previous layer count error.

$$\delta_j = \delta_net_j f'(Y_net_j) \quad (2.13)$$

Ket :

f'(Y_net $_j$) : Y_net j * (t – Y_net j) (2.14)

t : *output* yang diharapkan

Step 7: calculate the calculation of weighting changes ΔW1 (2.17)

Step 8: calculate the update weights in the hidden layer and output layer

$$W2_{j,k} = W2_{j,k}(lama) + \Delta W2_{j,k} \quad (2.15)$$

$$W1_{i,j} = W1_{i,j}(lama) + \Delta W1_{i,j} \quad (2.16)$$

2.2 *Fixed Size Momentum*

Fixed size momentum is basically designed to overcome some of the limitations associated with the backpropagation training. In order to accelerate the training, many researchers add weight updates based on the update weights before. This effectively increases the time of learning [2]. Many algorithm uses information from the previous weights update to determine how large an update that can be made without diverging.

Fixed size momentum can increase speeds exceeding standard momentum. Fixed size designed to overcome the momentum of a long history of updates of the existing weighting for each existing relations in neural networks. By using this additional information stored, fixed size momentum can

increase the speed of learning and also improve the accuracy of [2].

Standard rules update weights used in backpropagation is:

$\Delta W_{i,j} = \alpha \delta_j X_{i,j}$

Where:

I = index of the source node

j = index targets

α = learning rate

δj = backpropagated error term

x = the value of the input weights

How to update rules this weighting is very slow and takes a long time.

Fixed size momentum using fixed size window to get more information than using standard momentum. By using more memory can allow to overcome some of the limitations issue.

Fixed size momentum can remember update weights which often appear and use that information for other weights. By the standards of the momentum, the fault of the previous partial updates will be applied to the next to worst case some samples will often have the opposite update. Situasu this can disrupt the momentum that has been built and can even be made longer in the train. fixed size momentum can see history.

Fixed size formula for momentum is as follows.

$\Delta W1_{i,j}(t) = \alpha \delta_j X_i + f(\alpha \delta_j X_i, \Delta W1_{ij}(t - 1), \Delta W1_{i,j}(t-2), \dots, \Delta W1_{i,j}(t-k))$ (2.17)

$\Delta W2_{j,k}(t) = \alpha \delta_k Y_j + f(\alpha \delta_j Y_j, \Delta W2_{j.k}(t - 1), \Delta W2_{j.k}(t-2), \dots, \Delta W2_{j,k}(t-k))$ (2.18)

And for the function f is as follows.

$$f(z, \delta_1, \dots, \delta_k) = \begin{cases} -z & if,\ z * \sum \delta_i < 0 \\ 0 & if,\ z * \sum \delta_i \geq 0 \end{cases}$$

(2.19)

There is the argument of k + 1 for the function f, the first is an update at this time and the rest is an update of the previous k where k is the window size algorithm for fixed size momentum. δ 1

The formula above can help in faster convergence and also improve the accuracy of classification and increase the size of the history window which is used to train the network

2.3 jAudio *Feature Extractor*

jAudio is a digital signal processing that are constructed to provide an easy to use program to perform audio extraction feature in [5].

The audio feature in the extract like MFCC, LPC, Beat Histogram and others. jAudio Feature Extractor can be used in a graphical or command line. The system is made by way of executing jAudio is integrated with the command line, then the results are stored in an xml file and is taken by the system as input.

2.4 Discussion

The system includes two stages, namely the process of training and testing process. The training process is done to get the desired weight is met. If the desired conditions already fulfilled then the next can proceed to the testing process. In the process of testing, will be entered test data to classify the results will be compared from the previous training weights. An overview of the system can be seen in Figure 1 below.

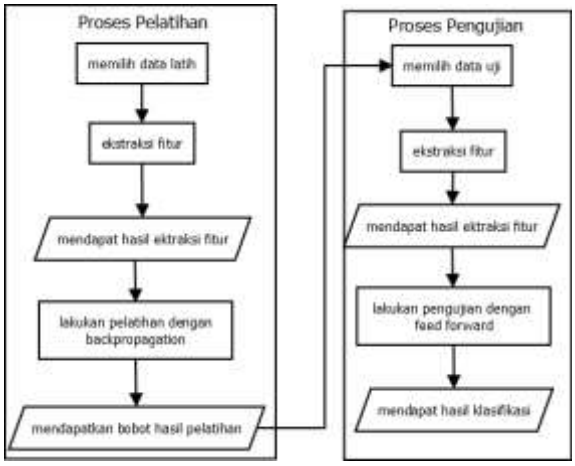


Figure 1 Gambaran Umum Sistem

In the image above can be seen in the first stage, namely conducting training with backpropagation after training is completed then the result of weight training is used to test the process in order to get the results of the classification.

Enter the data used is format wav, then performed the extraction using jAudio MFCC method and choose as a feature extraction. After extracted then the result will then come out in the form of array and stored in xml format which will then be called by the system when necessary. Enter the data used has the following format.

1. Data is formatted audio (.wav)
2. audio Length between 2-5 seconds.
3. data size between 200 – 600 kb.

Following are the results of the extraction for example one data using jAudio and MFCC feature extraction as speechxenvironment1.1.wav.

[-131.0, 15.97, 7.758, 3.273, 0.1709, 1.04, 0.05388, -0.477, 0.091, 0.8632, 1.175, 0.2048, 0.01695].

Table 1 Index dari setiap kelas

Index	Kelas	Output
1	Speech dan environment	0
2	Speech dan music	1

The above class division based on the research of the classification of the mixed type audio with support vector machine, this research will test the accuracy of each of the above classes.

Method of backpropagation architecture consists of three layers in the process analytical study, i.e. the input layer, hidden layer, and the output layer.

1) Input layer

Input layer has a number of neurons which are already affected by the amount of entries on a case,

input layer (Xi) on this case has 13 neurons, because on this research method using MFCC features as his. In the input layer X input signal delivery occurs to the hidden layer.

2) Hidden layer

There is no definite terms in determining the number of neurons in the hidden layer, but the number of hidden layer effect on accuracy and pace of learning of Backpropagation method. Hidden layer is a variable that can be personalised way to know the accuracy of this method, this method is used on analysis of 5 neurons in the hidden layer (use 5 hidden neurons it is early elections, other values will also be tested in the next chapter). Hidden layer on backpropagation serves as to correct and adjust the weights, so a new weight value obtained can be directed to approach the target with the desired output.

3) Output layer

The output layer (Zi) in this case had one neuron, because the number of output in this case in the form of a code of each class. The neurons used to identify learning and also as the output should be.

Network architecture in this case can be seen inFigure 2.

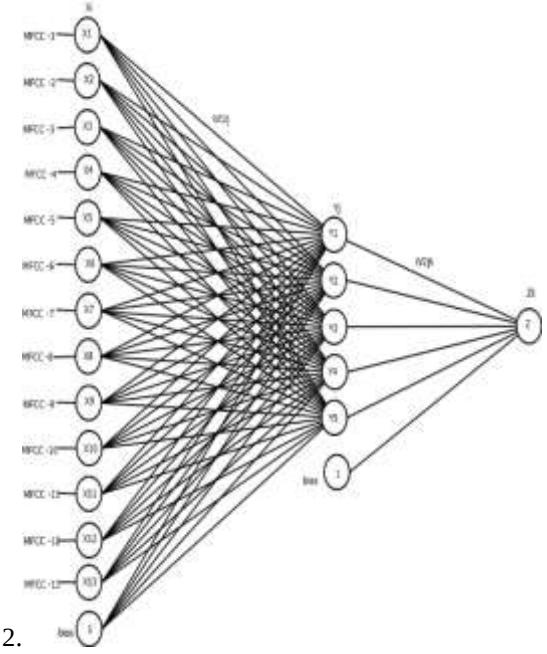


Figure 2 Arsitektur Jaringan Metode Backpropagation *Mixed Type Audio Classification*

Description:
XI: the input consists of 13 layers of neurons
W1i, a: weights on hidden layers
Yj: hidden Layer consists of 5 neurons
W2j, k.: the output layer weights
I: variable that represents the input layer
a: the variable that represents the hidden layer
k: variable that represents an output layer
ZK: output consists of 3 layers of neurons
1: constant bias



Figure 3 *Flowchart* pelatihan BPFSM

Explanation of the flowchart will be implemented on the example of the case below.

Determination Of Variable Training

The first stage in the backpropagation method is implemented on the classification of mixed type variable is the determination of audio training. There are three training variables is specified, i.e., learning rate, windowed size, and a minimum of errors.

The determination of the value of the variable training is affected by the learning rate (the rate of learning) and window size for fixed size momentum. The value of learning rate does not match resulted in a less than optimal classification results. If the value of learning rate is too large, the network will achieve converging in a short time but the error of classification to be great. Conversely if the learning rate is too low, the result of the classification accuracy is good but training process takes a long time, and the target epoch boundary error serves to limit the training, where both this variable will be compared with the condition training. Here are the variables that will be used for training on the analysis method of the example this is the case.

Learning rate

Learning rate used is 0.1, because it takes a big change on every value in order to achieve the expected value. As in the study [2] are already testing some of the learning rate parameter values. The value of learning rate does not match resulted in a less than optimal classification results. Other values on this parameter will also be tested on the next chapter.

Minimum error

Minimum error used is 0.01, because the smaller the error then the better accuracy will come by [2], the use of a minimum error of 0.01 is early elections, other values will also be tested on the next chapter.

Limits of the epoch

Limits of the epoch used is 1000000, the electoral boundary of the epoch with the requirements for the system.

Window size

Window size is 3, use 3 because of previous research using the 3 windows size [2].

Data retrieval training

The training data used in this method a pair signal input (X) with the target output (t) that form a pattern. Signal input (X) value is the result of parsing as demonstrated in the analysis of data input, this value is obtained from the results of the extraction are either MFCC method feature. The input signal will be used in the process of propagation of advanced and improved weighting. Target Output (t) will be used to calculate the error and also used for the propagation of the retreat.

The input data used is the result of a feature extraction method using the MFCC have 13 neurons, there is still a difference of numbers is still very much between one feature with other features. Thus the results of subsequent extraction done normalization on each feature, the normalization is used because it can make the results of the features to be not too far off the difference.

Here's how normalization:

From each feature taken greatest results (max) and terkecilnya (min), use the following normalization formula.

X'=(X - 〖Xmink〗) / (〖Xmaxk〗 - 〖Xmink〗) [6]

X ' = the result of the normalization

X = value to be normalized

Xmink = the smallest value from each of the features

Xmaxk = the greatest value from each of the features

K = index features (1,2, ..., n)

Do the normalization on each individual input data.

Here are the steps for backpropagation training with fixed size momentum.

Step 1: the selected data is going through the process of extraction of features using jAudio beforehand.

Step 2: after the initial weights, weights and biases do Initialization for all weights and biases that connects between the input layer to hidden layer and hidden layer to output layer. The weighting that connects between the input layer to hidden layer is represented by j. W1i Weights that connects between the hidden layer to output layer is represented by W2j, k. Initialization is done by giving the smallest random value between-1 to 1 to all weights and biases.

Step 3: once the initialization is done the weighting then the next is menginputkan's been extracted data to the input layer.

Step 4: after that will do the process feed forward, where in the process it will activate a hidden value in the output layer and the layer by using the following equation.

Activation of hidden layer using the equation (2.5) and (2.6).

Activation of the output layer using the equation (2.8) and (2.10).

Step 5: once the process has finished and the forward feed gets an activation on hidden layer and output layer, then it will start the process of finding the value i.e. backpropagation errors that used to calculate the change in weights using the equation (2.11) to equation (2.18).

Step 6: perform the process of changing weights by using fixed size calculation is made on the momentum equation (2.17) to change the weights occurred in W1 and equations (2.18) to W2.

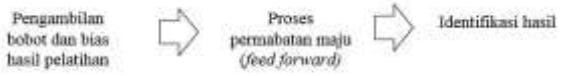
Step 7: after having obtained a new weighting changes then the next new weights will calculate where in the process it will use equation (2.15) and (2.16).

Step 8: then be checked whether there is still data that has not been trained, if still no then the process will be repeated back to step 3.

Step 9: If all data already trained then checked the terms of stop training, training will be terminated if one of the terms are met i.e. the first if the second epoch, maxEpoch > error < minError.

Step 10: If one of the conditions are met then the training is finished, but if you haven't fulfilled it will be kemablai in step 2.

Furthermore if the training has been completed and the results obtained by weighting the end of the training, then the next will proceed with testing. Weights used on training i.e. weighted end of the training. Here is a test to classify mixed flow type audio.



Gambar 4 alur proses pengujian

Step 1: take the initial weights of the weight of the final result of the training that was previously done.

Step 2: input data to be tested, then the data will be done using feature extraction jAudio just like training.

Step 3: results obtained after extraction will be done the process feed forward which results from the activation of the output layer is used as a result to get the classification. The activation process using the formula equation in step 4 the training process i.e. (2.5) to calculate the activation of hidden layer and the equation (2.8) to get the activation of the output layer.

2.5 Resullt

This testing is done for classifying 2 classes, namely class music environment speech and environment, with a total of 60 to 40 data audio data as data training and 20 data as test data.

2.5.1 *Hidden Layer Testing*

The number of neurons in the hidden layer was also influential in the success of learning. Testing this parameter using 0.2 on the learning rate. As for the test results to the number of neurons in the hidden layer can be seen in table 2.

Table 2 Hasil pengujian *Hidden Layer*

Data Latih	40 data		
Data Uji	20 data (10 data per kelas)		
Pengujian ke -	***Hidden layer***	**Akurasi**	**Jumlah iterasi**
1.a	3	75%	1000000
1.b	4	90%	1000000
1.c	5	95%	1000000

2.5.1 *Learning Rate Testing*

Learning rate used also affect the success of the training. Testing these parameters using the 5 hidden layer, and on at least 0.01 error. As for the learning rate test results can be seen in table 3.

Table 3 Hasil Pengujian *Learning Rate*

Data Latih	40 data		
Data Uji	20 data (10 data per kelas)		
Pengujian ke -	***Learning rate***	**Akurasi**	**Jumlah iterasi**
2.a	0.1	70%	168440
2.b	0.2	80%	1000000
2.c	0.3	90%	102800
2.d	0.4	80%	56640
2.e	0.5	85%	58320

2.5.2 *Windows Size Testing*

Windows size used also affect the results of training, testing the parameters it uses 0.3 learning rate, and 5 hidden layer. As for the results of the testing windows size can be seen in table 4.

Table 4 Hasil Pengujian *Windows Size*

Data Latih	40 data		
Data Uji	20 data (10 data per kelas)		
Pengujian ke -	***Windows size***	**Akurasi**	**Jumlah iterasi**
3.a	2	90%	1000000
3.b	3	95%	283120
3.c	4	75%	1000000
3.d	8	70%	1000000
3.e	10	50%	1000000

2.5.3 Comparison Testing Of BP and BPFSM

Based on the results from testing previously obtained values of parameters that are already considered good enough to be used as a parameter for both methods i.e. using 5 hidden layer learning rate, 0.3 and 3 windows size. As for the results of the testing of backpropagation with backpropagation with FSM can be seen in table 5.

Table 5 Hasil Pengujian Perbandingan BP dan BPFSM

Data Latih	40 data				
Data Uji	20 data (10 data per kelas)				
Pengujian	***Hidden layer***	***Learning rate***	***Windows size***	**akurasi**	**Jumlah iterasi**
BP	5	0.3	0	80%	22440
BPFSM	5	0.3	3	95%	146520

Based on the algorithm parameter testing has been done, can be summed up some of the things that is.

1) based on testing looking for the number of neurons in the hidden layer that yields the highest accuracy is 5 with accuracy of 80%.
2) based on testing the Learning Rate that generates the highest accuracy is 0.3 with accuracy by 90%.
3) based on testing windows size that produces the highest accuracy is 3 with 95% accuracy.
4) based on testing of backpropagation and backpropagation with fixed size momentum get higher accuracy is the backpropagation with fixed size momentum generates 95% accuracy.

4.CONCLUSION

Based on the results obtained in this research, then it can be concluded that the classification of the mixed type audio using backpropagation with fixed size momentum can increase the accuracy of 95% when compared with backpropagation only get 80% of accuracy, but requires more iterations. More

windows size used on a fixed size diminishing accuracy makes the momentum, and make more iterations.

As for suggestions that are proposed to be able to develop a similar study among other things:

1. Can apply the classification of mixed type audio using Backpropagation with fixed size momentum to classify the class of mixed type on pieces of film clips.

2. Can test back on windwos size on the fixed size momentum in order to find a solution to problems that occur at iteration.

DAFTAR PUSTAKA

[1] L. Chen, S. Gunduz, M. T. Ozsu, "Mixed Type Audio Classification With Support Vector Machine," IEEE International Confrence on Multimedia and Expo, 2006.

[2] N. Prabhu, A. Anodkar, and R. Kenkre, "Music Genre Classification Using Improved Artificial Neural Network With Fixed Sized Momentum," International Journal of Comupter Application, vol. 101, no. 14, 2014.

[3] Y. Somantri, "Pengenalan Pembicara Dengan Ekstrasi Ciri MFCC," 2006.

[4] R. Rojas, "Neural Networks A Systematic Introduction," Springer-Verlag, Berlin, 1996.

[5] D. McEnnis, C. McKay, "jAudio Official Website," [Online]. Available: http://jaudio.sourceforge.net. [Accessed 4 October 2016].

[6] P. Davis and S. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.