Pengenalan Ucapan Bahasa Indonesia Menggunakan MFCC dan Recurrent Neural Network

Panggih Tridarma[*1)], Sukmawati Nur Endah[*2)]

[*]Informatics Department, Universitas Diponegoro
[1)]tridarmapanggih@gmail.com, [2)]sukmawati020578@gmail.com

Abstrak

Pengenalan ucapan (speech recognition) merupakan perkembangan teknologi dalam bidang suara. Pengenalan ucapan memungkinkan suatu perangkat lunak mengenali kata-kata yang diucapkan oleh manusia dan ditampilkan dalam bentuk tulisan. Namun masih terdapat masalah untuk mengenali kata-kata yang diucapkan, seperti karakteristik suara yang berbeda, usia, kesehatan, dan jenis kelamin. Penelitian ini membahas pengenalan ucapan bahasa Indonesia dengan menggunakan Mel-Frequency Cepstral Coefficient (MFCC) sebagai metode ekstraksi ciri dan Recurrent Neural Network (RNN) sebagai metode pengenalannya dengan membandingkan arsitektur Elman RNN dan arsitektur Jordan RNN. Pembagian data latih dan data uji dilakukan dengan menggunakan metode k-fold cross validation dengan nilai k=5. Hasil penelitian menunjukkan bahwa arsitektur Elman RNN pada parameter 900 hidden neuron, target error 0.0005, learning rate 0.01, dan maksimal epoch 10000 dengan koefisien MFCC 20 menghasilkan akurasi terbaik sebesar 72.65%. Sedangkan hasil penelitian untuk arsitektur Jordan RNN pada parameter 500 hidden neuron, target error 0.0005, learning rate 0.01, dan maksimal epoch 10000 dengan koefisien MFCC 12 menghasilkan akurasi terbaik sebesar 73.55%. Sehingga berdasarkan hasil penelitian yang didapat, arsitektur Jordan RNN memiliki kinerja yang lebih baik dibandingkan dengan arsitektur Elman RNN dalam mengenali ucapan Bahasa Indonesia berjenis continuous speech.

Keywords : *Pengenalan ucapan, Mel-Frequency Cepstral Coefficient, Recurrent Neural Network, Elman RNN, Jordan RNN*

Abstract

Speech recognition is a technological development in the field of speech. Speech recognition supports a software that contains words spoken by humans and in written languages. However, there are still problems in assuming the words spoken, such as different voices, age, health, and gender. This study discusses the recognition of Indonesian speech using the Mel-Frequency Cepstral Coefficient (MFCC) as a feature extraction method and the Recurrent Neural Network (RNN) as a method of recognition by comparing the Elman RNN architecture and the Jordan RNN architecture. The distribution of training data and data testing was carried out using the k-fold cross validation method with a value of k = 5. The results showed that the Elman RNN architecture on parameters of 900 hidden neurons, target error 0.0005, learning rate 0.01, and a maximum epoch of 10000 with the MFCC coefficient of 20, the best value is 72.65%. While the results of research for Jordan RNN architecture on parameters of 500 hidden neurons, target error 0.0005, learning rate 0.01, and a maximum epoch of 10000 with MFCC coefficient 12 yield the best value of 73.55%. So that based on the research results obtained, the Jordan RNN architecture has a better performance than the Elman RNN architecture in the Indonesian language statement of the continuous speech type.

Keywords : *Speech recognition, Mel-Frequency Cepstral Coefficient, Recurrent Neural Network, Elman RNN, Jordan RNN*

1 PENDAHULUAN

Ucapan merupakan bentuk komunikasi manusia yang paling efisien untuk berkomunikasi dengan manusia lainnya. Bentuk komunikasi ini merupakan salah satu cara yang efektif untuk menyampaikan maksud dan tujuan seseorang dalam menyampaikan sebuah informasi. Seiring dengan perkembangan teknologi, komunikasi yang dilakukan oleh manusia tidak hanya terbatas pada komunikasi antar manusia tetapi juga sudah berkembang komunikasi antara manusia dengan komputer.

Pengenalan ucapan (*speech recognition*) merupakan salah satu contoh dari perkembangan teknologi dalam bidang suara. Pengenalan ucapan memungkinkan suatu perangkat lunak untuk mengenali dan memahami kata-kata yang diucapkan oleh manusia. Kata-kata yang diucapkan diubah bentuknya menjadi sinyal digital dengan cara mengubah gelombang ucapan menjadi sekumpulan angka yang kemudian disesuaikan dengan kode-kode tertentu untuk mengidentifikasi kata-kata tersebut. Hasil dari identifikasi kata yang diucapkan dapat ditampilkan dalam bentuk tulisan.

Penelitian pengenalan ucapan sebelumnya [2,3] masih sebatas membahas mengenai pengenalan ucapan jenis *isolated speech*, dimana jenis ini hanya mengenali satu kata setiap di *running*. Untuk itu pengenalan ucapan Bahasa Indonesia perlu dikembangkan lebih lanjut, sehingga dapat digunakan untuk mengenali dengan banyak kata atau kalimat.

Pengenalan ucapan jenis *continuous speech* terdapat beberapa tahapan yaitu *pre-processing*, ekstraksi ciri, dan pengenalan. *Pre-processing* merupakan tahapan awal dimana suara akan diproses sedemikian rupa sehingga siap untuk diekstraksi ciri. *Pre-processing* sendiri digunakan untuk meningkatkan kualitas sinyal ucapan [4]. Tahap sesudah dilakukan *pre-processing* adalah proses ekstraksi ciri. Proses ini dilakukan untuk mendapatkan ciri khas dari ucapan tersebut dengan menggunakan metode MFCC yang mengkonversi sinyal ucapan ke dalam beberapa vektor data yang berguna bagi proses pengenalan ucapan.

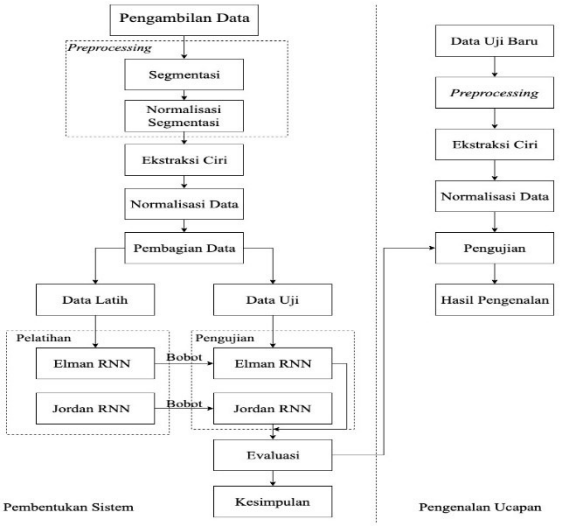
Hasil ekstraksi yang sudah didapatkan, kemudian dilakukan tahapan pengenalan. Metode yang dapat digunakan adalah *Recurrent neural network* (RNN) yang dapat bekerja dengan baik dalam pengenalan ucapan dikarenakan kemampuannya untuk mengenali suatu informasi dengan banyaknya varian waktu seperti sinyal ucapan [5]. RNN adalah jaringan yang mempunyai minimal satu *feedback loop* yang mana suatu *recurrent network* bisa terdiri atas satu lapisan *neuron* tunggal dengan masing-masing *neuron* memberikan kembali *output*-nya sebagai *input* pada semua neuron yang lain [6]. RNN mampu menyimpan memori/ingatan (*feedback loop*) yang memungkinkan untuk mengenali pola data dengan baik. Arsitektur RNN yang sering digunakan adalah arsitektur Elman RNN dan arsitektur Jordan RNN. Penelitian sebelumnya oleh Widya et al, (2018) kedua arsitektur tersebut dapat memprediksi kualitas udara dengan cukup baik [7]. Sehingga penelitian ini akan membandingkan arsitektur Elman RNN dan arsitektur Jordan RNN untuk digunakan pada metode pengenalan ucapan Bahasa Indonesia.

Berbagai penelitan sebelumnya yang menggunakan RNN sebagai metode pengenalan, di antaranya penelitian oleh Alotaibi (2004) yang menghasilkan akurasi 99.47% untuk mengenali ucapan angka Bahasa Arab [8]. Penelitian Kurniadi (2018)

menerapkan metode RNN untuk mengenali ucapan kata Bahasa Indonesia yang mampu menghasilkan akurasi 83.88% [9]. Penelitian Fachrie et al, (2015) menggunakan metode RNN untuk mengenali ucapan angka Bahasa Indonesia yang menghasilkan akurasi 99.30% [10]. Dari beberapa penelitian yang dikemukakan sebelumnya, metode RNN mampu memberikan hasil yang baik untuk metode pengenalan ucapan, namun kinerja dari RNN masih belum diketahui untuk kasus pengenalan ucapan secara *continuous* dalam Bahasa Indonesia.

2 METODE PENELITIAN

Penelitian Pengenalan Ucapan Bahasa Indonesia Menggunakan *Mel-Frequency Cepstral Coefficient* dan *Recurrent Neural Network* ini bertujuan untuk menentukan akurasi terbaik dengan membandingkan dua arsitektur RNN (arsitektur Elman RNN dan arsitektur Jordan RNN) pada proses pengenalan ucapan, serta menentukan parameter terbaik dari kedua arsitektur tersebut. Tahapan-tahapan penyelesaian masalah ditunjukkan pada Gambar 1.



Gambar 1 Blok Proses Tahapan-tahapan Penyelesaian Masalah

Penelitian ini memiliki beberapa tahapan proses diantaranya pengumpulan data, *preprocessing*, *partition data*, pelatihan, pengujian, dan evaluasi. Penjelasan mengenai tiap proses sebagai berikut:

a. Pengumpulan data

Pengumpulan data untuk pengenalan ucapan dilakukan dengan menggunakan data perekaman suara yang sudah direkam untuk data pelatihan, data uji, dan data suara langsung. Data perekaman suara untuk data pelatihan dan data uji direkam dengan suara lima orang yang berbeda dalam bentuk kalimat, terdiri dari lima kata, dengan frekuensi 44100 Hz, *channel mono* 16 bit, dan disimpan dalam *file* berformat *.wav. Sedangkan untuk data suara langsung digunakan untuk pengenalan ucapan secara rekaman langsung.

b. *Pre-processing*

Tahap *pre-processing* yang dilakukan adalah proses segmentasi dan normalisasi. Proses segmentasi dilakukan pemecahan sinyal ucapan *continuous speech* menjadi beberapa kata. Sedangkan proses normalisasi dilakukan pada sinyal ucapan agar didapat keseragaman data

c. Ekstraksi Ciri

Data sinyal ucapan akan mengalami serangkaian proses yang bertujuan untuk mendapatkan ciri yang merepresentasikan sinyal ucapan tersebut. Metode yang digunakan pada tahap ekstraksi ciri menggunakan metode *Mel-Frequency Cepstral Coefficient* (MFCC). Ada tujuh proses pada metode MFCC, yaitu *pre-emphasize, frame blocking, windowing, fast fourier transform, mel-frequency wrapping, discrete cosine transform,* dan *cepstral liftering* [11].

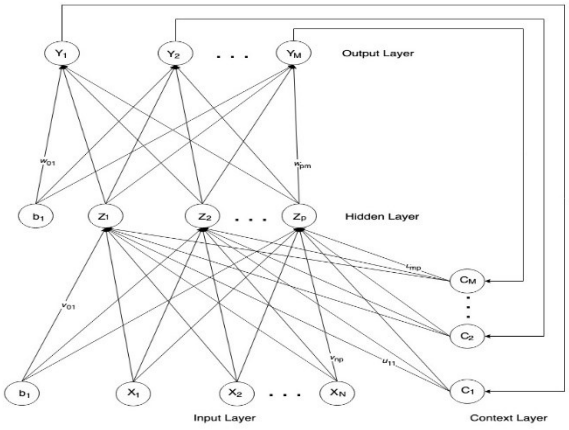
d. Pembagian Data

Data penelitian dibagi menjadi data latih dan data uji. Data latih digunakan dalam

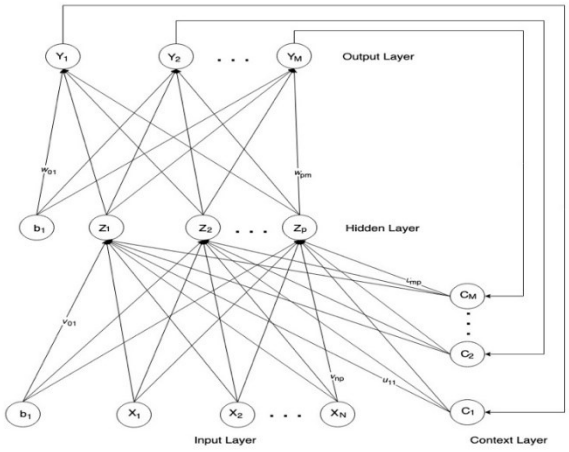
proses pelatihan, sedangkan data uji digunakan untuk validasi model dari proses pelatihan yang berupa pengujian keakuratan dari arsitektur yang dilatih. Data yang digunakan berjumlah 2000 data suara yang berasal dari 5 orang berbeda, masing-masing mengucapkan 50 kata sebanyak 8 kali. Pembagian data dilakukan dengan menggunakan metode *k-fold cross validation* dengan nilai *fold* = 5, sehingga seluruh dataset suara dikelompokkan menjadi 5 kelompok. Setiap kelompok terdiri dari 400 data sinyal ucapan, dimana setiap kelompok terdiri dari 50 kelas yang masing-masing terdapat 8 data sinyal ucapan.

e. Arsitektur Recurrent Neural Network

Arsitektur yang digunakan pada penelitian ini terdiri dari dua arsitektur, yaitu arsitektur Elman RNN dan arsitektur Jordan RNN. Kedua arsitektur tersebut dilakukan tahapan-tahapan seperti tahap *feedforward* (propagasi maju), tahap *backward* (propagasi mundur), tahap *error backpropagation* terhadap setiap bobot mulai dari bobot *input-hidden* (*w*), *hidden-output* (*v*), dan *context-hidden/input* (*u*) pada setiap *timestep*, dan tahap modifikasi bobot dan bias [12]. Arsitektur umum Elman RNN dan arsitektur Jordan RNN dapat dilihat pada Gambar 2 dan Gambar 3 dengan banyak lapisan sebagai berikut :



Gambar 2 Arsitektur Elman Recurrent Neural Network



Gambar 3 Arsitektur Jordan Recurrent Neural Network

1) Input Layer

Pada *layer* ini terdapat satu lapisan yang terdiri dari *N neuron* dan satu *neuron* bias. Nilai *N* merupakan jumlah *input neuron* dengan banyak sesuai hasil dari sinyal proses MFCC. Jika nilai koefisien MFCC adalah 12, 16, atau 20 maka nilai *N* untuk jumlah *input neuron* berturut-turut adalah 1700, 2200, atau 2700.

2) Hidden Layer

Hidden layer berjumlah satu lapisan, dimana jumlah *hidden neuron* berjumlah *P* dan satu satu *neuron* bias. Nilai *P* merupakan nilai yang akan dijadikan parameter uji dalam penelitian ini. Untuk menentukan banyaknya jumlah *hidden neuron* pada penelitian ini digunakan nilai 1/3 dan ¼ dari jumlah *input neuron*. Jika *input neuron* adalah 1700 maka nilai *P* adalah 400 dan 500 (hasil pembulatan). Jika *input neuron* adalah 2200 maka nilai *P* adalah 500 dan 700. Jika *input neuron* adalah 2700 maka nilai *n* adalah 700 dan 900

3) Output Layer

Output layer berjumlah satu lapisan yang memiliki 6 *output neuron* yaitu nilai biner dari 1 sampai 50. Jumlah *output neuron* tersebut dapat mewakili 50 kata yang akan dikenali oleh RNN.

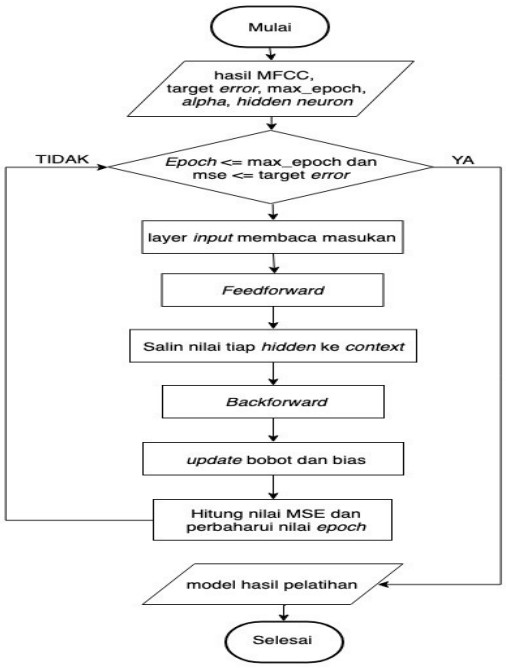
4) Context Layer

Context layer berjumlah satu lapisan dimana jumlah *context neuron* pada arsitektur Elman RNN berjumlah *P* dengan nilai *P* merupakan salinan dari *hidden layer*. Sedangkan pada arsitektur Jordan RNN berjumlah *M* dengan nilai *M* merupakan salinan dari *output layer*.

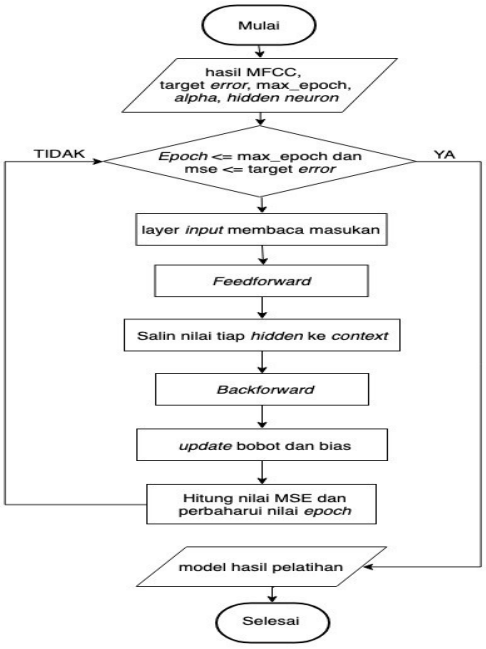
Proses pelatihan RNN memiliki 50 target berbeda yang menggambarkan 50 kata yang akan digunakan saat pengenalan. Jika hasil dari pelatihan RNN digunakan untuk proses pengenalan, maka nilai *threshold* untuk *output layer* adalah 0.5, dimana jika nilai *output* lebih besar dari 0.5 maka nilai *output* tersebut adalah 1. Sedangkan jika nilai *output* lebih kecil dari 0.5 maka nilai *output* tersebut adalah 0.

f. Pelatihan & Pengujian

Tahap pelatihan menggunakan data latih dan dibagi menjadi dua fase yaitu fase propagasi maju (*feedforward*) dan fase propagasi mundur (*backforward*). Sedangkan pada tahap pengujian menggunakan data uji dan dilakukan hanya pada fase propagasi maju dengan menggunakan bobot akhir hasil dari pelatihan. Gambar *flowchart* tahap pelatihan dan pengujian arsitektur RNN dapat dilihat pada Gambar 4 dan Gambar 5.



Gambar 4 *Flowchart* Pelatihan Elman RNN



Gambar 5 Flowchart Pelatihan Jordan RNN

3 HASIL DAN PEMBAHASAN

Penelitian ini memiliki beberapa tahapan proses diantaranya pengumpulan data, *preprocessing*, *partition data*, pelatihan, pengujian, dan evaluasi. Penjelasan mengenai tiap proses sebagai berikut:

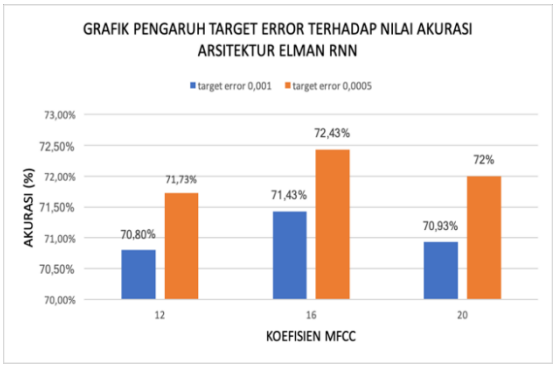
a. Skenario Uji

Pengujian ini digunakan untuk mencari arsitektur RNN yang optimal dari setiap RNN pada masing-masing skenario dengan mengubah koefisien MFCC (12, 16, dan 20) dan beberapa parameter RNN yaitu, 2 *hidden neuron* ($\frac{1}{3}$ dari jumlah *input neuron* dan $\frac{1}{4}$ dari jumlah *input neuron*), 3 *learning rate* (0.01, 0.005, dan 0.001), dan 2 *target error* (0.001 dan0.0005), dan maksimal *epoch* 10000. Penelitian ini menggunakan 3 skenario uji, yaitu :

1) Skenario 1: Skenario 1 bertujuan untuk mengetahui kinerja arsitektur Elman RNN terhadap pengenalan ucapan Bahasa Indonesia dengan menghitung nilai akurasi hasil.
2) Skenario 2: Skenario 2 bertujuan untuk mengetahui kinerja arsitektur Jordan RNN terhadap pengenalan ucapan Bahasa Indonesia dengan menghitung nilai akurasi hasil.
3) Skenario 3: Skenario 3 bertujuan untuk membandingkan kinerja arsitektur Elman RNN dan arsitektur Jordan RNN berdasarkan hasil dari skenario 1 dan skenario 2. Pada skenario ini dilakukan perbandingan parameter (hidden neuron, learning rate, target error) yang digunakan oleh kedua arsitektur untuk menghasilkan akurasi yang baik.

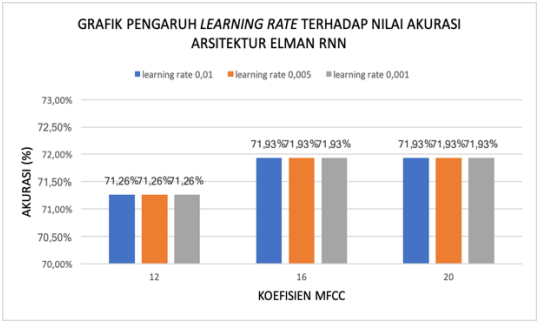
b. Hasil Eksperimen

1) Skenario 1 : Hasil pengujian skenario 1 dapat dijelaskan analisis hasil berdasarkan tiap parameter yang digunakan sebagai berikut :



i. Learning Rate

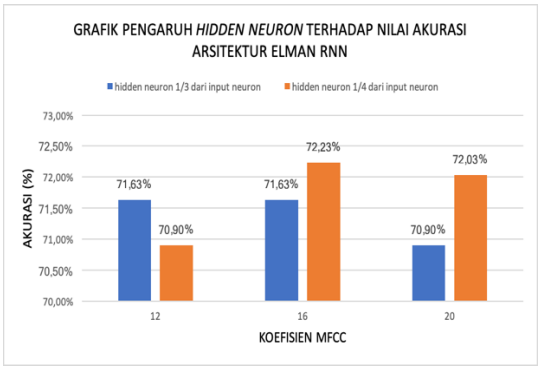
Berdasarkan Gambar 6 *learning rate* tidak berpengaruh pada akurasi pengenalan ucapan Bahasa Indonesia.



Gambar 6 Grafik Pengaruh *Learning Rate* Terhadap Nilai Akurasi Elman RNN

ii. *Hidden Neuron*

Berdasarkan Gambar 7 *hidden neuron* dengan jumlah 1/4 dari *input neuron* menghasilkan akurasi yang lebih baik dibandingkan dengan *hidden neuron* dengan jumlah 1/3 dari *input neuron*, dengan rata-rata akurasi tertinggi yaitu 72.23%.



Gambar 7 Grafik Pengaruh *Hidden Neuron* Terhadap Nilai Akurasi Elman RNN

iii. *Target Error*

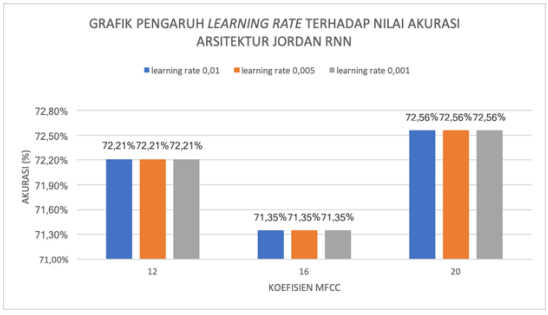
Berdasarkan Gambar 7 target *error* dengan nilai 0.0005 menghasilkan akurasi yang lebih baik dibandingkan dengan target *error* dengan nilai 0.001. Akurasi Elman RNN dengan nilai target *error* yang kecil membuat nilai *output* semakin mendekat dengan nilai target, sehingga arsitektur Elman RNN lebih mudah mengenali suara sesuai dengan target yang ditentukan.

Gambar 7 Grafik Pengaruh *Target Error* Terhadap Nilai Akurasi Elman RNN

2) Skenario 2: Hasil pengujian skenario 2 dapat dijelaskan analisis hasil berdasarkan tiap parameter yang digunakan sebagai berikut :

i. *Learning Rate*

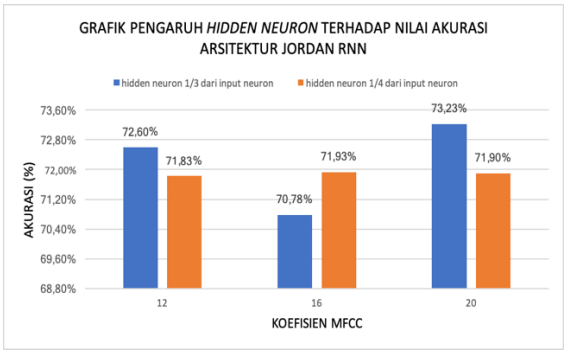
Berdasarkan Gambar 8 *learning rate* tidak berpengaruh pada akurasi pengenalan ucapan Bahasa Indonesia.



Gambar 8 Grafik Pengaruh *Learning Rate* Terhadap Nilai Akurasi Jordan RNN

ii. *Hidden Neuron*

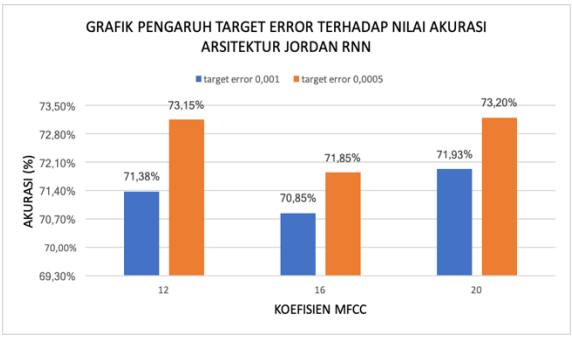
Berdasarkan Gambar 9 *hidden neuron* dengan jumlah 1/3 dari *input neuron* menghasilkan akurasi yang lebih baik dibandingkan dengan *hidden neuron* dengan jumlah ¼ dari *input neuron*, dengan rata-rata akurasi tertinggi yaitu 73.23%.



Gambar 9 Grafik Pengaruh *Hidden Neuron* Terhadap Nilai Akurasi Jordan RNN

iii. *Target Error*

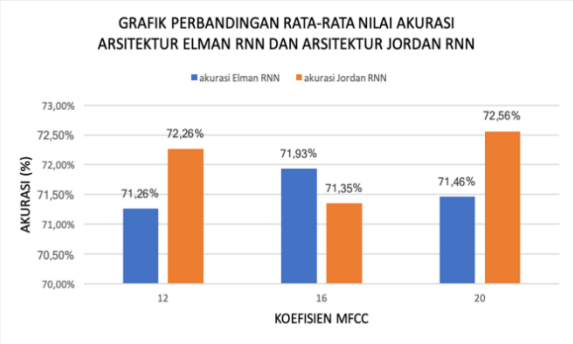
Berdasarkan Gambar 10 target *error* dengan nilai 0.0005 menghasilkan akurasi yang lebih baik dibandingkan dengan target *error* dengan nilai 0.001. Target *error* terbaik dihasilkan ketik koefisien MFCC 20 dengan rata-rata akurasi 73,20%. Akurasi Jordan RNN dengan nilai target *error* yang kecil membuat nilai *output* semakin mendekat dengan nilai target.



Gambar 10 Grafik Pengaruh *Target Error* Terhadap Nilai Akurasi Jordan RNN

3) Skenario 3: Berdasarkan hasil pengujian pada pembahasan skenario 1 dan hasil pengujian pada pembahasan skenario 2, hasil akurasi terbaik diperoleh pada pengujian arsitektur Jordan RNN dengan jumlah koefisien MFCC 12 dan parameter yaitu, jumlah *hidden neuron* 500, target *error* 0.0005, dan maksimal *epoch* 10000, yaitu 73.35%. Hasil perbandingan nilai akurasi Elman

RNN dan Jordan RNN dapat ditunjukkan pada Gambar 11.



Gambar 11 Grafik Perbandingan Rata-rata Nilai Akurasi Arsitektur Elman RNN dan Arsitektur Jordan RNN

c. Analisis Hasil

Hasil akurasi pengujian parameter kedua arsitektur secara umum memiliki nilai yang belum optimal. Nilai rata-rata akurasi tertinggi yang dapat dicapai pada penelitian ini adalah 73.55%. Ada beberapa faktor yang menyebabkan belum optimalnya akurasi dari kedua arsitektur tersebut yaitu :

1) Hasil yang di-*input*-kan dipengaruhi oleh faktor suara manusia, seperti usia, cara pengucapan, kesehatan yang mempengaruhi kualitas suara, perasaan seseorang, dan lain-lain.
2) Kata yang sama diucapkan oleh orang yang berbeda cenderung menghasilkan sinyal ucapan yang berbeda. Hal ini menyebabkan sulitnya RNN dalam mengenali kata yang sama.
3) Jumlah ciri yang dihasilkan ketika proses MFCC tiap kata berbeda, sedangkan pada RNN jumlah *input neuron* harus sama, hal ini menyebabkan perlu dilakukan proses *padding* (penambahan nilai 0) agar jumlah ciri tiap katanya sama. Proses penambahan *padding* yang didapat mempengaruhi hasil akurasi.

4 KESIMPULAN

Berdasarkan hasil nilai akurasi yang didapat pada proses pengujian, dapat disimpulkan bahwa metode *recurrent neural network* dengan arsitektur Elman RNN menghasilkan nilai rata-rata akurasi tertinggi, yaitu 72.65%. Akurasi didapat pada saat koefisien MFCC 20 dan parameter RNN dengan jumlah *hidden neuron* 900 (1/3 *input neuron*), target *error* 0.0005, dan maksimal *epoch* 10000.

Arsitektur Jordan RNN terbukti memiliki kinerja yang lebih baik dibandingkan dengan arsitektur Elman RNN dilihat dari nilai rata-rata akurasi tertinggi yang dihasilkan pada penelitian ini yaitu 73.55%. Akurasi didapat pada saat koefisien MFCC 12 dan parameter RNN dengan jumlah *hidden neuron* 500 (1/3 *input neuron*), target *error* 0.0005, dan maksimal *epoch* 10000.

DAFTAR PUSTAKA

[1] Sukmawati NE, Satrio A, Sutikno. Integrated System Design for Broadcast Program Infringement Detection. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2015; 13(2): 571-577.

[2] Norhaslinda K, Abdul WAR, Khairul IMH, Muhammad HIM. Driver Behaviour State Recognition Based on Speech. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2018; 16(2): 852-861.

[3] Widya MS, Sukmawati NE. Suitable Recurrent Neural Network for Air Quality Prediction With Backpropagation Through Time. *International Conference on Informatics and Computational Sciences (ICICos)*. 2018(2): 1-6.

[4] Muhammad F, Agus H. Robust Indonesian Digit Speech Recognition Using Elman Recurrent Neural Network. *Konferensi Nasional Informatika (KNIF)*. 2015; 15(1): 49-54.
[5] Michael IJ. Parallel Distruted Processing Approach. California: ICS. 1986: 10.
[6] Saliza I, Abdul M.Recurrent Neural Network with Backpropagation Through Time Algorithm for Arabic Recognition. *European Simulation Multiconference*. 18(1): 1-5.
[7] Widya MS, Sukmawati NE. Suitable Recurrent Neural Network for Air Quality Prediction With Backpropagation Through Time. *International Conference on Informatics and Computational Sciences (ICICos)*. 2018(2): 1-6.
[8] Yousef AA. Spoken Arabic Digits Recognizer Using Recurrent Neural Networks. *IEEE*. 1(1): 1-5.
[9] Wildan K, Irfan M. Speech Recognition Menggunakan Elman Recurrent Neural Network Untuk Kata Dalam Bahasa Indonesia. *Universitas Komputer Indonesia*. 2018: 1-10.
[10] Muhammad F, Agus H. Robust Indonesian Digit Speech Recognition Using Elman Recurrent Neural Network. *Konferensi Nasional Informatika (KNIF)*. 2015; 15(1): 49-54.
[11] Sukmawati NE, Satrio A, Sutikno. Comparison of Feature Extraction MFCC and LPC in Automatic Speech Recognition for Indonesian. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2017; 15(1): 292-298.