Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/JTT
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

PENGENALAN POLA FONEM VOKAL MENGGUNAKAN *SHORT TIME FOURIER TRANSFORM* (STFT) DAN FITUR *MEL FREQUENCY CEPSTRAL COEFFICIENT* (MFCC)

Ahmad Rio Adriansyah[1], Kurniawan Dwi Prasetyo[2], Hamdan Ainul Atmam Al Faruqi[3]

[1, 3] Teknik Informatika, Sekolah Tinggi Teknologi Terpadu Nurul Fikri
[2] Sistem Informasi, Sekolah Tinggi Teknologi Terpadu Nurul Fikri
Depok, Jawa Barat, Indonesia 16451
arasy@nurulfikri.ac.id, kurniawan@nurulfikri.ac.id, atmam58@gmail.com

Abstract

Phonemes are the building blocks of every oral language. Every utterance is composed of one or more phonemes. To improve the accuracy of acoustic models, the researchers attempted to identify the pattern of vowel phonemes in bahasa Indonesia using STFT and MFCC features. This paper analyzes 398 audio files gathered from 51 participants and explores the difference of phonemes a, i, u, e,o. Using SVM and Neural Network, the features are classified and tested. The result gave 93.8% accuracy using SVM with radial based kernel.

Keywords: *phonemes classification, vowel, STFT, MFCC*

Abstrak

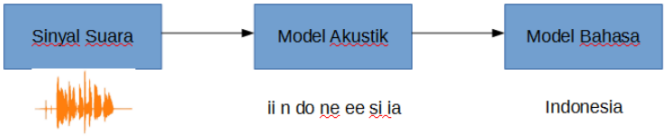
Fonem adalah bagian yang menyusun semua bahasa lisan. Setiap kata dan kalimat yang diutarakan terdiri dari satu fonem atau lebih. Untuk meningkatkan akurasi dari model akustik, peneliti mencoba mengidentifikasi pola fonem vokal dalam bahasa Indonesia menggunakan STFT dan Fitur MFCC. Dalam penelitian ini, peneliti menganalisis data dari 398 file suara yang bersumber dari 51 orang partisipan dan mengeksplorasi perbedaan pola dari fonem vokal a,i,u,e,o. Dengan menggunakan SVM dan JST, fitur tersebut diklasifikasikan dan diuji. Hasil pengujian memberikan akurasi 93,8% menggunakan SVM dengan kernel radial.

Kata kunci: klasifikasi fonem, fonem vokal, STFT, MFCC

1. PENDAHULUAN

Pengenalan suara otomatis (ASR) adalah salah satu teknologi yang memudahkan manusia dalam berinteraksi dengan komputer. Model maupun aplikasi terapannya telah banyak dikembangkan dan diterapkan ke bidang-bidang seperti layanan kesehatan, pemerintahan, dan lain sebagainya [1].

Tahapan yang dilakukan dalam proses pengenalan suara ditunjukkan pada **Error! Reference source not found.**.



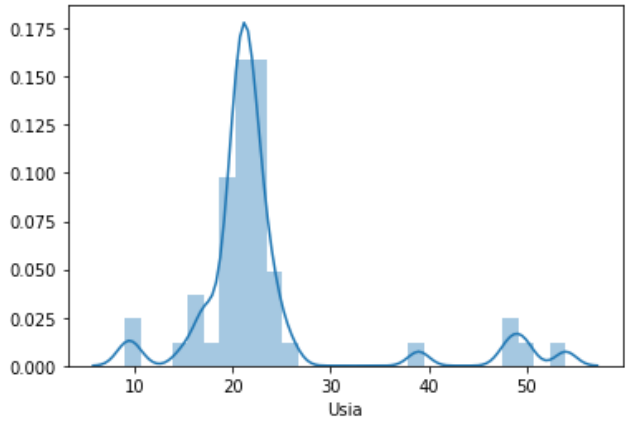
Gambar 1. Alur Pengenalan Suara

Selama ini, sistem pengenalan suara banyak dikembangkan dengan kalimat lengkap. Pada penelitian ini, penulis mencoba melakukan pendekatan berbeda yaitu dengan pola fonem vokal untuk meningkatkan akurasi model akustiknya. Fonem vokal digunakan karena bentuk tersebut adalah bentuk paling dasar dari sebuah utaran. Vokal yang dipilih adalah fonem vokal bahasa Indonesia (a, i, u, e, o) yang disederhanakan. Variasi silabe seperti ‘e’ pada ‘teh’, dan ‘e’ pada ‘teman’ dianggap sama, begitu juga pada vokal yang lainnya.

Penelitian ini dilakukan dengan data vokal murni yang sengaja diambil (bukan dari kalimat lengkap) untuk mengenali fonemnya. Hal ini karena data serupa yang bersifat terbuka tidak tersedia, maka peneliti mengumpulkan data tersebut dari masyarakat penutur bahasa Indonesia. Data suara dikumpulkan dari 51 orang partisipan dengan rentang usia dari 9 hingga 54 tahun. Rata-

Diterima 07 Desember 2020, Direvisi 08 Februari 2021, Diterima untuk publikasi 12 Juli 2021

rata usia partisipan adalah 23 tahun dengan simpangan baku 9 tahun.



Gambar 2. Distribusi Usia Partisipan

Peneliti mencoba mengeksplorasi pola fonem vokal menggunakan *Short Time Fourier Transform* (STFT) dan fitur *Mel Frequency Cepstral Coefficient* (MFCC). Dari data tersebut, dibuat beberapa buah model klasifikasi sederhana dan dibandingkan hasilnya. Klasifikasi dilakukan dengan SVM dan JST lalu dibandingkan akurasinya.

2. TINJAUAN PUSTAKA

Pengenalan suara adalah bidang riset yang sudah berkembang dari tahun 1950an. Berawal dari pengenalan angka pada teknologi yang dikembangkan oleh Bell Laboratories (Audrey), pengenalan kata yang berdasarkan fonem oleh IBM's Shoebox, hingga sistem yang dikembangkan oleh Google, Apple (Siri), Amazon (Alexa), dan Microsoft (Cortana) di dekade kedua abad 21 [2]. Model pengenalan akustik yang diterapkan pun beragam, dari HMM yang popular sejak diterapkan oleh IBM di tahun 1980an, RNN, LSTM, hingga model *end-to-end* seperti CTC [3].

Pada bagian ini akan dijelaskan mengenai klasifikasi fonem bahasa Indonesia, serta *Short Time Fourier Transform* (STFT) dan *Mel Frequency Cepstral Coefficient* (MFCC) sebagai dasar perancangan sistem pengenalan yang dibuat.

2.1 Fonem Bahasa Indonesia

Fonem adalah sebuah bunyi fungsional. Misalnya pada bahasa Inggris, 't' dalam 'stop' dan 'th' dalam *top* merupakan bunyi yang sama secara fungsional. Jadi 't' dan 'th' merupakan dua bentuk yang berbeda dari fonem yang sama [4].

Fonem identitasnya hanya berlaku dalam sebuah bahasa yang sama. Sebuah fonem yang berbeda dalam suatu bahasa bisa jadi sama dalam bahasa lainnya. Contohnya fonem 'r' dan 'l' pada bahasa Indonesia berbeda, karena itu fungsionalitasnya dalam kata 'palu' dan 'paru' juga berbeda. Lain halnya dengan bunyi 'r' dan 'l' bahasa Jepang, kedua tersebut bukan merupakan fonem yang berbeda sehingga kata 'biru' dan 'bilu' memiliki arti yang sama.

Jumlah bunyi fonem vokal dalam bahasa Indonesia bervariasi menurut pendapat para ahli. Marsono [5] mengategorikan fonem vokal menjadi 9, Dardjowidjoyo dan Samsuri mengategorikan menjadi 8 bunyi, dan Subardi menkategorikannya menjadi 10 bunyi dengan mendasarkan suara yang ada pada bahasa Jawa [6]. Halim dan Lapoliwa dalam Wahyuni [6] mengategorikan fonem vokal bahasa Indonesia menjadi 6.

Tabel 1. Daftar Fonem Vokal Bahasa Indonesia

Peneliti	Banyak Fonem	Daftar
Marsono	9	/a/, /i/, /u/, /e/, /ɛ/, /ə/, /o/, /ɔ/, dan /ɑ/
Dardjowidjoyo & Samsuri	8	/a/, /i/, /u/, /U/, /e/, /ɛ/, /ə/, dan /o/
Subardi	10	/a/, /i/, /I/, /u/, /U/, /e/, /ɛ/, /ə/, /ɔ/, dan /o/
Halim, Lapoliwa, Wahyuni	6	/a/, /i/, /u/, /e/, /ə/, dan /o/

Dalam penelitian ini, digunakan 5 kategori fonem vokal dengan meniadakan 'ə'dari kategori yang diberikan oleh Halim, Lapoliwa, dan Wahyuni. Hal ini dilakukan karena dalam penuturan huruf vokal tanpa didampingi konsonan dan semi konsonan, hanya fonem 'e' yang digunakan.

2.2 STFT

Transformasi *Fourier* (*Fourier Transform*) adalah sebuah transformasi matematis yang mendekomposisi fungsi dalam domain waktu ke frekuensi-frekuensi penyusunnya. Transformasi ini banyak digunakan terutama dalam pemrosesan sinyal (*signal processing*).

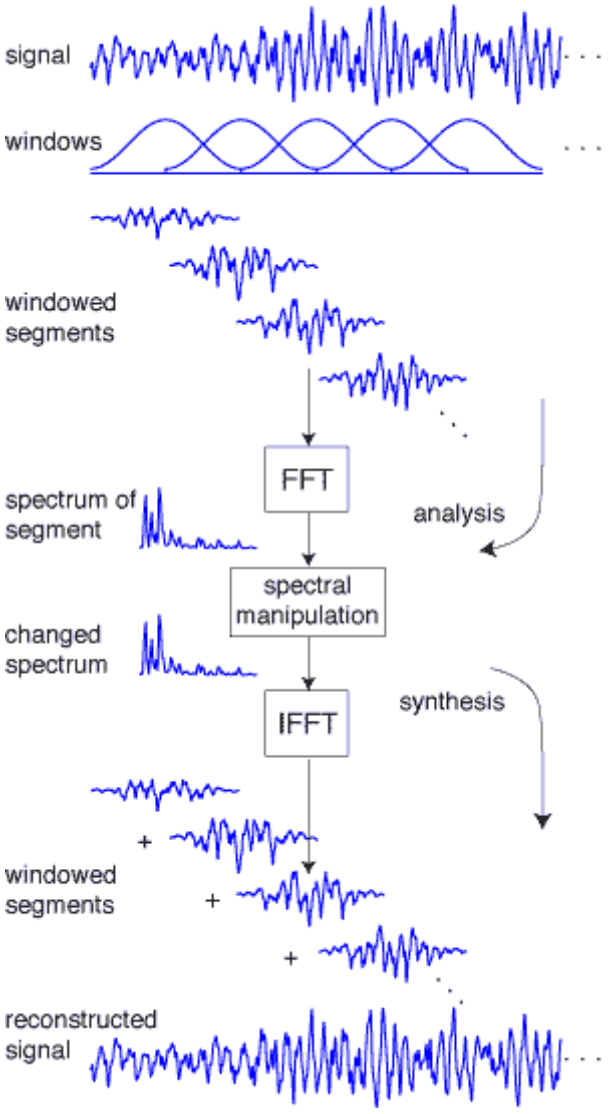
Didefinisikan dalam persamaan (1), transformasi *fourier* mengubah fungsi dalam domain waktu *f(t)* ke dalam fungsi lain dalam domain frekuensi *F(x)*.

$$F(x) = \int_{-\infty}^{\infty} f(t) e^{-ixt} dt \qquad (1)$$

Transformasi lain yang terkait dengan tranformasi *fourier* juga digunakan untuk mengubah dari domain waktu ke frekuensi baik untuk data yang kontinu maupun yang diskrit. Beberapa contohnya adalah transformasi hankel, Transformasi *Fourier* Diskrit (DFT), dan *Short Time Fourier Transform* (STFT).

STFT adalah transformasi *fourier* yang digunakan untuk menentukan frekuensi *sinusoidal* pada bagian lokal dari sinyal seiring dengan berubahnya sinyal tersebut terhadap waktu. Dengan kata lain, STFT adalah transformasi *fourier*

di sinyal berjendela (*windowed signal*). STFT memberikan informasi lokal (terhadap waktu) dari komponen frekuensi, berbeda dengan transformasi *fourier* standar yang menyediakan informasi frekuensi di sepanjang interval waktunya [7].



Gambar 3. *Short Time Fourier Transform*

Perubahan sinyal menggunakan STFT bersifat dapat diinverskan (*invertible*), artinya kita dapat mensintesis kembali sinyal asal dengan menggunakan *invers* dari prosesnya. Diagramnya dapat dilihat pada **Error! Reference source not found.** [8].

STFT diformulasikan sebagai perkalian antara sinyal dengan sebuah pembobot (yang disebut dengan *window* atau jendela) h(t) dimana τ adalah indeks spektro-temporal seperti pada persamaan (2).

$$F_\tau(x) = \int_{-\infty}^{\infty} f(t).h(t-\tau).e^{-ixt}dt \qquad (2)$$

Jendela h(t) dapat dipilih dari berbagai macam fungsi *sinusoidal* yang ada. Pada penelitian ini digunakan *Hanning Window* sebagai pembobotnya, ditunjukkan pada persamaan (3) dimana N adalah lebar jendelanya.

$$h(t) = sin^2\left(\frac{\pi t}{N}\right) \qquad (3)$$

2.3 MFCC

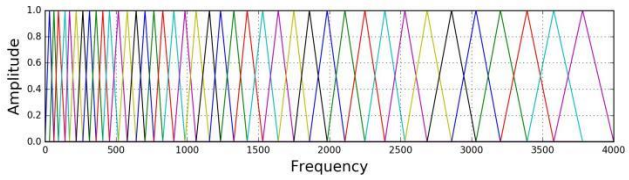
Fitur MFCC dikembangkan mengikuti sifat koklear (*cochlear*) dalam mempersepsikan suara di telinga manusia, yaitu dengan lebih mendiskriminasikan suara di frekuensi rendah dan kurang diskriminatif di frekuensi yang tinggi.

MFCC mengubah frekuensi f dalam satuan Hz ke dalam skala mel m (dengan satuan Mel) dan sebaliknya dengan fungsi pada persamaan (4) dan (5).

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad (4)$$

$$f = 700\left(10^{\frac{m}{2595}}-1\right) \qquad (5)$$

Skala tersebut diterapkan ke dalam spektogram sebagai *filter bank* berupa sekumpulan filter berbentuk segitiga yang saling beririsan seperti pada **Error! Reference source not found.**.



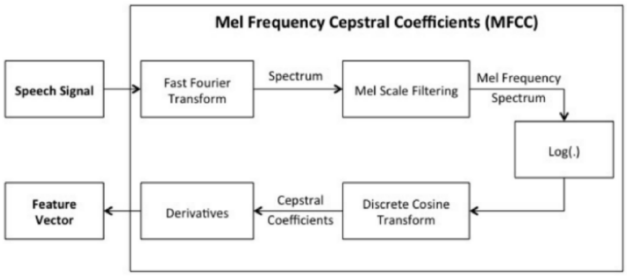
Gambar 4. *Filter Bank* MFCC

Filter tersebut dapat dimodelkan sebagai persamaan (6) berikut ini.

$$H_m(k) = \begin{cases} 0 & ,k < f(m-1) \\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & ,f(m-1) \le k < f(m) \\ 1 & ,k = f(m) \\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & ,f(m) < k \le f(m+1) \\ 0 & ,k > f(m+1) \end{cases}$$

MFCC digunakan untuk mengidentifikasi monosilabel dalam kalimat yang diutarakan secara kontinu [1]. Meskipun menurut Shrawankar dan Thakare [9], salah satu kelemahan MFCC adalah sensitifitasnya terhadap derau (*noise*) karena ketergantungannya terhadap bentuk *spectral*.

Pengambilan fitur menggunakan MFCC terdiri dari beberapa langkah komputasi seperti yang ditunjukkan dalam **Gambar 5**. Proses Pengambilan Fitur MFCC.



Gambar 5. Proses Pengambilan Fitur MFCC

2.4 Penelitian Terkait

Pada penelitian dengan objek penutur bahasa Indonesia atau Melayu, MFCC banyak digunakan untuk penentuan fitur silabel tertentu dalam bahasa Indonesia dan bahasa Arab. Contohnya untuk membedakan bacaan *gunnah* pada Al-Quran oleh Heriyanto [10] atau untuk mengklasifikasikan pengucapan huruf hijaiyah oleh Adiwijaya dkk. [11]. Syahroni dkk menggunakan MFCC untuk mengekstaksi fitur suku kata bahasa Indonesia [12] dan juga untuk membedakan suara digit [13]. Effendi juga menggunakan MFCC untuk mengenali pengaruh suara konsonan terhadap vokal [14]. Fitur yang didapatkan dari MFCC cukup *representative* dan sering dibandingkan dengan fitur Wavelet. Akurasi yang didapatkan tergantung dari data dan model klasifikasi yang dilakukan.

Banyak penelitian tentang pengenalan fonem seperti menggunakan *Time Delay Neural Network* [15], *Large Hierarchical Reservoirs* [16], *Convolution Neural Network* [17] [18], dan lain sebagainya. Sebagian yang lain memanfaatkan *Hidden Markov Model* untuk mendapatkan representasi model akustik dari data suara.

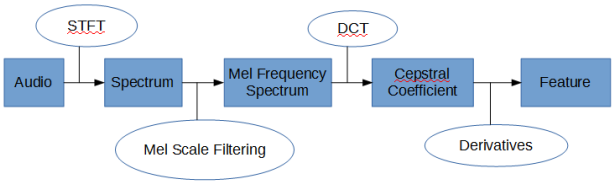
Menggunakan model yang manapun, untuk dapat meningkatkan akurasinya dibutuhkan pemahaman yang lebih terhadap objek yang akan dikenali, karena itulah pada penelitian ini peneliti mencoba mengeksplorasi tentang fonem vokal menggunakan MFCC.

3. METODE PENELITIAN

Penelitian ini dilakukan dalam tiga tahapan besar, yaitu pengumpulan data, pengolahan, dan eksplorasi.

Data dikumpulkan dari 51 partisipan, baik laki-laki maupun perempuan. Setiap responden diminta untuk bersuara a, i, u, e, dan o selama beberapa detik. Alat perekam menggunakan mikrofon kardioid yang dibawa oleh petugas lapangan atau menggunakan mikrofon dari *smartphone* yang tersedia.

Praproses dilakukan terhadap suara yang telah dikumpulkan untuk menyamakan formatnya.



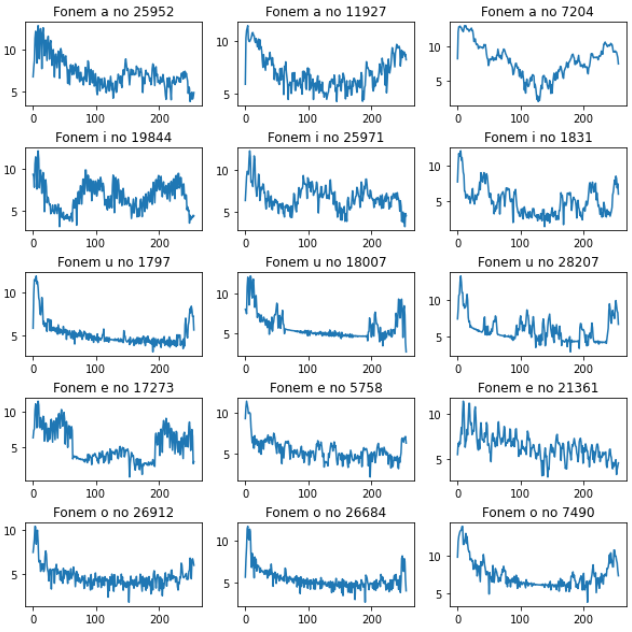
Gambar 6. Proses Pengolahan Data Audio

Tahapan pengolahan data audio ditunjukkan pada **Gambar 6.** Proses Pengolahan Data Audio. Suara yang dikumpulkan ditransformasi menggunakan STFT dan difilter menggunakan *Mel Scale Filtering* dan DCT untuk mendapatkan fiturnya.

Fitur yang terdiri dari n-buah koefisien mel, dinormalisasi dan dieksplorasi polanya satu sama lain. Data tersebut dibagi menjadi 2 kategori secara acak menjadi data *training* dan data *testing* dengan perbandingan 80:20. Dari data tersebut, dibuat beberapa buah model klasifikasi sederhana menggunakan algoritma SVM dan JST lalu dibandingkan akurasinya. Akurasi diuji menggunakan data *testing*.

4. HASIL DAN PEMBAHASAN

Dari 398 file audio yang berhasil dikumpulkan, dilakukan pra-proses untuk menyamakan frekuensi *sampling* menjadi 16.000 Hz dan format dalam bentuk wav. Suara di awal dan akhir file audio yang tidak berbunyi dipotong sehingga didapatkan file audio yang berisi fonem yang seragam dari awal hingga akhir.

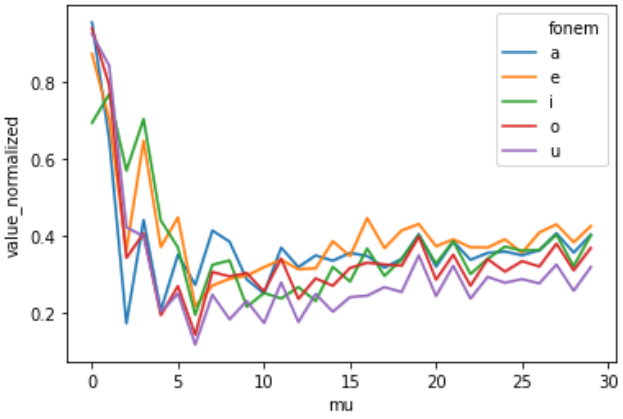


Gambar 7. Contoh Audio Pendek Fonem Vokal

Menggunakan *Hanning Window* dengan lebar 512, didapatkan 164.720 data audio pendek. Hasil *sampling* dari

data audio tersebut dapat dilihat pada **Gambar 7**. Contoh Audio Pendek Fonem Vokal.

Data tersebut lalu difilter menggunakan MFCC dengan 30 buah koefisien mel.



Gambar 8. Mel Koefisien (mu-0 s.d. mu-29)

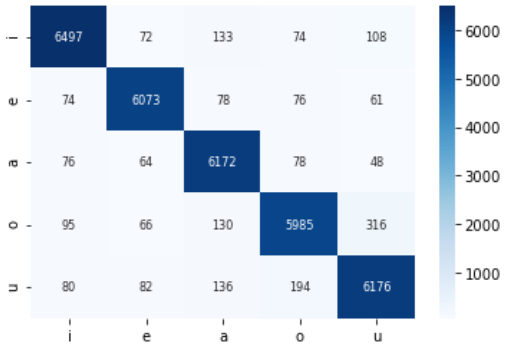
Hagen [19] mengatakan bahwa 10 hingga 12 koefisien mel sudah cukup untuk mengkodekan sinyal wicara. Dari visualisasi pada **Gambar 8.** Mel Koefisien (mu-0 s.d. mu-29) masih tampak perbedaan yang signifikan di koefisien ke-25 sehingga pada penelitian ini tetap menggunakan 30 koefisien mel sebagai fiturnya.

Model *Support Vector Machine* dengan 4 buah kernel yang berbeda dan Jaringan Syaraf Tiruan digunakan untuk mengklasifikasikan data tersebut. Dengan pembagian 80:20 untuk data *training* dan *testing*-nya, didapatkan hasil seperti ditunjukkan pada **Tabel 2.** Hasil Klasifikasi Fonem Vokal.

Tabel 2. Hasil Klasifikasi Fonem Vokal

No	Model	Akurasi
1	SVM kernel linear	74.56%
2	SVM kernel polynomial derajat 3	92.29%
3	SVM kernel radial	93.80%
4	SVM kernel sigmoid	27.62%
5	JST dengan 2 hidden layer	84.98%

Hasil yang paling tinggi akurasinya didapat dari model SVM dengan kernel radial. *Confusion matrix* dari model tersebut ditampilkan dalam **Gambar 9**. *Confusion Matrix* dari SVM dengan Kernel Radial berikut.



Gambar 9. *Confusion Matrix* dari SVM dengan Kernel Radial

5. KESIMPULAN

Fitur MFCC dapat membedakan fonem vokal dengan cukup jelas. Dengan akurasi tertinggi sebesar 93.8% yang diperoleh dari SVM kernel radial.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Direktorat Riset dan Pengabdian Masyarakat, Kementrian Riset, Teknologi, dan Pendidikan Tinggi atas dukungan yang diberikan berupa bantuan dana penelitian yang menunjang berlangsungnya penelitian ini dengan baik.

DAFTAR PUSTAKA

[1] N. K. b. A. M. R. Alim Sabur Ajibola, "*Some Commonly Used Speech Feature Extraction Algorithms*," 2018.

[2] N. v. d. Velde, 25 October 2018. [Online]. Available: https://www.globalme.net/blog/speech-recognition-software-history-future/.

[3] G. L. Song Wang, "*Overview of End-To-End Speech Recognition*," in *Journal of Physics: Conference Series*, 2019.

[4] J. Verhaar, "Asas-Asas Linguistik Umum," Yogyakarta: UGM Press, 1996.

[5] Marsono, "Fonetik," Yogyakarta: UGM Press, 1986.

[6] R. W. Primasari Wahyuni, "Kajian Fonetik Bunyi Vokal Bahasa Indonesia oleh Penutur Bahasa Indonesia di Wilayah Timur," in *PIBSI XXXIX*, Semarang, 2017.

[7] N. Kehtarnavaz, "*Frequency Domain Processing*," in *Digital Signal Processing System Design (Second Edition)*, Academic Press, 2008.

[8] W. A. Sethares, "*Rhythm and Transforms, Perception and Mathematics*," 2007.

[9] V. T. Urmila Shrawankar, "*Techniques for Feature Extraction In Speech Recognition System: A Comparative Study*," 2013.

[10] O. S. S. Heriyanto, "Identifikasi Suara Hukum Bacaan *Gunnah* menggunakan MFCC," in *Prosiding LPPM UPN Veteran*, Yogyakarta, 2016.

[11] M. N. A. M. S. M. W. U. N. a. F. N. Adiwijaya, "*A Comparative Study of MFCC-KNN and LPC-KNN for Hijaiyyah Letters Pronounciation Classification System*," in *5th International Conference on Information and Communication Technology (ICoIC7)*, Melaka, Malaysia, 2017.

[12] R. H. T. B. A. Syahroni Hidayat, "Sistem Pengenal Tutur Bahasa Indonesia berbasis Suku Kata menggunakan MFCC, Wavelet, dan HMM," Yogyakarta, 2015.

[13] S. H. Zaurarista Dyarbirru, "Metode Wavelet-MFCC dan Korelasi dalam Pengenalan Suara Digit," *Jurnal Teknologi Informasi dan Multimedia*, vol. 2, no. 2, pp. 100-108, 2020.

[14] M. M. Effendi, "Pengenalan Pengaruh Suara Konsonan terhadap Vokal menggunakan MFCC dan SVM," *IT For Society*, vol. 3, no. 2, 2018.

[15] T. H. G. H. K. S. a. K. L. A. Waibel, "*Phoneme Recognition using Time-delay Neural Networks*," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, p. 328–339, 1989.

[16] A. J. B. S. Fabian Triefenbach, "*Phoneme Recognition with Large Hierarchical Reservoirs*," in *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2010.

[17] R. C. M. M.-D. Dimitri Palaz, "*End-to-end Phoneme Sequence Recognition using Convolutional Neural Networks*," 2013.

[18] J. W. G. C. N. D. N. C. Cornelius Glackin, "*Convolutional Neural Networks for Phoneme Recognition*," in *7th International Conference on Pattern Recognition Applications and Methods*, 2018.

[19] C. D. P. B. Hagen A., "*The Analysis and Design of Architecture Systems for Speech Recognition on Modern Handheld Computing Devices*," 2003.