Analisis Fitur dan Convolutional Neural Network pada Pengenalan Aksen Ucapan

Dwi Sari Widyowaty[1], Andi Sunyoto[2], Hanif Al Fatta[3]
Universitas Amikom Yogyakarta[1,2,3]
dwi.1234@students.amikom.ac.id[1], andi@amikom.ac.id[2], hanif.a@amikom.ac.id[3]

Abstrak – Setiap negara memiliki ciri khas dan budaya masing-masing, salah satu ciri khas tersebut yaitu aksen ucapan, dengan mendengarkan aksen ucapan seseorang, maka dapat dikenali asal negara dari pembicara tersebut. Penelitian mengenai pengenalan aksen termasuk pada Teknologi Automatic Speech Recognition (ASR) yang sekarang ini sedang berkembang, contoh dari pemanfaatan teknologi ASR yaitu Asisten Virtual, pengembangan penelitian ini dapat menuju Asisten Virtual yang lebih cerdas karena dapat mengenali aksen dari seorang pembicara. Pada penelitian ini, penulis mencoba mengklasifikasikan aksen dari berbagai Negara (5 kelas) yaitu *English, Spanish, Mandarin, French* dan *Arabic*. Dataset yang digunakan pada Penelitian ini berjumlah 1231 rekaman suara yang terdiri dari *English* 627 audio, *Spanish* 220 audio, *Mandarin* 132 audio, *French* 80 audio, dan *Arabic* 172 audio, dimana seluruh pembicara mengucapkan kalimat yang sama dalam bahasa Inggris. Pada penelitian ini fitur audio yang digunakan yaitu Mel – Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), dan Energy (pada librosa disebut RMS). Ekstraksi Fitur audio menghasilkan array dari setiap audio, hasil ekstraksi fitur audio akan menjadi masukan Metode Convolutional Neural Network (CNN) untuk mengklasifikasikan aksen tersebut. Penelitian ini menghasilkan akurasi 51.30 % pada fitur MFCC, 48.05 % pada fitur ZCR, dan 51,95 % pada fitur Energy. Fitur Energy mendapatkan akurasi yang baik, kemudian diikuti dengan fitur MFCC dan ZCR.

Kata kunci: *Pengenalan Aksen; MFCC; Zero Crossing Rate; Energy; CNN.*

Abstract - Each country has its characteristics and culture, one of these characteristics is the accent of speech. By listening to someone's accent, we can identify the country of origin of the speaker. Research on accent recognition includes Automatic Speech Recognition (ASR) Technology which is currently being developed, an example of ASR technology, namely Virtual Assistant, the development of this research can be more intelligent Virtual Assistant because it can provide an accent from a speaker. In this study, the authors tried to classify accents from various countries (5 classes), namely English, Spanish, Mandarin, French and Arabic. The dataset used in this study consists of English 627 audio, Spanish 220 audio, Mandarin 132 audio, French 80 audio, and Arabic 172 audio, where all sentences are the same sentence in English. In this study, the audio features used are Mel - Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and Energy (in librosa it is called RMS). Audio feature extraction generates an array of each audio, the result of audio feature extraction will be the Convolutional Neural Network (CNN) Method input for classifying the accent. This research resulted in 51.30% accuracy for the MFCC feature, 48.05% for the ZCR feature, and 51.95% for the Energy feature. The Energy feature gets good accuracy, followed by the MFCC and ZCR features.

Keyword : Accent Recognition; MFCC; Zero Crossing Rate; Energy; CNN.

1. Latar Belakang

Bahasa Inggris adalah bahasa yang digunakan secara luas di belahan dunia, hampir semua penduduk di dunia wajib mempelajari bahasa Inggris, bahkan menggunakan bahasa Inggris di setiap hari. Ada sekitar 61 negara yang menjadikan bahasa Inggris sebagai bahasa resmi. Setiap Negara memiliki aksen yang berbeda, misalnya Amerika Serikat dan India memiliki aksen yang berbeda meskipun keduanya berbahasa Inggris. Perbedaan aksen ini dapat dipengaruhi oleh lingkungan, budaya, tempat lahir, usia, dan jenis kelamin [1].

Belakangan ini speech recognition sedang menjadi trend, contohnya pada virtual assistant, seperti google assist, siri, dan alexa namun asisten virtual yang ada saat ini belum mampu mengenali aksen pembicara dan negara asalnya. Oleh karena itu, penelitian tentang pengembangan sistem pengenalan aksen sangat diperlukan untuk menuju asisten virtual yang lebih canggih[2].

Beberapa penelitian telah dilakukan pada pengenalan suara, tahapan pengenalan aksen yaitu ekstraksi ciri dari sinyal audio menjadi array angka, kemudian klasifikasi suara.

Beberapa ekstraksi fitur audio yang telah dilakukan penelitian yaitu MFCC, Spectral Centroid and spread, Spectral Entropy, Spectral Flux, Chroma Vector, dan Spectral Roll Off [3]. Sedangkan metode yang telah digunakan, yaitu SVM, Naïve Bayes, Softmax Regression, GDA, GMM, dan K-Means [4]. Pada penelitian sebelumnya, dilakukan perbandingan ekstraksi fitur berdasarkan domain frekuensi [5] yang menghasilkan fitur MFCC menjadi fitur terbaik dengan akurasi paling tinggi, sedangkan pada penelitian tersebut, belum menganalisa ekstraksi fitur berdasarkan domain waktu (Zero Crossing Rate dan Energy), sehingga pada penelitian ini dilakukan perbandingan ekstraksi fitur MFCC (diambil dari fitur terbaik pada penelitian sebelumnya), Zero Crossing Rate, dan energy.

Metode kalsifikasi yang kami usulkan pada penelitian ini yaitu 2-layer CNN, outputnya adalah mengklasifikasikan *Native Language*. Penelitian ini disusun sebagai berikut : Kajian Pustaka di bagian 2, Implementasi Sistem dan Hasil di bagian 3, Kesimpulan ditunjukkan di bagian 4, dan Pustaka di bagian 5.

2. Kajian Pustaka

Analisis Fitur audio berdasarkan domain waktu telah dilakukan penelitian untuk clusterisasi berkas musik tradisional Indonesia, penelitian ini menyampaikan bahwa Zero Crossing Rate dan energy menghasilkan custerisasi yang baik[6]. Sedangkan beberapa penelitian tentang pengenalan aksen pernah dilakukan menggunakan dua kelas yaitu US dan Non US, dan metode yang diusulkan adalah KNN dan SVM, metode tersebut mencapai akurasi yang tinggi, KNN 96,05% dan SVM 95,09% [7]. Penelitian lain pada dua kelas yaitu menggunakan metode GMM dimana penelitian ini mengklasifikasikan Aksen Alquran yaitu Haf Al Asim dan Al Kisaie. Metode yang diusulkan mencapai akurasi tinggi 99,3% [8].

Penelitian berikutnya menggunakan banyak kelas yaitu lima kelas yang terdiri dari aksen bahasa Arab, Inggris, Prancis, Mandarin, dan Spanyol. Penelitian pertama [4] menyimpulkan bahwa GDA dan Naïve Bayes merupakan metode klasifikasi terbaik dengan akurasi 42%. Sedangkan penelitian kedua [5] menyimpulkan bahwa pada penggunaan dataset dengan panjang segmen berupa paragraf, MFCC dan CNN mampu mendapatkan akurasi sebesar 48,24%.

a. Mel-Frequency Cepstral Coefficients (MFCC)

Tahapan ekstraksi fitur MFCC terdiri dari Pre-emphasis, Windowing, Fast Fourier Transform, Mel Filter Bank, Descrete Cosine Transform, dan Delta Feature [9].

- Pre-emphasis

  Pada tahap ini sinyal suara akan disaring. Metode penyaringan ini mengurangi nilai frekuensi sinyal, hanya sinyal frekuensi tinggi yang dapat lolos penyaringan, pada langkah ini noise akan dikurangi, sehingga hanya data sinyal suara yang sebenarnya yang dapat ditangkap oleh sistem.
- Windowing

  Fungsi windowing untuk meminimalkan diskontinuitas pada frame awal dan akhir. Sebagian besar penelitian menggunakan hamming window karena kesederhanaan rumus dan nilai kerja windowing. Hamming window dapat dilihat pada persamaan (1) berikut:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N-1 \quad (1)$$

- Fast Fourier Transform (FFT)

  FFT adalah algoritma yang mengimplementasikan Discrete Fourier Transform (DFT), DFT mengubah setiap frame dari domain waktu ke domain frekuensi. Persamaan DFT dapat dilihat pada persamaan (2) berikut:

$$X_k = \sum_{n=0}^{N-1} X_n e^{-f2\frac{\pi}{N}kn} \quad (2)$$

- Mel Filter Bank

  Frekuensi linier yang diperoleh dari FFT diubah menjadi skala Mel-Frequency untuk mendapatkan limit bank filter berdasarkan Skala Frekuensi Mel. Persamaan Frekuensi Filter Mel dapat dilihat pada persamaan (3) berikut:

$$F_{mel} = 2595\ log\left(1 + \frac{f}{700}\right) \quad (3)$$

- Descrete Cosine Transform

  Hasil log dari domain waktu DCT disebut MFCC. Persamaan MFCC dapat dilihat pada persamaan (4) berikut:

$$X_k = \sum_{n=0}^{N-1} 2x[n] \quad \cos\left(\frac{\pi}{2N}k(2n+1)\right) \quad (4)$$

- Delta Feature

  Hasil perhitungan delta akan ditambahkan ke vektor ciri, sehingga menghasilkan vektor ciri yang lebih besar untuk meningkatkan akurasi sistem Automated

Speech Recognition (ASR). Metode ini akan menghasilkan koefisien delta sebesar koefisien ceptral yang dihasilkan oleh MFCC [10]. Fitur Delta dapat dihitung dengan persamaan (5) berikut.

$$d_{t=} \frac{\sum_{n=1}^{N} n(C_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (5)$$

b. Zero Crossing Rate

Zero Crossing Rate termasuk ke dalam metode ekstraksi fitur sinyal audio berdasarkan domain waktu.Zero Crossing Rate dapat dikatakan jumlah berapa kali dalam interval waktu/ bingkai sebuah amplitudo melawati nilai nol [3]. Persamaan Zero Crossing Rate dapat dilihat pada persamaan (6) dan (7) berikut :

$$Z_i = \frac{1}{2W_L}\sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \quad (6)$$

Dimana sgn adalah

$$sgn[xi(n)] = \quad (7)$$

$$\begin{cases} 1, xi(n) \geq 0, \\ -1, xi(n) < 0 \end{cases}$$

Keterangan :
Z : nilai *Zero Crossing*
X(n) : nilai amplitude pada data ke-n
WL : jumlah total bit yang ada pada frame

sgn : apabila lebih dari sama dengan 0 maka 1, dan apabila kurang dari 0 maka -1

c. Energy

Energy adalah salah satu ekstraksi fitur sinyal audio berdasarkan domain waktu.Energy juga dapat dikatakan nilai dari gabungan dari urutan frame [11] atau dapat juga dijelaskan seberapa tinggi dari sinyal audio. Biasanya Energy dinormalisasikan untuk menghilangkan ketergantungan pada panjang frame, Short-Time Energy dapat dihitung dengan persamaan (8) berikut:

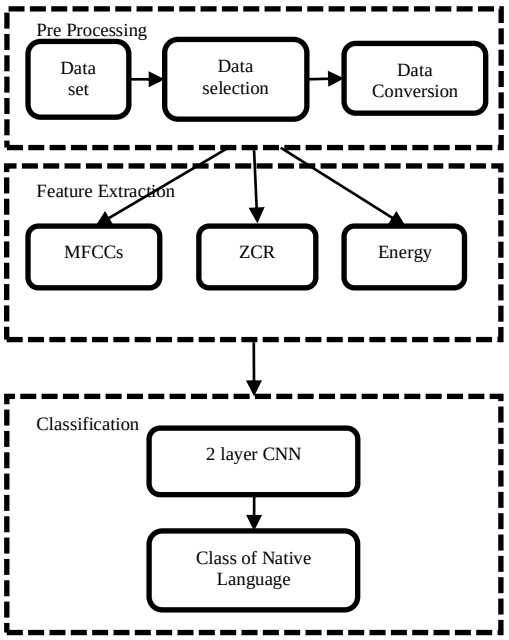
$$E(i) = \frac{1}{W_L}\sum_{n=1}^{WL} |x_i(n)|^2 \quad (8)$$

Keterangan :
E : nilai fitur short time energy
WL : jumlah sample yang ada pada frame i
X(n) : data sinyal pada window dengan panjang WL

3. Implementasi Sistem dan Hasil

Penelitian ini melalui beberapa tahapan yang dapat dilihat pada gambar 1. Tahap pertama yaitu pre-processing. Pada tahap pre-processing, dilakukan pengumpulan dataset, kemudian pemilihan data, dan koversi file Mp3 menjadi wav. Alasan pemilihan data yaitu agar menyamai jumlah dataset pada penelitian sebelumnya [5] sehingga nilai akurasi yang dihasilkan dapat dibandingkan menggunakan dataset yang sama. Sedangkan alasan dilakukan konversi dari Mp3 ke Wav karena file yang diperlukan pada library librosa adalah bertipe Wav. Tahap berikutnya yaitu melakukan ekstraksi fitur MFCC, ZCR, dan Energy. Hasil dari masing-masing fitur akan menjadi input dari Metode CNN, sehingga akan didapatkan 3 hasil penelitian, yaitu MFCC-CNN, ZCR-CNN, dan Energy-CNN.

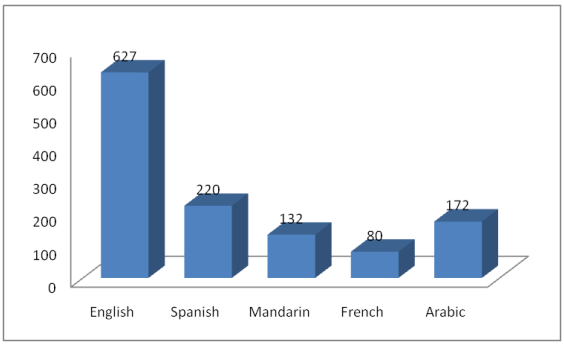
(3)



Gambar 1. Alur Penelitian

a. Dataset

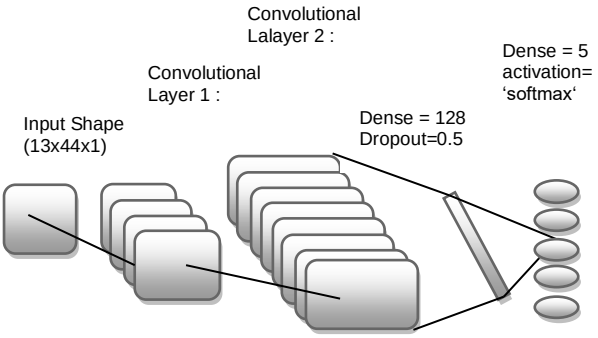
Penelitian ini menggunakan dataset publik yaitu The Speech Accent Archive dari George Mason University [12]. Total audio yang digunakan berjumlah 1231 rekaman suara dengan rincian dataset ((1) dilihat pada gambar 2 berikut.



Gambar 2. Jumlah Dataset

b. Convolutional Neural Netwok

Penelitian ini menggunakan 2-layer CNN, dengan input shape 13x44x1, Convolutional Layer 1 : 32 filters, kernel_size=3,3, activation = 'relu', MaxPooling2D=2,2, Strides=2,2, Padding ='same'. Convolutional Layer 2 : 64 filters, kernel_size=3,3, activation = 'relu', MaxPooling2D=2,2, Strides=2,2, Padding ='same', Dropout=0.25. Flatten Dense = 128, Dropout=0.5. Dense = 5, activation='softmax'. Arsitektur lengkap dapat dilihat pada gambar 3 berikut.



Gambar 3. Arsitektur CNN

c. Hasil Percobaan

Pada percobaan kali ini pembagian dataset yang digunakan yaitu data uji 25% dan data Validasi 20% dengan epoch 30. Kami menemukan bahwa epoch yang tinggi tidak menghasilkan akurasi yang baik, sehingga diambil 30 epoch saja. Hasil percobaan ini dapat dilihat pada tabel 1.

Tabel 1. Hasil Percobaan

Fitur	Akurasi	Loss
MFCC-CNN	51,30	1,57
ZCR-CNN	48,05	1,32
Energy-CNN	51,95	1,18

Tabel 1 menjelaskan bahwa Fitur Energy dan MFCC merupakan fitur yang terbaik, MFCC-CNN mampu menghasilkan akurasi sebesar 51,30 % dan Energy-CNN mampu menghasilkan akurasi sebesar 51,95 %, sedangkan fitur ZCR-CNN mendapatkan akurasi yang kurang baik yaitu 48,05 %.

4. Kesimpulan

Penelitian ini telah melakukan percobaan pada dataset yang sama dengan penelitian sebelumnya, namun peneliti menganalisis fitur lainnya yang belum dilakukan pada penelitian sebelumnya. Dalam penelitian sebelumnya, dataset diuji dan mencapai akurasi 48,24 % dengan fitur terbaik yaitu MFCC. Sedangkan pada penelitian ini dilakukan percobaan fitur lainnya yaitu fitur berdasarkan domain waktu (Zero Crossing Rate dan Energy) dibandingkan dengan MFCC (karena MFCC adalah fitur terbaik dari penelitian sebelumnya). Hasil dari perbandingan ini yaitu MFCC dan Energy mampu menghasilkan tingkat akurasi di atas 51 %. Sedangkan, Zero Crossing Rate menghasilkan akurasi yang kurang baik yaitu 48,05 %. Penelitian selanjutnya dapat memperbaiki sistem dengan meningkatkan nilai akurasi, dan menguji peran dropout untuk menghindari overfitting.

6. Pustaka

[1] B. D. Barkana and A. Patel, "Analysis of vowel production in Mandarin/Hindi/American- accented English for accent recognition systems," *Appl. Acoust.*, vol. 162, p. 107203, 2020.

[2] D. Honnavalli and Shylaja, "Supervised Machine Learning Model for Accent Recognition in English Speech using Sequential MFCC Features," *AIDE 2019*, 2019.

[3] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: a Matlab approach*. 2014.

[4] M. Bryant, A. Chow, and S. Li, "Classification of Accents of English Speakers by Native Language," pp. 1–5, 2014.

[5] Y. Singh, A. Pillay, and E. Jembere, "Features of speech audio for deep learning accent recognition," pp. 4–6, 2019.

[6] A. G. Jondya and B. H. Iswanto, "Analisis dan Seleksi Fitur Audio pada Musik Tradisional Indonesia," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 77, 2018.

[7] Z. Ma and E. Fokoué, "A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs," *Open J. Stat.*, vol. 04, no. 04, pp. 258–266, 2014.

[8] N. Kamarudin, S. A. R. Al-Haddad, S.

J. Hashim, M. A. Nematollahi, and A. R. Hassan, “Feature extraction using Spectral Centroid and Mel Frequency Cepstral Coefficient for Quranic Accent Automatic Identification,” *2014 IEEE Student Conf. Res. Dev. SCOReD 2014*, pp. 0–5, 2014.

[9] M. A. Imtiaz and G. Raja, “Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN,” *Proc. - APMediaCast 2016*, pp. 106–110, 2017.

[10] A. Winursito, R. Hidayat, and A. Bejo, “Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition,” *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018–Janua, pp. 379–383, 2018.

[11] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, “Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy,” *Adv. Tech. Comput. Sci. Softw. Eng.*, pp. 279–282, 2010.

[12] G. Mason and University, “The Speech Accent Archive.”