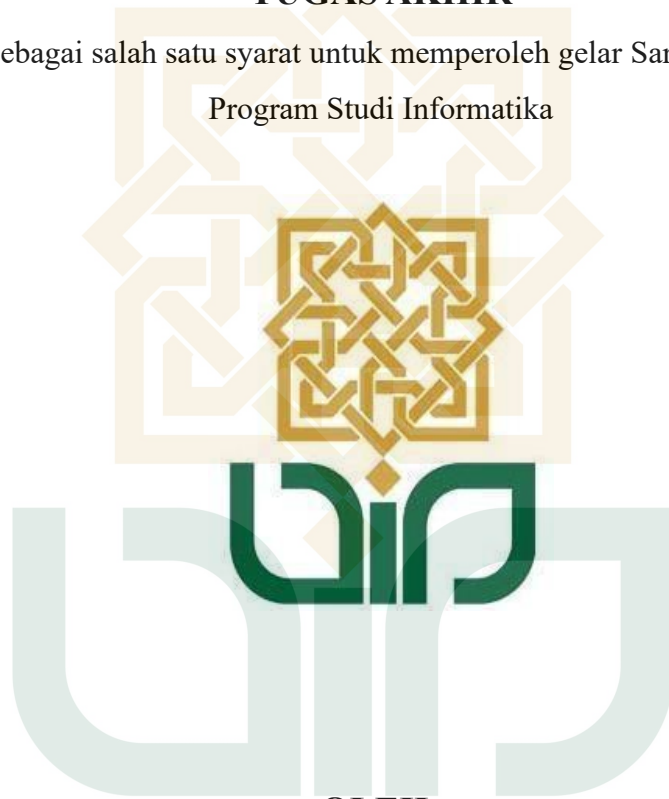


**Perbandingan Akurasi Ekstraksi Fitur Mel-Spektrogram dan
MFCC dalam Pengenalan Bunyi Konsonan Bilabial
Menggunakan CNN**

TUGAS AKHIR

Sebagai salah satu syarat untuk memperoleh gelar Sarjana S-1
Program Studi Informatika



OLEH:

PUTRA ANDIKA MEIFANI

NIM: 21106050031

**PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
YOGYAKARTA**

2026

HALAMAN PENGESAHAN



KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI

Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-2352/Un.02/DST/PP.00.9/09/2026

Tugas Akhir dengan judul : Perbandingan Akurasi Ekstraksi Fitur Mel-Spektrogram dan MFCC dalam Pengenalan Bunyi Konsonan Bilabial Menggunakan CNN

yang dipersiapkan dan disusun oleh:

Nama : PUTRA ANDIKA MEIFANI
Nomor Induk Mahasiswa : 21106050031
Telah diujikan pada : Senin, 31 Agustus 2026
Nilai ujian Tugas Akhir : A-

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR



Ketua Sidang
Dr. Syifaun Nafisah, S.T., MT.
SIGNED

Valid ID: 6ab1dc3a1e952



Penguji I
Ir. Muhammad Didik Rohmad Wahyudi, S.T., MT.
SIGNED

Valid ID: 6aacb7d0e42bc



Penguji II
Nurochman, S.Kom., M.Kom
SIGNED

Valid ID: 6aa3c8edaceef



Yogyakarta, 31 Agustus 2026
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi
Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6ab1e38da045b

PERNYATAAN KEASLIAN SKRIPSI

SURAT PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Putra Andika Meifani
NIM : 21106050031
Program Studi : Informatika
Fakultas : Sains dan Teknologi

Menyatakan dengan sesungguhnya, bahwa skripsi saya yang berjudul Analisis Akurasi Ekstraksi Fitur Mel-Spektrogram dan MFCC dalam Pengenalan Bunyi Konsonan Bilabial Menggunakan CNN merupakan hasil pekerjaan penulis sendiri sepanjang pengetahuan penulis, bukan duplikasi atau saduran dari karya orang lain kecuali bagian tertentu yang penulis ambil sebagai bahan acuan. Apabila terbukti pernyataan ini tidak benar, sepenuhnya menjadi tanggung jawab penulis.

Yogyakarta, 27 Agustus 2026



Putra Andika Meifani
21106050031

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

LEMBAR PERSETUJUAN SKRIPSI/TUGAS AKHIR



Universitas Islam Negeri Sunan Kalijaga



FM-UINSK-BM-05-03/RO

SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi / Tugas Akhir

Lamp : -

Kepada

Yth. Dekan Fakultas Sains dan Teknologi

UIN Sunan Kalijaga Yogyakarta

Di Yogyakarta

Assalammu 'alaikum wr. wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka saya selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama : Putra Andika Meifani

NIM : 21106050031

Judul Skripsi : Analisis Akurasi Ekstraksi Fitur Mel-Spektrogram dan MFCC dalam Pengenalan Bunyi Konsonan Bilabial Menggunakan CNN

sudah dapat diajukan kembali kepada Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Informatika.

Dengan ini kami berharap agar skripsi/tugas akhir Saudari dapat segera dimunaqasyahkan. Atas perhatiannya saya ucapkan terima kasih.

Wassalammu 'alaikum wr. wb.

Yogyakarta, 27 Agustus 2026

Pembimbing

Dr. Svifaun Nafisah, S.T., M.T.

NIP. 19781226 200801 2 017

LEMBAR PEDOMAN PENGGUNAAN TUGAS AKHIR

Tugas akhir ini tidak dipublikasikan, tetapi tersedia di perpustakaan dalam lingkungan UIN Sunan Kalijaga Yogyakarta, diperkenankan dipakai sebagai referensi kepustakaan, tetapi pengutipan harus seizin penyusun, dan harus menyebutkan sumbernya sesuai dengan kebiasaan ilmiah. Dokumen tugas akhir ini merupakan hak milik UIN Sunan Kalijaga Yogyakarta.



ABSTRAK

Pengenalan konsonan bilabial merupakan salah satu bagian dalam pengolahan sinyal suara yang dapat diterapkan pada sistem *Automatic Speech Recognition* (ASR). Representasi fitur suara memiliki peran penting dalam menentukan kemampuan model dalam melakukan klasifikasi. Penelitian ini bertujuan untuk menganalisis dan membandingkan performa representasi fitur Mel-Spektrogram dan *Mel-Frequency Cepstral Coefficients* (MFCC) dalam mengenali konsonan bilabial /p/, /b/, dan /m/ menggunakan *Convolutional Neural Network* (CNN).

Metode penelitian dilakukan melalui pengumpulan data rekaman suara, *preprocessing* audio, ekstraksi fitur Mel-Spektrogram dan MFCC, pembagian dataset menggunakan *Stratified Sample-Level Partitioning* dan *Speaker-Level Partitioning*, pelatihan model CNN, serta evaluasi model. Setiap model dievaluasi menggunakan 24 data uji dengan metrik akurasi dan *Root Mean Square Error* (RMSE). Kedua representasi fitur diuji pada masing-masing skenario pembagian dataset untuk mengetahui perbedaan performanya dalam melakukan klasifikasi konsonan bilabial.

Hasil penelitian menunjukkan bahwa Mel-Spektrogram memperoleh akurasi lebih tinggi dibandingkan MFCC pada kedua skenario pembagian dataset. Hasil terbaik berdasarkan akurasi diperoleh Mel-Spektrogram pada *Speaker-Level Partitioning* sebesar 83,33%, sedangkan hasil terbaik berdasarkan RMSE diperoleh MFCC pada *Speaker-Level Partitioning* sebesar 0,291898. Pada skenario *Stratified Sample-Level Partitioning*, Mel-Spektrogram memperoleh akurasi sebesar 75,00% dengan RMSE 0,362090, sedangkan MFCC memperoleh akurasi sebesar 70,83% dengan RMSE 0,344311. Dengan demikian, Mel-Spektrogram menunjukkan performa lebih baik berdasarkan akurasi, sedangkan MFCC menghasilkan nilai RMSE lebih rendah. Selain itu, *Speaker-Level Partitioning* menghasilkan performa evaluasi lebih baik dibandingkan *Stratified Sample-Level Partitioning* pada kedua representasi fitur dalam eksperimen yang dilakukan.

Kata Kunci : Mel-Spektrogram, MFCC, CNN, Fonetik dan Konsonan Bilabial

ABSTRACT

The recognition of bilabial consonants is one of the areas of speech signal processing that can be applied to Automatic Speech Recognition (ASR) systems. Speech feature representations play an important role in determining the ability of a model to perform classification. This study aims to analyze and compare the performance of Mel-Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) feature representations in recognizing the bilabial consonants /p/, /b/, and /m/ using a Convolutional Neural Network (CNN).

The research method consists of speech recording data collection, audio preprocessing, feature extraction using Mel-Spectrogram and MFCC, dataset partitioning using Stratified Sample-Level Partitioning and Speaker-Level Partitioning, CNN model training, and model evaluation. Each model was evaluated using 24 test data with accuracy and Root Mean Square Error (RMSE) as evaluation metrics. Both feature representations were tested under each dataset partitioning scenario to determine their performance differences in bilabial consonant classification.

The results show that Mel-Spectrogram achieved higher accuracy than MFCC under both dataset partitioning scenarios. The best accuracy was achieved by Mel-Spectrogram under Speaker-Level Partitioning, with an accuracy of 83.33%, while the best RMSE was achieved by MFCC under Speaker-Level Partitioning, with an RMSE of 0.291898. Under Stratified Sample-Level Partitioning, Mel-Spectrogram achieved an accuracy of 75.00% with an RMSE of 0.362090, while MFCC achieved an accuracy of 70.83% with an RMSE of 0.344311. Therefore, Mel-Spectrogram demonstrated better performance in terms of accuracy, whereas MFCC produced a lower RMSE. In addition, Speaker-Level Partitioning resulted in better evaluation performance than Stratified Sample-Level Partitioning for both feature representations in the conducted experiments.

Keywords: Mel-Spectrogram, MFCC, CNN, Phonetics, and Bilabial Consonants

HALAMAN MOTTO

“Demi waktu dhuha, dan demi waktu malam apabila telah sunyi, tuhanmu (Nabi Muhammad) tidak meninggalkan dan tidak (pula) membencimu.”

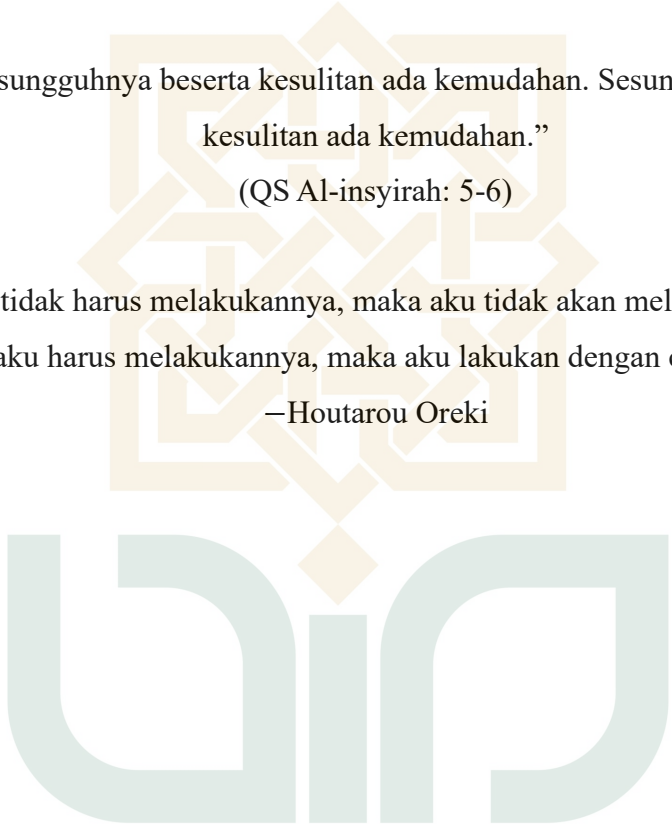
(QS Ad-dhuha: 1-3)

“Maka, sesungguhnya beserta kesulitan ada kemudahan. Sesungguhnya beserta kesulitan ada kemudahan.”

(QS Al-insyirah: 5-6)

“Jika aku tidak harus melakukannya, maka aku tidak akan melakukannya. Jika aku harus melakukannya, maka aku lakukan dengan cepat.”

—Houtarou Oreki



STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

HALAMAN PERSEMBAHAN

Atas karunia Allah Swt. skripsi ini penuli persembahkan
kepada:

MOCH. Fadil, Anik, dan Damai Daffa Alamsyah
Keluarga tercinta

Sahabat dan teman seperjuangan
Yang selalu memberikan dukungan untuk penulis

Dan
Almamater tercinta

Program Studi Informatika
Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

KATA PENGANTAR

Assalamu'alaikum warahmatullahi wabarakatuh.

Segala puji dan syukur penulis panjatkan ke hadirat Allah Swt. atas limpahan rahmat, taufik, dan hidayah-Nya, sehingga penulis dapat menyelesaikan penyusunan skripsi ini dengan baik. Sholawat serta salam senantiasa tercurah kepada Nabi Muhammad saw. yang telah membawa umat manusia dari zaman kegelapan menuju cahaya kebenaran. Skripsi yang berjudul “Perbandingan Akurasi Ekstraksi Fitur Mel-Spektrogram dan MFCC dalam Pengenalan Bunyi Konsonan Bilabial Menggunakan CNN” ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Informatika, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga Yogyakarta. Penyusunan skripsi ini tentu tidak terlepas dari dukungan, bantuan, dan doa dari berbagai pihak. Oleh karena itu, penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Noorhaidi Hasan, S.Ag., M.A., M.Phil., Ph.D. selaku Rektor Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
2. Ibu Prof. Dr. Dra. Hj. Khurul Wardati, M.Si. selaku Dekan Fakultas Sains dan Teknologi.
3. Bapak Dr. Muhammad Mustakim, S.T., M.T. sebagai Ketua Program Studi Informatika yang telah memberikan dukungan selama masa perkuliahan.
4. Bapak Dr. Ir. Aulia Faqih Rifa'i, M.Kom. sebagai dosen pembimbing akademik yang selalu membimbing dan mengarahkan selama menjalani perkuliahan.
5. Ibu Dr. Syifaun Nafisah, S.T., M.T. sebagai dosen pembimbing skripsi yang telah memberikan dukungan luar biasa, meluangkan waktu, bimbingan, serta motivasi untuk menyelesaikan skripsi ini.
6. Keluarga tercinta Bapak MOCH. Fadil, Ibu Anik, dan adik Damai Daffa Alamsyah, yang selalu menjadi sumber semangat dan kekuatan bagi penulis dalam setiap langkah, atas doa, cinta, dan dukungan tanpa henti.
7. Rekan-rekan seperjuangan di Program Studi Informatika serta semua pihak yang telah memberikan bantuan dalam berbagai bentuk, baik secara

langsung maupun tidak langsung, selama masa perkuliahan hingga proses penyusunan skripsi ini.

8. 35 teman-teman yang telah berkenan menjadi responden untuk penelitian ini, tanpa bantuan kalian mungkin penulis tidak dapat menyelesaikan penelitian ini.
9. Seluruh teman-teman yang tidak dapat penulis sebutkan satu per satu, penulis menyampaikan terima kasih atas segala bentuk bantuan, doa, dan dukungan yang telah diberikan.



DAFTAR ISI

HALAMAN PENGESAHAN.....	i
PERNYATAAN KEASLIAN SKRIPSI.....	ii
LEMBAR PERSETUJUAN SKRIPSI/TUGAS AKHIR.....	iii
LEMBAR PEDOMAN PENGGUNAAN TUGAS AKHIR	iv
ABSTRAK	v
HALAMAN MOTTO	vii
HALAMAN PERSEMBAHAN	viii
KATA PENGANTAR	ix
DAFTAR ISI.....	xi
DAFTAR GAMBAR	xiv
DAFTAR TABEL.....	xv
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Batasan Masalah.....	5
1.4 Tujuan dan Manfaat	5
BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI	7
2.1 Tinjauan Pustaka	7
2.2 Landasan Teori.....	11
2.2.1 Automatic Speech Recognition (ASR)	11
2.2.2 Fonem.....	12
2.2.3 Konsonan	13
2.2.4 Tempat Artikulasi.....	14
2.2.5 Ekstraksi Fitur Audio	14
2.2.6 Mel-Spektrogram	15
2.2.7 MFCC.....	21
2.2.8 CNN	22
2.2.9 Evaluasi Model.....	24
BAB III METODE PENELITIAN.....	28

3.1	Jenis Penelitian.....	28
3.2	Tahapan Penelitian	29
3.2.1	Pengumpulan Data	30
3.2.2	Preprocessing Data Audio.....	32
3.2.3	Ekstraksi Fitur	35
3.2.4	Pembagian Dataset	39
3.2.5	Pelatihan Model CNN	40
3.2.6	Evaluasi Model.....	43
BAB IV HASIL DAN PEMBAHASAN		45
4.1	Hasil Pengumpulan Data.....	45
4.2	Hasil Preprocessing Audio.....	45
4.3	Skenario Pengujian.....	46
4.4	Hasil Ekstraksi Fitur.....	47
4.4.1	Hasil Ekstraksi Fitur Mel-Spektrogram	47
4.4.2	Hasil Ekstraksi Fitur MFCC.....	50
4.5	Hasil Pembagian Dataset.....	52
4.5.1	Hasil Pembagian Menggunakan Stratified Sample-Level Partitioning 52	
4.5.2	Hasil Pembagian Menggunakan Speaker-Level Partitioning	53
4.6	Hasil Pelatihan Model CNN.....	54
4.6.1	Hasil Pelatihan dengan Stratified Sample-Level Partitioning.....	54
4.6.2	Hasil Pelatihan dengan Speaker-Level Partitioning.....	56
4.7	Hasil Evaluasi Model	57
4.7.1	Hasil Evaluasi Stratified-Level Partitioning	57
4.7.2	Hasil Evaluasi Speaker-Level Partitioning	58
4.8	Perbandingan Hasil	59
4.8.1	Perbandingan Hasil Berdasarkan Representasi Fitur	60
4.8.2	Perbandingan Hasil Berdasarkan Metode Pembagian Dataset	61
4.9	Pembahasan.....	63
4.9.1	Perbandingan Mel-Spektrogram dan MFCC	63
4.9.2	Pengaruh Metode Pembagian Dataset.....	64
4.9.3	Hubungan Hasil Pelatihan dan Evaluasi	64

4.9.4 Hasil Terbaik Penelitian	65
BAB V PENUTUP.....	66
5.1 Kesimpulan	66
5.2 Saran.....	67
DAFTAR PUSTAKA	69
DAFTAR RIWAYAT HIDUP.....	75



DAFTAR GAMBAR

Gambar 3.1 Diagram Alir Tahapan Penelitian	29
Gambar 3.2 Contoh Waveform Bunyi P	34
Gambar 3.3 Contoh Waveform Bunyi B	35
Gambar 3.4 Contoh Waveform Bunyi M	35
Gambar 3.6 Contoh Mel-Spektrogram Bunyi B	37
Gambar 3.5 Contoh Mel-Spektrogram Bunyi P	37
Gambar 3.7 Contoh Mel-Spektrogram Bunyi M	37
Gambar 3.9 Contoh MFCC Bunyi B	38
Gambar 3.8 Contoh MFCC Bunyi P	38
Gambar 3.10 Contoh MFCC Bunyi M	38
Gambar 4.1 Mel-Spektrogram bunyi /p/ pada responden P_001	48
Gambar 4.2 Mel-Spektrogram bunyi /b/ pada responden B_001	49
Gambar 4.3 Mel-Spektrogram bunyi /m/ pada responden M_001	49
Gambar 4.4 MFCC bunyi /p/ pada responden P_001	50
Gambar 4.5 MFCC bunyi /b/ pada responden B_001	51
Gambar 4.6 MFCC bunyi /m/ pada responden M_001	51

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

DAFTAR TABEL

Tabel 2.1 Tinjauan Pustaka	8
Tabel 3.1 Hasil Preprocessing	33
Tabel 4.1 File Audio Digunakan	45
Tabel 4.2 Pembagian Dataset dengan Stratified Sample-Level Partitioning	53
Tabel 4.3 Pembagian Dataset dengan Speaker-Level Partitioning	54
Tabel 4.4 Hasil Pelatihan CNN dengan Mel-Spektrogram	55
Tabel 4.5 Hasil Pelatihan CNN dengan MFCC	56
Tabel 4.6 Hasil Pelatihan CNN dengan Mel-Spektrogram	56
Tabel 4.7 Hasil Pelatihan CNN dengan MFCC	57
Tabel 4.8 Hasil Evaluasi Pada Skenario Stratified Sample-Level Partitioning	58
Tabel 4.9 Hasil Evaluasi Pada Skenario Speaker-Level Partitioning.....	59
Tabel 4.10 Perbandingan Hasil Pelatihan Berdasarkan Representasi Fitur	60
Tabel 4.11 Perbandingan Hasil Evaluasi Berdasarkan Representasi Fitur.....	61
Tabel 4.12 Hasil Evaluasi Mel-Spektrogram Berdasarkan Metode Pembagian Dataset.....	62
Tabel 4.13 Hasil Evaluasi MFCC Berdasarkan Metode Pembagian Dataset.....	62

STATE ISLAMIC UNIVERSITY
SUNAN KALIJAGA
YOGYAKARTA

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam komunikasi lisan sehari-hari, kesalahan dalam memahami bunyi yang diucapkan oleh seseorang dapat terjadi ketika suatu bunyi memiliki karakteristik akustik yang mirip dengan bunyi lainnya. Kondisi tersebut dapat menyebabkan pendengar salah menangkap atau menafsirkan bunyi yang disampaikan, terutama ketika ujaran diucapkan dengan cepat, terdapat gangguan suara di lingkungan sekitar, atau terdapat perbedaan karakteristik suara antarpemutar. Sebagai contoh, bunyi /p/, /b/, dan /m/ memiliki kesamaan dalam tempat artikulasi karena ketiganya dihasilkan dengan melibatkan kedua bibir, tetapi memiliki perbedaan pada cara artikulasi dan karakteristik suara yang dihasilkan. Perbedaan karakteristik tersebut dapat menjadi informasi penting untuk membedakan satu bunyi dengan bunyi lainnya. Fenomena ini menunjukkan bahwa kemampuan dalam mengenali dan membedakan bunyi ujaran merupakan bagian penting dalam memahami informasi yang disampaikan melalui komunikasi lisan. Agar perbedaan karakteristik akustik tersebut dapat dikenali secara sistematis oleh komputer, sinyal suara perlu direpresentasikan ke dalam bentuk fitur yang mampu mempertahankan informasi akustik yang relevan.

Permasalahan dalam mengenali dan membedakan karakteristik bunyi tersebut juga menjadi salah satu aspek yang dapat ditangani melalui teknologi pengenalan ucapan. Pengenalan ucapan merupakan salah satu bidang penelitian yang terus berkembang dalam teknologi kecerdasan buatan dan pembelajaran mesin. Teknik ini memungkinkan mesin untuk memahami dan memproses input suara manusia, sehingga membuka peluang pemanfaatan di berbagai sektor seperti identifikasi pembicara, autentikasi pengguna, hingga pemrosesan informasi berbasis bahasa lisan. Namun, mengingat sinyal suara manusia bersifat sangat variatif dan kompleks, pemilihan metode ekstraksi fitur akustik menjadi langkah penting dalam analisis pengenalan ucapan [1].

Mel-Spektrogram dan *Mel-Frequency Cepstral Coefficients* (MFCC) adalah dua metode ekstraksi fitur akustik yang dapat digunakan untuk merepresentasikan karakteristik sinyal suara. Mel-Spektrogram merepresentasikan distribusi energi suara terhadap waktu dan frekuensi pada skala mel, sehingga pola perubahan energi dan frekuensi pada sinyal ucapan dapat dipertahankan. Sementara itu, MFCC menghasilkan representasi karakteristik spektral suara yang lebih ringkas dan berkaitan dengan informasi fonetik. Perbedaan karakteristik representasi tersebut memberikan pendekatan yang berbeda dalam menggambarkan informasi akustik suara. Oleh karena itu, penggunaan Mel-Spektrogram dan MFCC memungkinkan karakteristik akustik bunyi /p/, /b/, dan /m/ dianalisis dari dua bentuk representasi fitur yang berbeda [2].

Sistem pengenalan ucapan otomatis atau biasa disebut *Automatic Speech Recognition* (ASR) yang memanfaatkan fitur-fitur tersebut diklasifikasikan menjadi beberapa kategori berdasarkan unit pemrosesannya, yaitu berbasis fonem (satuan bunyi terkecil), berbasis suku kata (*syllable*), berbasis kata (*word*), berbasis frasa (*phrase*), dan berbasis kalimat (*sentence*), setiap kategori ini memiliki karakteristik dan aplikasi tersendiri sesuai dengan kompleksitas bahasa dan kebutuhan sistem [3][4][5]. Di antara berbagai kategori tersebut, unit proses berbasis fonem menempati posisi paling fundamental karena fonem merupakan satuan bunyi terkecil yang membedakan makna dalam suatu bahasa, serta menawarkan keunggulan strategis berupa generalisasi yang lebih baik untuk kata di luar *vocabulary*, efisiensi data latih, skalabilitas untuk *Large Vocabulary Continuous Speech Recognition* (LVCSR), dapat diterapkan pada sistem multibahasa, dan kemampuan menangani variasi *co-articulation* melalui *modeling di-phone* dan *tri-phone* [3][6]. Meskipun sistem berbasis kata memiliki *success rate* lebih tinggi untuk *vocabulary* kecil, keterbatasannya pada skalabilitas dan kebutuhan data menjadikan pendekatan berbasis fonem sebagai salah satu pendekatan yang paling banyak digunakan dalam sistem ASR modern. [7].

Dalam sistem ASR berbasis fonem, pemilihan metode ekstraksi fitur menjadi faktor penting karena fitur yang dihasilkan harus mampu merepresentasikan karakteristik akustik setiap fonem secara efektif. Studi sebelumnya menunjukkan

bahwa kedua teknik ekstraksi fitur tersebut memiliki keunggulan tersendiri dalam merepresentasikan karakteristik akustik bunyi konsonan [8]. Salah satu bunyi konsonan yang menjadi fokus kajian dalam penelitian pengenalan ucapan adalah konsonan bilabial, karena artikulasinya terjadi pada kedua bibir sehingga menghasilkan petunjuk visual yang kuat dan informasi visual dari bibir terbukti meningkatkan *intelligibility* serta akurasi ASR [9]. Konsonan bilabial yang diproduksi melalui kontak langsung antara kedua bibir merupakan salah satu kategori tempat artikulasi yang secara konsisten dianalisis dalam penelitian karakteristik artikulasi konsonan untuk sistem pengenalan ucapan. Kajian terhadap konsonan ini tidak hanya memperkaya pemahaman linguistik, tetapi juga mendukung pengembangan teknologi pengenalan ucapan yang berbasis fitur artikulasi [10]. Pada sistem ASR konvensional, proses pengenalan ucapan umumnya dilakukan menggunakan metode seperti *Hidden Markov Model* (HMM), *Gaussian Mixture Model* (GMM), dan *Dynamic Time Warping* (DTW). Meskipun metode-metode tersebut telah banyak digunakan dan menunjukkan kinerja yang baik pada berbagai tugas pengenalan ucapan, performanya masih bergantung pada kualitas ekstraksi fitur yang dirancang secara manual serta memiliki keterbatasan dalam memodelkan hubungan nonlinier yang kompleks pada sinyal ucapan [11].

Perkembangan *deep learning* mendorong penggunaan berbagai model berbasis kecerdasan buatan, seperti *Artificial Neural Network* (ANN), *Long Short-Term Memory* (LSTM), dan *Convolutional Neural Network* (CNN), yang mampu mempelajari representasi fitur secara otomatis dari data. ANN memiliki kemampuan untuk memodelkan hubungan nonlinier yang kompleks, namun arsitektur *fully connected* yang digunakan kurang optimal dalam memanfaatkan keterkaitan lokal antarfitur akustik [12]. Di sisi lain, LSTM dirancang untuk menangkap dependensi temporal jangka panjang pada data sekuensial sehingga banyak digunakan pada tugas pengenalan ucapan berkelanjutan [13]. Akan tetapi, penelitian ini berfokus pada klasifikasi konsonan berdasarkan karakteristik akustiknya, bukan pada pemodelan urutan ucapan yang panjang. CNN dipilih karena mampu mengekstraksi pola lokal secara otomatis melalui mekanisme konvolusi dan *pooling*, sehingga efektif dalam mempelajari karakteristik akustik

dari representasi fitur ucapan [6]. Kemampuan ini sesuai dengan penggunaan Mel-Spektrogram dan MFCC yang sama-sama mengandung informasi penting mengenai distribusi energi dan karakteristik frekuensi sinyal ucapan. Selain itu, CNN mampu membangun representasi fitur secara hierarkis sehingga dapat mengenali pola akustik yang relevan untuk membedakan kelompok konsonan berdasarkan tempat artikulasinya [12]. Oleh karena itu, penelitian ini memanfaatkan arsitektur CNN untuk mengklasifikasikan konsonan bilabial berdasarkan fitur Mel-Spektrogram dan MFCC.

Meskipun beberapa studi telah membandingkan performa Mel-Spektrogram dan MFCC dalam pengenalan ucapan secara umum, seperti studi yang telah dilakukan oleh Meghanani, et al. membahas terkait pengenalan demensia alzheimer dari ucapan spontan [14]. Namun, kajian yang secara spesifik berfokus pada konsonan bilabial masih sangat terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi potensi kedua metode tersebut, menggunakan CNN sebagai pengklasifikasi utama untuk mendukung analisis yang lebih mendalam. Penelitian ini tidak hanya memberikan kontribusi dalam pengenalan ucapan konsonan bilabial secara spesifik, tetapi juga membuka peluang untuk pengembangan lebih lanjut dalam skala yang lebih kompleks. Dengan menggunakan huruf sebagai fokus utama, hasil penelitian ini dapat dijadikan dasar untuk mengembangkan sistem pengenalan ucapan yang mampu mengenali kata, menyusun kalimat, hingga menganalisis makna emosional yang terkandung dalam kalimat tersebut. Fokus pada konsonan bilabial dalam penelitian ini juga memiliki potensi untuk memperluas cakupan aplikasi, seperti dalam pengembangan sistem pengenalan emosi berbasis suara, terapi bicara berbasis teknologi, dan interaksi manusia-mesin yang lebih intuitif [15]. Dengan demikian, penelitian ini berfungsi sebagai pijakan awal untuk menjembatani analisis perbandingan metode ekstraksi fitur akustik menuju aplikasi linguistik dan teknologi suara yang lebih canggih.

1.2 Rumusan Masalah

1. Bagaimana performa Mel-Spektrogram dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN?
2. Bagaimana performa MFCC dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN?
3. Bagaimana perbandingan performa Mel-Spektrogram dan MFCC dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN?

1.3 Batasan Masalah

Penelitian ini dibatasi pada pengenalan suara konsonan bilabial, yaitu /p/, /b/, dan /m/, yang diklasifikasikan berdasarkan tempat artikulasi. Fokus penelitian hanya mencakup konsonan bilabial karena bunyi ini memiliki karakteristik artikulasi yang jelas dan konsisten, sehingga relevan untuk dieksplorasi lebih lanjut dalam konteks pengenalan suara. Data yang digunakan merupakan rekaman audio yang dikumpulkan secara mandiri melalui Google Form, di mana partisipan diminta merekam pengucapan suku kata bilabial dalam format "pe", "be", dan "em", yang masing-masing mewakili konsonan /p/, /b/, dan /m/. Partisipan terdiri dari mahasiswa dan masyarakat umum dengan variasi usia, jenis kelamin, dan dialek untuk memastikan data yang representatif terhadap variasi alami pengucapan konsonan bilabial bahasa Indonesia.

1.4 Tujuan dan Manfaat

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa dua metode ekstraksi fitur akustik, yaitu Mel-Spektrogram dan MFCC, dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN. Evaluasi dilakukan untuk memahami sejauh mana kedua metode ini mampu merepresentasikan pola akustik yang relevan dan mendukung pengenalan suara secara akurat. Adapun tujuan penelitian dirumuskan sebagai berikut:

1. Menganalisis performa Mel-Spektrogram dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN.

2. Menganalisis performa MFCC dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN.
3. Membandingkan performa Mel-Spektrogram dan MFCC dalam mengenali konsonan bilabial (/p/, /b/, dan /m/) menggunakan CNN.

Penelitian ini memberikan manfaat yang dapat dibagi menjadi dua aspek utama, yaitu manfaat teoritis dan praktis, dengan rincian sebagai berikut:

1. Manfaat Teoritis:
 - a. Memberikan kontribusi terhadap pengembangan ilmu linguistik, khususnya dalam bidang linguistik terapan dan teknologi pengenalan suara.
 - b. Menjadi referensi bagi penelitian lanjutan yang berfokus pada:
 - 1.) Eksplorasi metode ekstraksi fitur akustik lainnya.
 - 2.) Implementasi teknologi pembelajaran mesin pada analisis suara.
2. Manfaat Praktis:
 - a. Mendukung pengembangan aplikasi berbasis suara, seperti:
 - 1.) Asisten virtual.
 - 2.) Sistem pengenalan suara otomatis.
 - 3.) Alat bantu komunikasi untuk kebutuhan khusus.
 - 4.) Sistem temu kembali informasi.
 - b. Fokus pada pengucapan konsonan bilabial bahasa Indonesia memungkinkan hasil penelitian diimplementasikan dalam teknologi yang digunakan untuk interaksi manusia.
 - c. Mendukung pengembangan sistem berbasis suara yang lebih akurat dan efisien.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai klasifikasi konsonan bilabial /p/, /b/, dan /m/ menggunakan *Convolutional Neural Network* (CNN) dengan fitur Mel-Spektrogram dan *Mel-Frequency Cepstral Coefficients* (MFCC), dapat diperoleh beberapa kesimpulan sebagai berikut:

1. Mel-Spektrogram mampu digunakan sebagai representasi fitur untuk mengenali konsonan bilabial menggunakan CNN. Berdasarkan hasil evaluasi, Mel-Spektrogram memperoleh hasil terbaik pada skenario *Speaker-Level Partitioning* dengan akurasi sebesar 83,33% dan RMSE sebesar 0,303359. Sementara itu, pada skenario *Stratified Sample-Level Partitioning*, Mel-Spektrogram memperoleh akurasi sebesar 75,00% dengan RMSE sebesar 0,362090.
2. MFCC juga mampu digunakan sebagai representasi fitur untuk mengenali konsonan bilabial menggunakan CNN. Hasil terbaik MFCC diperoleh pada skenario *Speaker-Level Partitioning* dengan akurasi sebesar 79,17% dan RMSE sebesar 0,291898. Pada skenario *Stratified Sample-Level Partitioning*, MFCC memperoleh akurasi sebesar 70,83% dengan RMSE sebesar 0,344311.
3. Hasil perbandingan menunjukkan bahwa Mel-Spektrogram memperoleh akurasi lebih tinggi dibandingkan MFCC pada kedua metode pembagian dataset. Namun, MFCC secara konsisten menghasilkan nilai RMSE yang lebih rendah. Dengan demikian, Mel-Spektrogram lebih unggul berdasarkan akurasi, sedangkan MFCC lebih unggul berdasarkan RMSE.
4. *Speaker-Level Partitioning* menghasilkan hasil evaluasi yang lebih baik dibandingkan *Stratified Sample-Level Partitioning* pada eksperimen yang dilakukan. Kombinasi *Speaker-Level Partitioning* dan Mel-Spektrogram menghasilkan akurasi tertinggi sebesar 83,33%, sedangkan kombinasi *Speaker-Level Partitioning* dan MFCC menghasilkan RMSE terendah

sebesar 0,291898. Namun, hasil tersebut terbatas pada dataset dan komposisi eksperimen yang digunakan dalam penelitian ini.

Secara keseluruhan, hasil penelitian menunjukkan bahwa karakteristik akustik konsonan bilabial /p/, /b/, dan /m/ dapat direpresentasikan menggunakan Mel-Spektrogram maupun MFCC dan dimanfaatkan oleh CNN untuk melakukan klasifikasi secara otomatis. Pendekatan tersebut menunjukkan bahwa perbedaan karakteristik akustik pada ketiga konsonan yang memiliki kesamaan tempat artikulasi dapat dimanfaatkan sebagai informasi dalam proses pengenalan bunyi. Namun, akurasi tertinggi yang diperoleh, yaitu 83,33%, menunjukkan bahwa masih terdapat kesalahan dalam proses klasifikasi sehingga kinerja model belum optimal. Akurasi tersebut masih memiliki peluang untuk ditingkatkan melalui pengembangan lebih lanjut, seperti penambahan dan keberagaman data rekaman, pengaturan parameter ekstraksi fitur, optimasi arsitektur dan hyperparameter CNN, serta penerapan teknik prapemrosesan dan augmentasi data yang lebih sesuai. Dengan demikian, hasil penelitian ini dapat menjadi dasar untuk pengembangan sistem klasifikasi konsonan bilabial yang memiliki kemampuan pengenalan lebih baik pada penelitian selanjutnya.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, beberapa saran untuk pengembangan selanjutnya adalah sebagai berikut:

1. Penelitian selanjutnya dapat menggunakan jumlah data dan jumlah responden yang lebih besar untuk memperoleh variasi karakteristik suara yang lebih beragam dan meningkatkan kemampuan model dalam mengenali pola pengucapan.
2. Cakupan kelas dapat diperluas dengan menambahkan konsonan atau fonem lain sehingga pengujian tidak hanya terbatas pada konsonan bilabial /p/, /b/, dan /m/.
3. Penelitian selanjutnya dapat menggunakan lebih dari satu skenario atau pengulangan pembagian dataset untuk melihat konsistensi hasil model terhadap perubahan komposisi data.

4. Pengembangan model dapat dilakukan melalui pengujian arsitektur atau konfigurasi yang berbeda, seperti perubahan parameter CNN, penambahan data, atau penggunaan metode klasifikasi lainnya.



DAFTAR PUSTAKA

- [1] K. I. Taher and A. M. Abdulazeez, "Deep Learning Convolutional Neural Network for Speech Recognition: A Review," *Int. J. Sci. Bus.*, vol. 5, no. 3, pp. 1–14, 2021, doi: 10.5281/zenodo.4475361.
- [2] G. T. Waleed and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," *Inf.*, vol. 16, no. 7, p. 518, 2025, doi: 10.3390/info16070518.
- [3] S. Karpagavalli and E. Chandra, "A Review on Sub-word unit Modeling in Automatic Speech Recognition," *IOSR J. VLSI Signal Process.*, vol. 6, no. 6, pp. 77–84, 2016, doi: 10.9790/4200-0606017784.
- [4] C. J. Watkins, "An Integrated Approach to Speech Recognition using Phrase-Based Units," University of East Anglia, 2010.
- [5] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tür, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Commun.*, vol. 32, no. 1–2, 2000.
- [6] T. Sercu, C. Puhersch, B. Kingsbury, and Y. LeCun, "Very Deep Multilingual Convolutional Neural Networks for LVCSR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), 2016, pp. 4955–4959.
- [7] R. Ashlyan, "Syllable Based Speech Recognition," in *Speech Technologies*, InTech, 2011, pp. 263–284.
- [8] R. Chen, A. Ghobakhlou, and A. Narayanan, "Interpreting CNN Models for Musical Instrument Recognition Using Multi-Spectrogram Heatmap Analysis: A Preliminary Study," *Front. Artif. Intell.*, no. December, pp. 1–14, 2024, doi: 10.3389/frai.2024.1499913.
- [9] I. Matthews, "Features for Audio-Visual Speech Recognition," University of East Anglia, 1998.
- [10] M. Algabri, H. Mathkour, M. M. Alsulaiman, and M. A. Bencherif, "Deep Learning-Based Detection of Articulatory Features in Arabic and English

- Speech,” *Sensors (MDPI)*, vol. 21, no. 4, p. 1205, 2021, doi: 10.3390/s21041205.
- [11] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.
- [12] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional Neural Networks for Speech Recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014, doi: 10.1109/TASLP.2014.2339736.
- [13] A. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.
- [14] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, “Recognition of Alzheimer’s Dementia From the Transcriptions of Spontaneous Speech Using fastText and CNN Models,” *Front. Comput. Sci.*, vol. 3, 2021, doi: 10.3389/fcomp.2021.624558.
- [15] B. J. Abbaschian, D. Sierra-sosa, and A. Elmaghraby, “Deep Learning Techniques for Speech Emotion Recognition , from Databases to Models,” *Sensors (MDPI)*, vol. 21, p. 1249, 2021, doi: 10.3390/s21041249.
- [16] A. P. Prathosh, A. G. Ramakrishnan, and T. V Ananthapadmanabha, “Classification of Place-of-Articulation of Stop Consonants Using Temporal Analysis,” in *ISCA*, ISCA, 2015, pp. 2655–2659. doi: 10.21437/Interspeech.2015-397.
- [17] S. S. Calık, A. Kucukmanisa, and Z. H. Kilimci, “An Ensemble-Based Framework for Mispronunciation Detection of Arabic Phonemes,” 2023.
- [18] P. Rashmi and M. P. Singh, “Convolution Neural Networks with Hybrid Feature Extraction Methods for Classification of Voice Sound Signals,” *World J. Adv. Eng. Technol. Sci.*, vol. 08, no. 02, pp. 110–125, 2023, doi: 10.30574/wjaets.2023.8.2.0083.

- [19] D. Jurafsky and J. H. Martin, "Automatic Speech Recognition and Text-to-Speech," Draft of D., 2021, pp. 1–29.
- [20] B. Editors, "Phoneme," Encyclopaedia Britannica. Accessed: Jun. 18, 2026. [Online]. Available: <https://www.britannica.com/topic/phoneme>
- [21] L. G. Duncan, "Language and Reading: The Role of Morpheme and Phoneme Awareness," *Curr. Dev. Disord. Reports*, vol. 5, no. 4, pp. 226–234, 2018, doi: 10.1007/s40474-018-0153-2.
- [22] C. Anderson *et al.*, "3.3 Describing Consonants: Place and Phonation," in *Essentials of Linguistics, 2nd Edition*, eCampusOntario, 2022, ch. 3.3. [Online]. Available: <https://ecampusontario.pressbooks.pub/essentialsoflinguistics2/chapter/3-3-describing-consonants-place-and-phonation/>
- [23] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. Wadsworth, Cengage Learning, 2011.
- [24] T. U. of Sheffield, "Articulatory Phonetics," The University of Sheffield. [Online]. Available: <https://sheffield.ac.uk/linguistics/home/all-about-linguistics/about-website/branches-linguistics/phonetics/what-do-phoneticians-study/articulatory>
- [25] D. Odden, *Introducing Phonology*. Cambridge University Press, 2005.
- [26] M. Yavaş, *Applied English Phonology*, 4th ed. John Wiley & Sons, Inc., 2020.
- [27] A. M. Learning, "Introduction to Feature Extraction," ApX Machine Learning. Accessed: Jun. 19, 2026. [Online]. Available: <https://apxml.com/courses/introduction-to-speech-recognition/chapter-2-processing-audio-signals/introduction-to-feature-extraction>
- [28] A. Project, "Automatic Speech Recognition, How It Works?," ATCO2. Accessed: Jun. 19, 2026. [Online]. Available: <https://www.atco2.org/news/automatic-speech-recognition-how-it-works>
- [29] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *J. Comput.*, vol. 2, no. 3, pp. 138–143,

- 2010, [Online]. Available:
<https://sites.google.com/site/journalofcomputing/>
- [30] I. The MathWorks, “Mel Spectrogram: Extract Mel Spectrogram from Audio,” The MathWorks, Inc. Accessed: Jun. 19, 2026. [Online]. Available:
<https://www.mathworks.com/help/audio/ref/melspectrogramblock.html>
 - [31] T. Bäckström *et al.*, “The Cepstrum, Mel-Cepstrum and Mel-Frequency Cepstral Coefficients (MFCCs),” in *Introduction to Speech Processing*, 2nd ed., Aalto University, 2022, ch. 3.8. doi: 10.5281/zenodo.6821775.
 - [32] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. Acoust.*, vol. ASSP-28, no. 4, pp. 357–366, 1980.
 - [33] A. Amidi and S. Amidi, “CS 230 - Convolutional Neural Networks Cheatsheet,” Stanford University. Accessed: Jun. 22, 2026. [Online]. Available: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks/>
 - [34] H. Sinha, V. Awasthi, and P. K. Ajmera, “Audio Classification Using Braided Convolutional Neural Networks,” *IET Signal Process.*, vol. 14, no. 7, pp. 448–454, 2020, doi: 10.1049/iet-spr.2019.0381.
 - [35] H. Phan, L. Hertel, M. Maass, and A. Mertins, “Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks.”
 - [36] D. Chicco, M. J. Warrens, and G. Jurman, “The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation,” *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/peerj-cs.623.
 - [37] T. O. Hodson, “Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not,” *Geosci. Model Dev.*, vol. 15, no. 2, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.
 - [38] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D*, 19th ed. ALFABETA, 2013.

- [39] S. Hershey *et al.*, “CNN Architectures for Large-Scale Audio Classification,” in *arXiv*, 2017, pp. 1–5.
- [40] N. Vasantha Raju and N. S. Harinarayana, “Online Survey Tools: A Case Study of Google Forms,” in *National Conference on Scientific, Computational & Information Research Trends in Engineering*, 2016, pp. 1–12.
- [41] P. R. Regmi, E. Waithaka, A. Paudyal, P. Simkhada, and E. van Teijlingen, “Guide to the Design and Application of Online Questionnaire Surveys,” *Nepal J. Epidemiol.*, vol. 6, no. 4, pp. 640–644, 2016.
- [42] S. G. Kwak and J. H. Kim, “Central Limit Theorem: The Cornerstone of Modern Statistics,” *Korean J. Anesthesiol.*, vol. 70, no. 2, pp. 144–156, 2017, doi: 10.4097/kjae.2017.70.2.144.
- [43] M. Z. Naser, “A review of machine learning with small and limited data,” *J. Big Data*, vol. 13, 2026, doi: 10.1186/s40537-025-01346-9.
- [44] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, “Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women,” *J. Voice*, vol. 32, no. 5, pp. 644.e1–644.e9, 2019, doi: 10.1016/j.jvoice.2017.08.003.
- [45] E. Pépiot, “Male and Female Speech: A Study of Mean f_0 , f_0 Range, Phonation Type and Speech Rate in Parisian French and American English Speakers,” in *Speech Prosody 7*, N. Campbell, D. Gibbon, and D. Hirst, Eds., International Speech Communication Association / ISCA, 2014, pp. 305–309. doi: 10.21437/SpeechProsody.2014-49.
- [46] R. Cole, Y. Muthusamy, and M. Fanty, “The ISOLET Spoken Letter Database,” 1990.
- [47] B. McFee *et al.*, “librosa : Audio and Music Signal Analysis in Python,” in *Proceedings of the 14th Python in Science Conference, Scipy*, 2015, pp. 18–24.
- [48] librosa D. Team, “librosa.feature.rms,” librosa. Accessed: Jul. 27, 2026. [Online]. Available: <https://librosa.org/doc/main/generated/librosa.feature.rms.html>

- [49] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv*, pp. 1–11, 2018.
- [50] R. Siddalingappa *et al.*, "Adaptive Phoneme State Learning Architecture for Enhanced Speech Recognition Using Backpropagation Neural," *F1000Research*, vol. 15, pp. 1–39, 2026, doi: 10.12688/f1000research.177414.2.
- [51] D. R. Megalmani, B. G. Shailesh, A. R. M. V, S. S. Jeevannavar, and P. K. Ghosh, "Unsegmented Heart Sound Classification Using Hybrid CNN-LSTM Neural Networks," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 713–717. doi: 10.1109/EMBC46164.2021.9629596.
- [52] H. Shi, X. Shi, and S. Dogan, "Speech Inpainting Based on Multi-Layer Long Short-Term Memory Networks," *Futur. Internet*, vol. 2, no. 16, 2024, doi: 10.3390/fi16020063.

