Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Arikel Penelitian

Pemodelan Akustik HMM–GMM untuk Pengenalan Ucapan Kata Darurat Bahasa Indonesia

Candra Bella Vista [a,*], *Endah Septa Sintiya* [b], *Wilda Imama Sabilla* [c], *Adevian Fairuz Pratama* [d]

[a,b,c,d] *Politeknik Negeri Malang, Jl Soekarno Hatta No 9, Malang, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 17 Oktober 2025
Revisi Akhir: 11 Mei 2026
Diterbitkan *Online*: 03 September 2026

KATA KUNCI

Sistem Pengenal Ucapan,
HMM-GMM,
Model Akustik,
Ucapan Darurat

KORESPONDENSI

E-mail: bellavista@polinema.ac.id*

A B S T R A C T

Keamanan dan keselamatan merupakan kebutuhan dasar manusia. Situasi darurat seperti pencurian, kebakaran, maupun ancaman lainnya sering kali memerlukan respons yang cepat untuk mengurangi dampak negatifnya. Seiring berkembangnya teknologi, sistem keamanan berbasis suara menjadi solusi potensial untuk meningkatkan efektivitas deteksi dini dan tanggapan terhadap keadaan darurat. Penelitian ini bertujuan untuk mengembangkan model akustik berbasis *Hidden Markov Model–Gaussian Mixture Model* (HMM–GMM) guna mendeteksi kata-kata darurat dalam Bahasa Indonesia, seperti tolong, jangan, maling, kebakaran, dan kecelakaan. Dataset yang digunakan terdiri atas 1.100 berkas audio yang seimbang antara kelas darurat dan non-darurat. Fitur akustik dari dataset diekstraksi menggunakan *Mel-Frequency Cepstral* Coefficients (MFCC) beserta fitur turunan delta dan delta-delta. Evaluasi dilakukan dengan metode 10-*fold cross-validation* dan model terbaik diujikan dengan data *holdout test* untuk memastikan kemampuan generalisasi model. Hasil penelitian menunjukkan bahwa model HMM–GMM dengan 4 *state* dan 16 komponen Gaussian memberikan performa terbaik dengan akurasi 94,5%, presisi 94,3%, recall 94,5%, dan F1-score 94,4% pada *hold*. Temuan ini membuktikan bahwa model HMM–GMM efektif dalam mengenali ucapan darurat Bahasa Indonesia secara akurat dan efisien.

1. PENDAHULUAN

Keamanan dan keselamatan merupakan kebutuhan mendasar dalam kehidupan. Di berbagai tempat baik di rumah, sekolah, tempat wisata, industri, dan area publik lainnya, keamanan merupakan aspek krusial yang harus diperhatikan. Situasi darurat seperti pencurian, kebakaran, atau ancaman keamanan dan keselamatan lainnya, sering kali memerlukan respons yang cepat untuk mengurangi dampak negatifnya. Berbicara atau berucap adalah cara paling natural untuk menyampaikan maksud, terutama dalam situasi mendesak di mana tindakan lain mungkin terbatas.

Dengan berkembangnya teknologi, sistem keamanan berbasis suara menjadi salah satu inovasi yang menjanjikan untuk meningkatkan efektivitas deteksi dan respons terhadap situasi darurat. Seperti pada penelitian [1] teknologi pengenalan ucapan menjadi solusi dalam mendeteksi situasi darurat bagi orang lanjut usia di rumah mereka sendiri. Penelitian [2] memanfaatkan pengenalan ucapan untuk mengenali kata "tolong" dalam Bahasa Mandarin untuk keamanan penumpang kereta bawah tanah.

Pengenalan ucapan telah menjadi topik penelitian yang berkembang dengan berbagai metode digunakan untuk meningkatkan akurasi dan efisiensi dalam memodelkan aspek akustik. Model akustik merupakan komponen utama dalam sistem pengenalan ucapan yang bertugas untuk menghubungkan sinyal suara dengan representasi linguistik, seperti fonem atau kata [3]. Model ini berfungsi untuk menganalisis dan mengenali pola akustik dari sinyal suara. Salah satu pendekatan yang terbukti efektif untuk memodelkan akustik adalah *Hidden Markov Model (HMM) - Gaussian Mixture Model* (GMM). HMM unggul dalam memodelkan urutan temporal suara [4], sedangkan GMM berperan dalam merepresentasikan distribusi probabilistik fitur suara yang diekstraksi [5]. Kombinasi kedua metode ini memungkinkan sistem untuk mengenali pola akustik

https://doi.org/10.25077/TEKNOSI.v12i2.2026.369-376

dengan lebih baik, terutama dalam kondisi lingkungan yang bervariasi [6].

Namun demikian, perkembangan terbaru menunjukkan dominasi metode berbasis *deep learning*, seperti *Long Short-Term Memory* (LSTM) [1] dan kombinasi MLP-LSTM [7] , yang mampu meningkatkan akurasi pengenalan ucapan secara signifikan. Selain itu, pendekatan *keyword spotting* (KWS) [8] juga dikembangkan untuk mendeteksi kata kunci tertentu dengan lebih akurat, terutama pada sistem berbasis perintah suara. Di sisi lain, penelitian terbaru seperti GTCRN (*Grouped Temporal Convolutional Recurrent Network)* berfokus pada peningkatan kualitas sinyal suara melalui *speech enhancement* dengan kebutuhan komputasi yang rendah [9]. Meskipun metode-metode tersebut menunjukkan performa yang tinggi, terdapat beberapa keterbatasan yang signifikan. Pendekatan *deep learning* umumnya membutuhkan dataset dalam jumlah besar serta sumber daya komputasi yang tinggi agar dapat bekerja secara optimal [10].

Penelitian dalam bidang pengenalan ucapan juga sangat erat kaitannya dengan karakteristik bahasa yang digunakan, karena setiap bahasa memiliki struktur fonetik yang unik [11]. Penelitian terkait pengenalan ucapan dalam Bahasa Indonesia, terutama dalam konteks pengenalan kata-kata darurat masih relatif terbatas. Salah satu penyebabnya adalah ketersediaan sumber data yang minim. Pendekatan HMM-GMM menjadi alternatif yang baik untuk memodelkan aspek akustik pada dataset yang berukuran kecil [12]. Pemodelan akustik merupakan proses awal dan esensial dalam pengenalan ucapan, karena menetapkan hubungan antara informasi akustik dan unit *linguistic* [13]. Performa sistem pengenalan ucapan sangat ditentukan oleh kualitas model akustiknya, sehingga penelitian ini fokus pemodelan akustik untuk mendapatkan pengenal ucapan yang handal. Pengembangan model akustik pada pengenal ucapan berkaitan dengan bahasa, karena setiap bahasa memiliki struktur fonetik yang unik[11]. Bahasa Indonesia tergolong sebagai bahasa dengan sumber daya terbatas (*low-resource language*), sehingga dataset yang tersedia masih jauh lebih sedikit dibandingkan Bahasa Inggris dan Mandarin, bahkan belum ada dataset ucapan yang spesifik terkait kata-kata darurat dalam Bahasa Indonesia. Faktanya penggunaan model akustik yang spesifik terhadap domain tertentu menghasilkan tingkat akurasi yang semakin baik dalam pengenal ucapan [14].

HMM-GMM menjadi alternatif untuk memodelkan akustik pada dataset yang berukuran kecil [12]. Penelitian [15] menunjukkan efektivitas model *hybrid* dalam meningkatkan performa sistem pengenalan ucapan. HMM-GMM pada penelitian [16] mendeteksi teriakan saat pencarian korban bencana. Menggabungkan GMM dan HMM meningkatkan akurasi dibandingkan baseline sistem yang hanya menggunakan HMM [6] [16]. GMM berperan dalam merepresentasikan distribusi probabilistik fitur suara yang diekstraksi dengan membedakan kuat dan lemahnya bunyi yang dikenali [5].

Lebih lanjut, sebagian besar penelitian yang ada berfokus pada domain umum seperti pembelajaran bahasa atau perintah suara, sehingga belum secara spesifik mengkaji pengenalan ucapan dalam konteks darurat. Padahal, ucapan darurat memiliki karakteristik tersendiri, baik dari sisi fonetik maupun konteks penggunaan, yang memerlukan pendekatan khusus. Selain itu, keterbatasan dataset ucapan darurat dalam Bahasa Indonesia juga menjadi tantangan dalam pengembangan sistem pengenalan ucapan.

Penelitian [1] mendeteksi situasi darurat bagi lansia di rumah. Metode LSTM digunakan mengenali bunyi-bunyi seperti gelas pecah, dentuman, dan ucapan “tolong” dalam Bahasa Korea. Nilai presisi dan *recall* sebesar 78% dan 90%. Penelitian [2] mengenali kata “tolong” dalam Bahasa Mandarin untuk keamanan penumpang kereta bawah tanah. Hasilnya LSTM lebih baik dibanding HMM-GMM pada data yang telah diproses untuk mengurangi *noise*.

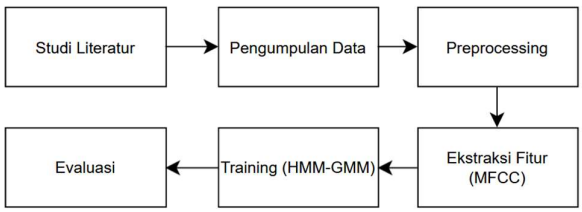
Metode *deep learning* memberikan akurasi lebih baik, namun membutuhkan data latih yang besar. Dalam kondisi *low-resource*, HMM-GMM lebih stabil, cepat, dan lebih mudah diterapkan dibandingkan model *deep learning* [12]. Berdasarkan analisis kesenjangan dari beberapa penelitian terdahulu, Pada penelitian ini akan fokus pada pengembangan model akustik menggunakan metode HMM-GMM yang mendukung kondisi data ucapan darurat Bahasa Indonesia yang terbatas.

Berdasarkan uraian di atas, model akustik berbasis HMM-GMM menawarkan keseimbangan antara kinerja, akurasi, dan efisiensi komputasi. Sehingga penelitian ini ditujukan untuk mengembangkan model akustik berbasis HMM-GMM untuk mengenali kata-kata darurat dalam Bahasa Indonesia. Diharapkan hasil dari penelitian ini dapat berkontribusi dalam penelitian pengenalan ucapan untuk domain ucapan darurat Bahasa Indonesia, penyusunan dataset ucapan darurat Bahasa Indonesia, serta dapat dimanfaatkan lebih lanjut untuk mengembangkan sistem respon situasi darurat yang lebih cerdas dan adaptif.

2. METODE

2.1. Tahapan Penelitian

Penelitian dimulai dari studi literatur sebagai langkah awal. Selanjutnya pengumpulan dataset ucapan baik kata-kata yang darurat maupun non darurat. Kumpulan data rekaman yang terkumpul kemudian melewati serangkaian praproses data sebelum pada akhirnya digunakan pada proses selanjutnya. Setelah data melewati tahapan praproses, data yang telah siap kemudian diekstrak untuk mendapatkan fitur-fitur penting untuk mengembangkan model akustik. Pelatihan model akustik menjadi tahapan selanjutnya dengan melibatkan kombinasi metode HMM dan GMM. Evaluasi model dilakukan dengan menggunakan *confussion matrix* untuk mendapatkan parameter akurasi, *recall*, dan *precision*. Gambar 1 menunjukkan alur penelitian.



Gambar 1. Tahapan Penelitian

https://doi.org/10.25077/TEKNOSI.v12i2.2026.369-376

2.2. Pengumpulan Data

2.2.1. Teknik Pengumpulan Data

Proses pengumpulan data dibagi menjadi 5 tahapan utama, yaitu inisialisasi pengumpulan kata darurat, pengumpulan kata darurat, pengumpulan kata non darurat, tahapan pelabelan, dan pembagian data latih dan data uji.

Tahapan pertama yaitu untuk pengumpulan data kata-kata yang sering diucapkan ketika berada dalam kondisi darurat. Pengumpulan kata-kata tersebut menggunakan media *google form* melalui penyebaran secara acak ke khalayak umum. Berdasarkan jawaban responden kata-kata yang paling sering digunakan adalah "tolong" dan kata lainnya yang muncul adalah "jangan". Menurut Kamus Besar Bahasa Indonesia (KBBI), kata "jangan" merupakan kata keterangan dengan definisi "menyatakan melarang, berarti tidak boleh; hendaknya tidak usah". Oleh karena itu definisi awal kata darurat akan diawali dengan kata "tolong" dan "jangan.

Tahapan kedua yaitu pengumpulan sekaligus penyempurnaan data yang akan digunakan sebagai dataset dalam pelatihan, validasi, dan pengujian sistem. Dalam prosesnya, pembuatan dua kelas utama atau label diperlukan agar model dapat mengenali mana yang merupakan ucapan darurat dan bukan darurat. Penyempurnaan data ucapan darurat dimulai melalui proses perekaman secara manual, khususnya pada ucapan "tolong" dan "jangan". Tiga puluh pembicara acak yang terdiri dari 15 laki-laki dan 15 perempuan dengan rentang usia yang beragam (5-58 tahun) akan direkam untuk mengucapkan kata "tolong" dan "jangan". Proses perekaman akan menggunakan gawai yang sama yaitu Iphone 11 dengan kondisi pengambilan suara minim environmental noise. Rujukan pengambilan sampel ucapan dari 30 pengguna didapatkan melalui evaluasi hasil dari penelitian yang dilakukan oleh [17] dimana pada penelitian tersebut, empat penutur akan direkam suaranya untuk mendeteksi ucapan yang digunakan sebagai akses keamanan. Selain itu, dalam rangka menyempurnakan sistem klasifikasi, diperlukan penambahan ucapan lainnya selain dua ucapan tersebut sehingga pengambilan ucapan lainnya didapatkan melalui dataset bernama "*Indonesian Words Audioset"* yang dibuat oleh [18]. Dataset tersebut merupakan koleksi ucapan kata berbahasa Indonesia yang didapatkan dari 70 orang acak. Berdasarkan pengumpulan data dan definisi dari KBBI tersebut, dirumuskanlah lima kata yang akan digunakan sebagai kata ucapan darurat untuk klasifikasi kondisi darurat, yaitu kata "tolong", "jangan", "maling", "kecelakaan", dan "kebakaran". Data yang terkumpul sebanyak 550 rekaman ucapan darurat.

Data ucapan non darurat didapatkan melalui rekaman suara podcast yang didapatkan dari video Youtube. Pengambilan melalui rekaman suara pada podcast dipilih melalui berbagai pertimbangan, diantaranya akses rekaman yang tersedia secara publik, kemudian konten podcast merupakan pembicaraan natural dan terarah, dan terakhir podcast merupakan platform popular dan banyak didengarkan [19]. Video podcast yang dijadikan sumber data non darurat merupakan podcast yang formal dan intelektual, sehingga ucapan yang terlontar akan minim dengan ucapan non-formal seperti tertawa, terbahak, dan lain sebagainya. Data yang terkumpul sebanyak 550 rekaman non darurat.

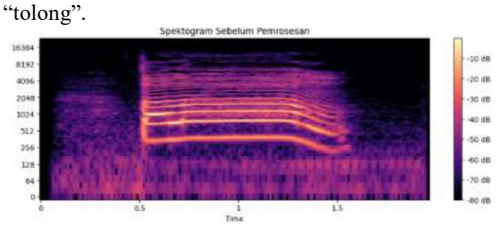
Sebelum data ucapan kata bukan darurat diproses lebih lanjut, dilakukan beberapa tahapan awal. Proses ini mencakup pemotongan rekaman menjadi per kata (trim perkata), diikuti dengan eliminasi rekaman yang dianggap kurang jelas atau tidak sesuai dengan dataset yang diinginkan. Tahapan pra-pemrosesan ini dilakukan untuk memastikan bahwa kualitas dataset tetap tinggi. Jika rekaman yang kurang sesuai tetap dimasukkan, hal ini dapat menurunkan tingkat akurasi dataset dan bahkan menyebabkan kesalahan dalam penilaian [19]. Beberapa kriteria rekaman yang akan dihapus, yaitu rekaman yang mengandung kata bahasa asing, suara kurang jealas, rekaman tanpa kata uang jelas, dan rekaman yang mengandung kata darurat.

2.2.2. Pelabelan

Rekaman hasil pengumpulan data disimpan dalam Google Drive dan diberi label berdasarkan kategori tertentu dengan mengadaptasi metode LJSpeech. Dalam proses ini, setiap rekaman dikelompokkan sesuai dengan kategori masing-masing untuk mempermudah pengelolaan serta pemanfaatan dataset dalam tahap pembelajaran model [19]. Berdasarkan penjelasan tersebut, nantinya struktur direktori penyimpanan dataset di Google Drive akan terstruktur sebagai berikut:

1. dataset/darurat/*.wav
2. dataset/tidak_darurat/*.wav

Dari struktur tersebut, setiap folder berfungsi sebagai tempat penyimpanan rekaman sesuai dengan kategorinya masing-masing. Dengan pendekatan ini, dataset dapat diorganisir secara sistematis sehingga lebih siap digunakan dalam tahap pemrosesan berikutnya. Gambar 2 menunjukkan contoh spektogram ucapan kata "tolong".



Gambar 2 Spektogram dari ucapan "tolong" sebelum praproses

2.2.3. Pembagian data latih dan data uji

Dataset yang terkumpul berjumlah 1100 rekaman (format .m4a). Dari 1100 file tersebut, 550 diantaranya merupakan rekaman ucapan darurat dan 550-nya lagi merupakan rekaman ucapan bukan darurat. Rencana pengumpulan 1100 rekaman ditujukan untuk memproporsionalkan pembagian dataset. Sebanyak 1000 rekaman atau 500 kelas darurat dan 500 kelas bukan darurat, akan digunakan sebagai *data training* dan *validation*. Kemudian 100 sisanya atau 50 kelas darurat dan 50 kelas bukan darurat akan digunakan sebagai data uji independen, atau disebut juga sebagai *holdout test set* (*external validation*). Data uji independen sendiri merupakan sebuah kumpulan data yang berada di luar data *training* dan *validation* dalam sebuah kerangka *cross validation*. Data ini diperlukan untuk membantu dalam pengujian lanjutan terhadap model yang telah dikembangkan [20]. Dalam jurnal [20], dijelaskan bahwasannya proses pengujian generalisasi model yang valid wajib menggunakan *holdout test set* ini, sebab

kinerja model akan lebih terlihat jika diujikan ke data yang tidak pernah dia lihat sama sekali.

Berdasarkan konsep tersebut, penulis akan membagi proporsi 1000 data untuk validasi struktur dengan *K-Fold Cross Validation* dan pembangunan model, kemudian 100 data lagi untuk proses validasi generalisasi model menggunakan *K- Fold Cross Validation - holdout test set*. Sehingga ekspektasi kinerja model akan bagus, mengingat tahapan pembangunan melalui proses validasi struktur dan validasi kinerja generalisasi model.

2.3. Preprocessing

Data yang didapatkan akan diolah terlebih dahulu dengan beberapa tahapan seperti yang ditunjukkan pada Gambar 3.

Gambar 3 Tahapan *Preprocessing*

Beberapa proses dilakukan di tahapan *preprocessing* antara lain:

1. Konversi format audio dari .m4a ke .wav

Preprocessing data ucapan dilakukan dengan mengubah data *voice content* (speech) menjadi format .wav, format WAV dipilih karena merupakan standar berkas audio yang dikembangkan oleh Microsoft dan IBM, serta memiliki kemampuan untuk menyimpan audio baik dalam bentuk terkompresi maupun tidak terkompresi [21]. Format ini umum digunakan untuk menjaga kualitas suara yang dihasilkan oleh berbagai perangkat, sehingga memastikan kejelasan dan ketepatan audio tetap terjaga.

2. Konversi audio stereo menjadi mono

Tahapan selanjutnya adalah konversi audio data yang semula berupa audio stereo (*two channels*) diubah menjadi mono (*one channel*). Hal ini dilakukan untuk menyederhanakan data. Sebagaimana disebutkan oleh [22], data audio stereo (dua saluran) diubah menjadi mono (satu saluran) untuk menyederhanakan data guna mempercepat komputasi pada tahapan selanjutnya.

3. Resampling ke frekuensi 16 kHz

Proses selanjutnya adalah melakukan pengaturan suara ke frekuensi 16 kHz dilakukan untuk memastikan data audio sesuai dengan standar telekomunikasi yang umum digunakan, yaitu 16.000 sampel per detik. Sejalan dengan penelitian yang dilakukan oleh [17], frekuensi sampel standar dalam audio telekomunikasi berada pada 16 kHz, di mana proses pengumpulan data mencakup proses merubah frekuensi sampel pada format data suara menjadi 16 kHz untuk menyesuaikan suara. Oleh karena itu, file dalam format WAV diubah ke frekuensi sampel 16 kHz untuk menjaga kualitas dan konsistensi data yang akan diproses lebih lanjut.

4. Trimming silence untuk menghilangkan bagian hening

Proses *truncate silence* dilakukan untuk mencegah kemungkinan adanya informasi tidak relevan dari data audio dengan menghilangkan bagian berupa keheningan. Langkah ini bertujuan meningkatkan efisiensi dan akurasi analisis data dengan hanya mempertahankan informasi yang relevan. Seperti dijelaskan dalam penelitian, setelah data suara diformat menjadi WAV, frekuensi sampel diubah menjadi 16 kHz kemudian menghilangkan jeda atau keheningan yang tidak diperlukan [17]. Proses ini akan berulang hingga data audio tidak memiliki jeda atau keheningan yang terlalu panjang.

5. Penyimpanan hasil preprocessing

Tahapan terakhir dari proses pengumpulan data adalah menyimpan data yang telah diproses ke dalam *Google Drive*. Seperti yang dijelaskan dalam penelitian [23], *Google Drive* merupakan media penyimpanan daring berbasis komputasi awan yang memungkinkan data tersimpan secara aman dan dapat diakses kapan saja tanpa batasan jarak dan waktu. Hal ini memastikan fleksibilitas dan efisiensi dalam pengelolaan serta penggunaan data selama penelitian berlangsung. Setiap tahapan dilakukan secara berurutan untuk memastikan kualitas data tetap optimal sebelum dilakukan ekstraksi fitur.

2.4. Ekstraksi Fitur

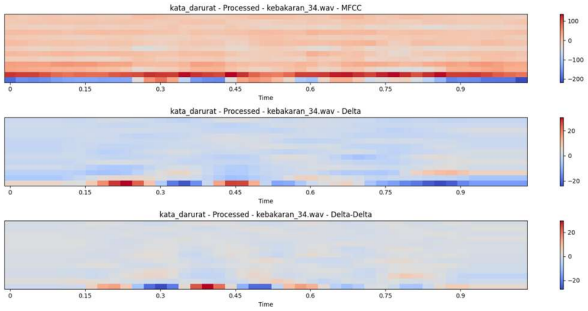
Proses mendapatkan informasi penting dari sinyal suara untuk identifikasi akustik disebut ekstraksi fitur [24]. Pada penelitian ini digunakan metode *Mel Frequency Ceptral Coefficient* (MFCC) yang memodelkan tingkat sensitifitas dari sistem pendengaran manusia menggunakan skala Mel non linier [25]. Setelah dataset melalui tahapan *preprocessing*, data tersebut akan dilakukan ekstraksi fitur dengan memanfaatkan *Mel Frequency Cepstral Coefficients* (MFCC) dari *library* librosa agar data bisa dijadikan input pada model. Input dari parameter MFCC adalah sinyal audio asli (y) yang telah melalui proses augmentasi sebelumnya. Selain itu, terdapat sampling rate (sr) yang menunjukkan jumlah sampel per detik dalam audio input.

Parameter penting lainnya adalah n_mfcc, yang menentukan jumlah koefisien MFCC yang akan dihasilkan sebagai fitur utama. Selain itu, terdapat n_fft, yang secara default bernilai 2048, berfungsi untuk menentukan panjang jendela dalam *Fast Fourier Transform* (FFT). Parameter hop_length, dengan nilai default 512, mengatur langkah pergeseran jendela dalam proses ekstraksi fitur. Selain MFCC, juga dihitung delta MFCC (turunan pertama) dan delta-delta MFCC (turunan kedua) untuk menangkap informasi perubahan temporal dari fitur-fitur tersebut. Hasil ekstraksi fitur berupa array dua dimensi, yang di mana:

1. Arah vertikal merepresentasikan jumlah 33 koefisien,
2. Arah horizontal merepresentasikan jumlah frame, yang diperoleh dari perhitungan berdasarkan durasi audio, sampling rate, dan hop length.

Ketiganya divisualisasikan secara berturut-turut pada Gambar yang masing-masing menampilkan karakteristik waktu-frekuensi berbeda dari audio input. Gambar 4 menunjukkan hasil output dari MCFF, delta, dan delta-delta.

https://doi.org/10.25077/TEKNOSI.v12i2.2026.369-376

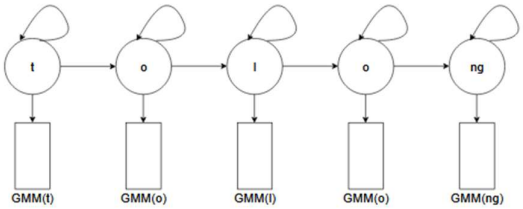


Gambar 4 Visualisasi hasil ekstraksi fitur MFCC, delta, dan delta-delta

2.5. *Model Akustik HMM-GMM*

Model akustik bertugas menghubungkan sinyal suara dengan representasi linguistik, seperti fonem atau kata [3]. Kombinasi HMM-GMM digunakan untuk memodelkan aspek akustik. Untuk mengenali kata "tolong" dalam Bahasa Indonesia, HMM akan memecah menjadi urutan unit suara terkecil (fonem), yaitu:
/t/ → konsonan letup alveolar nirsuara,
/o/ → vokal tengah madya,
/l/ → konsonan lateral alveolar,
/o/ → vokal tengah madya (diulang),
/ng/ → konsonan sengau velar (seperti "ng" dalam "bang"),
Dengan urutan fonem: /t/ → /o/ → /l/ → /o/ → /ng/.

Setiap fonem direpresentasikan sebagai *state* dalam HMM. Setiap *state* memiliki probabilitas transisi ke *state* berikutnya dan probabilitas *self-loop* untuk menangani variasi durasi. Setiap *state* memiliki distribusi probabilitas emisi yang dimodelkan oleh GMM untuk merepresentasikan variasi akustik, yang dalam penelitian ini melibatkan fitur dari hasil ekstraksi MFCC. Gambar 5 menujukkan ilustrasi model HMM-GMM pada kata "tolong".



Gambar 5 Ilustrasi model HMM-GMM pada kata "tolong"

Sebuah HMM didefinisikan dengan tiga parameter utama:

1. Matriks probabilitas transisi antar *state* (A)
$$A = \{a_{ij}\}, \quad a_{ij} = P(q_{t+1} = j \mid q_t = i) \tag{1}$$
dengan q_t adalah *state* pada waktu t.
2. Probabilitas observasi (B)
$$B = \{b_j(o_t)\}, \quad b_j(o_t) = P(o_t \mid q_t = j) \tag{2}$$
yang merepresentasikan distribusi probabilitas dari observasi o_t (fitur akustik) pada *state* j.
3. Probabilitas awal (π)
$$\pi_i = P(q_1 = i) \tag{3}$$
yang menunjukkan peluang memulai urutan dari *state* i.

Kombinasi beberapa Gaussian memungkinkan GMM merepresentasikan distribusi akustik yang bersifat multimodal. Sebagai contoh, fonem /a/ bisa diucapkan dengan variasi panjang, intensitas, dan artikulasi, sehingga tidak cukup dimodelkan dengan satu distribusi Gaussian. Dalam sistem HMM-GMM, HMM berfungsi memodelkan dinamika temporal ucapan, sementara GMM memodelkan distribusi fitur akustik pada setiap *state* HMM

Pendekatan yang digunakan untuk pelatihan data ucapan darurat dan non darurat adalah *Hidden Markov Model* (HMM) yang digabungkan dengan *Gaussian Mixture Model* (GMM). HMM dipilih karena mampu memodelkan sifat sekuensial dari ucapan manusia, sedangkan GMM digunakan untuk merepresentasikan distribusi probabilistik dari vektor fitur akustik (MFCC) pada setiap *state*. Setiap kata darurat dimodelkan sebagai HMM dengan jumlah *state* tertentu, bergantung pada panjang dan kompleksitas akustik kata. Struktur HMM yang digunakan adalah *left-to-right*, di mana transisi *state* hanya diizinkan dari *state* awal menuju *state* berikutnya tanpa mundur ke *state* sebelumnya. Hal ini sesuai dengan sifat ucapan yang berlangsung secara linier dari awal hingga akhir kata.

Jumlah *state* ditentukan berdasarkan kompleksitas kata yang diucapkan. Semakin panjang jumlah suku kata, semakin besar jumlah *state* yang diperlukan untuk memodelkan variasi akustik setiap bagian kata. Setiap *state* tidak hanya direpresentasikan oleh satu *Gaussian*, melainkan oleh campuran beberapa *Gaussian*. Hal ini bertujuan untuk mengakomodasi variasi akustik yang muncul dari perbedaan pembicara, intonasi, maupun kondisi rekaman.

Dalam penelitian ini, model HMM–GMM dikembangkan dengan dua arsitektur model, yaitu HMM dan HMM–GMM. Model HMM digunakan sebagai baseline system dengan distribusi Gaussian tunggal sebagai fungsi emisi pada setiap *state*, sementara HMM–GMM menggunakan campuran *Gaussian* untuk memodelkan distribusi emisi dengan lebih baik. Pada tahap inferensi, setiap rekaman diuji dengan menghitung *log-likelihood* terhadap kedua model, dan kelas ditentukan berdasarkan model dengan skor log-*likelihood* tertinggi.

2.6. *Evaluasi*

Evaluasi memanfaatkan *confussion matrix* untuk mendefinisikan performa model dengan mencatat jumlah prediksi yang benar dan salah [26]. Penelitian ini termasuk klasifikasi biner, yaitu mengenali kata-kata darurat dan non darurat, sehingga *confussion matrix* disusun dalam tabel 2x2 mencakup kategori *True Prositives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN).* TP adalah kasus di mana model secara tepat memprediksi kelas positif. TN adalah kasus di mana model secara tepat memprediksi kelas negatif. FP dikenal sebagai *Type* I *errors*, adalah kasus di mana model salah memprediksi kelas positif. FN dikenal sebagai *Type* II *errors*, terjadi ketika model salah memprediksi kelas negatif.

Berdasarkan nilai-nilai dalam confusion matrix, setelah model selesai dilatih, serangkaian tahap evaluasi dan pengujian akan dilakukan menggunakan *confusion matrix* dan metrik-metrik turunannya seperti *accuracy, precision, recall*, dan *F1-Score* untuk menilai performa model. Model dengan kinerja terbaik kemudian akan diintegrasikan ke dalam aplikasi web untuk pengujian lanjutan menggunakan data di luar dataset pelatihan.

3. HASIL

Dalam penelitian ini dikembangkan dua arsitektur untuk pelatihan model akustik HMM-GMM pada kata-kata darurat bahasa Indonesia, yaitu HMM dan GMM-HMM. Model HMM menggunakan distribusi Gaussian tunggal sebagai fungsi emisi pada setiap *state*, sementara HMM–GMMmenggunakan campuran *Gaussian* untuk memodelkan distribusi emisi dengan lebih baik. Model HMM dibangun dengan struktur transisi *left-to-right*. Untuk menguji pengaruh kompleksitas model terhadap performa, dilakukan pengujian terhadap variasi jumlah *state* HMM, yaitu 3 *state* dan 4 *state*. Jumlah *state* ini dipilih untuk merepresentasikan struktur fonetik kata darurat dalam Bahasa Indonesia yang umumnya terdiri dari dua hingga empat suku kata. Model dengan 3 *state* digunakan untuk menangkap pola temporal dari kata-kata pendek seperti tolong dan maling, sedangkan 4 *state* digunakan untuk kata yang lebih panjang seperti kebakaran dan kecelakaan.

Model HMM–GMM merupakan pengembangan dari HMM dengan menambahkan beberapa komponen Gaussian pada setiap *state* untuk merepresentasikan distribusi fitur akustik yang lebih kompleks. Pengujian dilakukan dengan dua menguji variasi jumlah *state* HMM (3 dan 4) untuk menguji sejauh mana model dengan kompleksitas temporal yang lebih tinggi dapat meningkatkan kinerja deteksi ucapan darurat. Sementara itu, jumlah komponen *Gaussian* pada distribusi emisi divariasikan menjadi 8, 16, dan 32 untuk mengukur *trade-off* antara kapasitas representasi distribusi akustik dan risiko *overfitting.*

Metode pengujian yang digunakan adalah 10-*fold cross-validation*, di mana seluruh dataset dibagi menjadi sepuluh subset (*fold*) dengan proporsi kelas darurat dan non-darurat yang seimbang. Setiap *fold* secara bergantian digunakan sebagai data uji, sementara sembilan *fold* lainnya digunakan sebagai data latih. Pendekatan ini bertujuan untuk memperoleh hasil evaluasi yang lebih objektif dan tidak bias terhadap satu pembagian data tertentu. Tabel 1 merangkum rata-rata hasil terbaik dari masing-masing konfigurasi berdasarkan pengujian 10-fold cross-validation.

Tabel 1 Rata-Rata Hasil Pengujian

Model	Akurasi	Presisi	Recall	F1-score
HMM (3)	90.9	90.8	90.8	90.9
HMM (4)	91.5	91.0	91,1	91.1
HMM–GMM(3x8)	93.3	93.4	92.2	92.3
HMM–GMM(3x16)	94.0	93.1	93.9	93.0
HMM–GMM(3x32)	93,4	93,3	93,4	93,3
HMM–GMM(4x8)	93,9	94,1	93,8	93,2
HMM–GMM(4x16)	**94.4**	**93.5**	**94.3**	**93.4**
HMM–GMM(4x32)	94.1	93.2	93.0	93.1

Rata-rata akurasi terbaik ditunjukkan dengan HMM-GMM konfigurasi 4 *state* dan 16 komponen Gaussian, yaitu mampu mencapai akurasi sebesar 94,4%, nilai presisi (93,5%), dan recall (94,3%). Hal ini menunjukkan keseimbangan yang sangat baik antara kemampuan model dalam mengenali ucapan darurat dan menghindari kesalahan deteksi pada ucapan non-darurat.

Setelah diperoleh konfigurasi model terbaik dari hasil 10-fold cross-validation, yaitu HMM–GMM dengan 4 *state* dan 16 komponen Gaussian, dilakukan pengujian akhir (holdout test) untuk mengevaluasi kemampuan model dalam mengenali data baru yang tidak pernah digunakan selama proses pelatihan maupun validasi. Pengujian dilakukan terhadap 10 berkas audio (50 ucapan darurat dan 50 ucapan non-darurat). Seluruh berkas audio pada tahap ini telah melalui proses prapengolahan (preprocessing) yang sama dengan data latih dan ekstraksi fitur MFCC beserta fitur turunan delta dan delta-delta. Tabel 2 menunjukkan hasil uji cross validation dan hold out test.

Tabel 2. Hasil uji cross validation dan holdout test

Tahap Pengujian	Akurasi (%)	Presisi (%)	Recall (%)	F1-score (%)
10-Fold Cross Validation	94,4	93,5	94,3	93,4
Holdout Test	94,5	94,3	94,5	94,4

4. PEMBAHASAN

Dari hasil pada Tabel 1, terlihat bahwa model HMM–GMM secara konsisten memberikan performa lebih baik dibandingkan HMM pada semua metrik evaluasi. Rata-rata peningkatan akurasi mencapai +2,5% hingga +3,0%, dengan peningkatan serupa pada nilai presisi, recall, dan F1-score. Peningkatan ini menunjukkan bahwa penggunaan beberapa komponen Gaussian pada setiap *state* secara signifikan memperbaiki kemampuan model dalam menangkap variasi akustik antarpenutur dan karakteristik fonetik antar kata.

Model HMM, meskipun memiliki struktur yang lebih sederhana, menunjukkan performa yang cukup stabil dengan akurasi di kisaran 91%. Namun, keterbatasannya muncul karena hanya mampu memodelkan distribusi tunggal, sehingga tidak cukup representatif untuk kondisi ucapan yang bervariasi seperti perbedaan intonasi, tekanan, atau lingkungan perekaman. Sebaliknya, model HMM–GMM mampu mengaproksimasi distribusi suara yang lebih kompleks melalui kombinasi beberapa Gaussian. Setiap *state* tidak hanya menggambarkan satu bentuk distribusi akustik, tetapi juga variasi dari beberapa penutur dan konteks pengucapan. Hal ini menyebabkan model menjadi lebih adaptif dan robust dalam mengenali ucapan darurat pada kondisi nyata.

Peningkatan jumlah *state* dari 3 menjadi 4 menghasilkan perbaikan kinerja yang konsisten untuk kedua jenis model. Model dengan 4 *state* memiliki akurasi rata-rata sekitar 0,5–0,7% lebih tinggi dibanding 3 *state*. Hal ini dikarenakan jumlah *state* yang lebih banyak memungkinkan model untuk merepresentasikan urutan temporal suara dengan resolusi yang lebih baik, terutama pada kata-kata panjang seperti kebakaran dan kecelakaan yang memiliki lebih banyak transisi antar fonem. Namun, peningkatan jumlah *state* lebih dari 4 tidak memberikan keuntungan signifikan, bahkan dapat menimbulkan penurunan performa akibat meningkatnya jumlah parameter yang harus dilatih. Dengan demikian, 4 *state* dianggap sebagai jumlah optimal untuk pengenalan kata darurat Bahasa Indonesia.

Tren peningkatan performa juga terlihat saat jumlah komponen Gaussian dinaikkan dari 8 ke 16. Pada titik ini, model

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.369-376

memperoleh kemampuan representasi yang optimal terhadap variasi distribusi fitur MFCC. Namun, ketika jumlah komponen dinaikkan lagi menjadi 32, peningkatan performa tidak lagi signifikan bahkan cenderung menurun sedikit. Fenomena ini mengindikasikan bahwa model telah mencapai kapasitas optimal pada 16 komponen Gaussian, sementara penambahan komponen berikutnya justru meningkatkan risiko overfitting dan memperpanjang waktu pelatihan tanpa memberikan manfaat berarti terhadap akurasi.

Nilai presisi (93,5%) dan recall (94,3%) pada model terbaik (HMM–GMM 4 *state*, 16 komponen) menunjukkan keseimbangan yang baik antara kemampuan mendeteksi sinyal darurat dan menghindari kesalahan klasifikasi. Nilai recall yang sedikit lebih tinggi dari presisi mengindikasikan bahwa sistem lebih sensitif dalam mengenali kondisi darurat, yang penting dalam konteks aplikasi keamanan atau tanggap bencana. Selain itu, F1-score sebesar 93,4% menunjukkan konsistensi kinerja sistem di berbagai fold pengujian. Keseimbangan ini memperlihatkan bahwa model tidak hanya akurat dalam pelatihan, tetapi juga mampu mempertahankan performa pada data uji yang berbeda.

Tabel 2 menunjukkan bahwa model HMM–GMM dengan konfigurasi 4 *state* dan 16 komponen Gaussian mampu mencapai akurasi sebesar 94,5% pada data uji yang benar-benar baru (holdout set). Nilai ini hanya berbeda sedikit dibandingkan dengan hasil rata-rata dari 10-fold cross-validation, yang menunjukkan bahwa model tidak mengalami overfitting dan memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya.

Nilai presisi (94,3%) dan recall (94,5%) juga menunjukkan keseimbangan yang sangat baik antara kemampuan model dalam mengenali ucapan darurat dan menghindari kesalahan deteksi pada ucapan non-darurat. Presisi yang tinggi menunjukkan bahwa sebagian besar prediksi "darurat" memang benar-benar darurat (minim *false positive*), sedangkan recall yang tinggi menunjukkan bahwa hampir seluruh ucapan darurat berhasil dikenali dengan benar (minim *false negative*). Selain itu, nilai F1-score sebesar 94,4% memperkuat bukti bahwa model memiliki performa yang konsisten antara sensitivitas dan ketepatan klasifikasi. Performa ini menunjukkan bahwa model memiliki stabilitas yang tinggi, kesalahan yang rendah, dan kemampuan generalisasi yang baik.

5. KESIMPULAN

Penelitian ini berhasil merancang dan mengimplementasikan model akustik berbasis HMM dan HMM–GMM untuk sistem pengenal ucapan darurat Bahasa Indonesia dengan total 1.100 data suara (550 darurat dan 550 non-darurat). Model HMM menunjukkan performa dasar yang stabil dengan rata-rata akurasi 91,3%, presisi 90,8%, recall 91,0%, dan F1-score 90,9%. Namun, model ini terbatas dalam menangkap variasi akustik antarpenutur karena setiap *state* hanya memuat satu distribusi Gaussian.

Model HMM–GMM memberikan peningkatan performa yang signifikan, dengan hasil terbaik pada konfigurasi 4 *state* dan 16 komponen Gaussian, mencapai akurasi 94,4%, presisi 93,5%, recall 94,3%, dan F1-score 93,4% pada 10-fold cross-validation. Pengujian akhir dengan data holdout menghasilkan performa konsisten, yaitu akurasi 94,5%, presisi 94,3%, recall 94,5%, dan F1-score 94,4%, menandakan bahwa model memiliki kemampuan generalisasi yang baik terhadap data baru. Secara keseluruhan, integrasi Gaussian Mixture Model ke dalam HMM terbukti meningkatkan akurasi deteksi ucapan darurat hingga sekitar +3% dibandingkan HMM, serta menjaga keseimbangan antara sensitivitas dan ketepatan klasifikasi.

Untuk pengembangan penelitian selanjutnya, beberapa saran yang dapat dipertimbangkan adalah penambahan variasi lingkungan rekaman, seperti latar bising (noise) dan jarak mikrofon yang berbeda, agar model lebih robust terhadap kondisi dunia nyata. Integrasi metode berbasis deep learning, seperti DNN-HMM atau CNN-GMM, untuk meningkatkan kemampuan ekstraksi fitur non-linear dari sinyal suara. Pengembangan sistem real-time, dengan optimasi model agar dapat dijalankan secara efisien pada perangkat dengan sumber daya terbatas seperti smartphone atau microcontroller dengan modul suara.

DAFTAR PUSTAKA

[1] J. Kim, K. Min, M. Jung, and S. Chi, "Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition," *Build. Environ.*, vol. 181, p. 107092, Aug. 2020, doi: 10.1016/j.buildenv.2020.107092.

[2] H. Chu *et al.*, "Call For Help Detection In Emergent Situations Using Keyword Spotting And Paralinguistic Analysis," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, New York, NY, USA: ACM, Oct. 2021, pp. 104–111. doi: 10.1145/3461615.3491111.

[3] L. Rabiner and B.-H. Juang, "Fundamentals of Speech Recognition," 1993. doi: 10.1002/ev.1647.

[4] C. Paul and P. Bora, "Digit Speech Recognition u sing Hidden Markov Model Toolkit," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 5, pp. 917–921, Mar. 2020, doi: 10.35940/ijitee.E2540.039520.

[5] S. M, S. Anusuya, and L. K. Narayanan, "Enhancing Automatic Speech Recognition Accuracy Using a Gaussian Mixture Model (GMM)," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5089158.

[6] P. L, P. Kumar, and D. Singh, "Speech Recognition Using Hmm and Combinations: A Review," *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.4488961.

[7] M. Orosoo *et al.*, "Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education," *Alexandria Engineering Journal*, vol. 111, pp. 21–32, Jan. 2025, doi: 10.1016/j.aej.2024.10.065.

[8] E. M. Tan and F. Utaminingrum, "Analisis Kombinasi Speech Enhancement Dan Keyword Spotting Lightweight Untuk Robust Speech Command Recognition Pada Smart Wheelchair," 2026. [Online]. Available: http://j-ptiik.ub.ac.id

[9] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, "GTCRN: A Speech Enhancement Model Requiring

Ultralow Computational Resources," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2024, pp. 971–975. doi: 10.1109/ICASSP48485.2024.10448310.

[10] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization," *Knowl. Based. Syst.*, vol. 277, p. 110851, Oct. 2023, doi: 10.1016/j.knosys.2023.110851.

[11] S. I. Ayuri, N. Nayla, N. Asa, S. H. Berutu, and A. Puteri, "Pentingnya Fonologi dan Peran Fonologi dalam Sistem Bahasa," *Jurnal Ilmiah Kajian Multidisipliner*, vol. 8, no. 12, 2024.

[12] A. Rouhe, A. Virkunen, J. Leinonen, and M. Kurimo, "Low Resource Comparison of Attention-based and Hybrid ASR Exploiting wav2vec 2.0," *Interspeech*, 2022.

[13] S. Bhatt, A. Jain, and A. Dev, "Acoustic Modeling in Speech Recognition A Systematic Review," *nternational Journal of Advanced Computer Science and Application (IJACSA),* , vol. 11, no. 4, 2020.

[14] C. B. Vista, C. H. Satriawan, D. P. Lestari, and D. H. Widyantoro, "Specific Acoustic Models for Spontaneous and Dictated Style in Indonesian Speech Recognition. ," *J. Phys. Conf. Ser.*, 2018.

[15] A. Rista and A. Kadriu, "Automatic Speech Recognition: A Comprehensive Survey," *SEEU Review*, vol. 15, no. 2, pp. 86–112, Dec. 2020, doi: 10.2478/seeur-2020-0019.

[16] F. Alifani, T. W. Purboyo, and C. Setianingsih, "Implementation of Voice Recognition in Disaster Victim Detection Using Hidden Markov Model (HMM) Method," in *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, IEEE, Aug. 2019, pp. 445–450. doi: 10.1109/ISITIA.2019.8937290.

[17] H. Isyanto, W. Ibrahim, R. Samsinar, and W. Sudarwati, "Accurate Speech Recognition AI using Model of Deep Learning for Security Access," *ICEREAM*, 2024.

[18] A. U. J. Dzulqornain, N. N. P. Utami, M. J. Kusuma, D. M. A. Arsana, M. A. Adyan, and P. M. Waliyullah, "Indonesian Words Audioset. ," https://www.kaggle.com/datasets/ahmadulfi/indonesian-words-audio-dataset.

[19] M. Novela and T. Basaruddin, "Dataset Suara dan Teks Berbahasa Indonesia Pada Rekaman Podcast dan Talk show," *JURNAL FASILKOM*, vol. 11, no. 2, pp. 61–66, Aug. 2021, doi: 10.37859/jf.v11i2.2628.

[20] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiol. Artif. Intell.*, 2023.

[21] E. N. Tamatjita and A. W. Mahastama, "Classification of Traditional and Modern Music using NCC and k-NN," in *Proceedings of the 1st International Conference on Intermedia Arts and Creative Technology*, SCITEPRESS - Science and Technology Publications, 2019, pp. 112–117. doi: 10.5220/0008527301120117.

[22] Y. K. Aini, T. B. Santoso, and T. Dutono, "Pemodelan CNN untuk deteksi emosi berbasis speech Bahasa Indonesia," *Jurnal Komputer Terapan*, vol. 1, no. 7, 2021.

[23] Z. Salsabila and A. Syarif, "Pemanfaatan media Google Drive dalam pengelolaan dokumen elektronik Komisi Aparatur Sipil Negara," *Serasi: Jurnal Sekretari dan Administrasi*, vol. 2, no. 20, 2022.

[24] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition*. Upper Saddle River, New Jersey: Prentice-Hall, 2008.

[25] S. Santoso, R. Hartayu, C. Anam, and Dimas Abdul Aziz, "Simulasi Simulasi Ekstraksi Fitur Suara menggunakan Mel-Frequency Cepstrum Coefficient," *Jurnal Sains dan Informatika*, vol. 8, no. 1, pp. 80–87, Jul. 2022, doi: 10.34128/jsi.v8i1.357.

[26] N. Manouchehri and N. Bouguila, "Human Activity Recognition with an HMM-Based Generative Model," *Sensors*, vol. 23, no. 3, p. 1390, Jan. 2023, doi: 10.3390/s23031390.

BIODATA PENULIS

Candra Bella Vista, S.Kom., M.T. merupakan dosen pada Program Studi Sarjana Terapan Teknik Informatika, Politeknik Negeri Malang yang memiliki minat penelitian dan publikasi ilmiah, dengan fokus pada pengolahan data, speech recognition, dan natural language processing.

Endah Septa Sintiya, S.Pd., M.Kom merupakan dosen pada Program Studi Sarjana Terapan Teknik Informatika, Politeknik Negeri Malang yang memiliki minat besar pada bidang Data Mining, Machine Learning, dan sistem pendukung keputusan.

Wilda Imama Sabilla, S.Kom., M.Kom. merupakan dosen pada Program Studi Sarjana Terapan Sistem Informasi Bisnis, Politeknik Negeri Malang yang memiliki minat pada bidang *computational intelligence*, *machine learning*, dan pemrosesan citra digital.

Adevian Fairuz Pratama, S.Kom., M.T. merupakan dosen pada Program Studi Sarjana Terapan Teknik Informatika, Politeknik Negeri Malang yang memiliki minat pada bidang sistem informasi, sistem pendukung keputusan, dan pemrosesan citra.