JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

Review: Metode-Metode Ekstraksi Ciri dan Klasifikasi Identifikasi Pembicara

Faisal Dharma Adhinata*, Nur Ghaniaviyanto Ramadhan

Rekayasa Perangkat Lunak, Institut Teknologi Telkom Purwokerto, Indonesia
Email: 1,*faisal@ittelkom-pwt.ac.id, 2ghani@ittelkom-pwt.ac.id
Email Penulis Korespondensi: faisal@ittelkom-pwt.ac.id

Abstrak−Pengenalan identitas seseorang masih sering menggunakan ID Card (KTP, SIM, paspor, dsb). Cara ini mempunyai kelemahan karena ID Card mudah rusak dan hilang. Sistem pengenalan biometric memberi solusi dengan menggunakan bagian tubuh manusia sebagai pengenalan identitas. Suara merupakan informasi biometric yang mudah didapat. Pengenalan pola suara digunakan untuk proses identifikasi pembicara untuk memperoleh identitas seseorang yang berbicara. Pada paper ini mereview beberapa metode ektraksi ciri dan klasifikasi yang sering digunakan dalam identifikasi pembicara. Pemilihan metode ekstraksi ciri dan klasifikasi berfungsi dalam komputasi serta tingkat akurasi sistem identifikasi pembicara. Berdasarkan survey dataset yang diaplikasikan dengan metode ekstraksi ciri, metode Mel Frequency Cepstral Coefficients (MFCC) memiliki akurasi yang tinggi meskipun dengan input suara ber-noise. Kemudian dalam klasifikasi, metode Gaussian Mixture Model (GMM) paling sering digunakan karena mampu bekerja dalam suara ber-noise. Akhir-akhir ini dikembangkan hybrid classifier yang membuat nilai akurasi semakin meningkat.

Kata Kunci: Identifikasi Pembicara; MFCC; GMM; Hybrid Classifier

Abstract− Identifying a person's identity still often uses an ID card (KTP, SIM, passport, etc.). This method has a weakness because the ID Card is easily damaged and lost. Biometric recognition systems provide a solution by using human body parts as identity recognition. Sounds are readily available biometric information. Voice pattern recognition is used for the speaker identification process to obtain the identity of someone speaking. This paper reviews several feature extraction and classification methods that are often used in speaker identification. The selection of feature extraction methods and classification functions in computation and the level of accuracy of the speaker identification system. Based on the survey dataset applied with the feature extraction method, the Mel Frequency Cepstral Coefficients (MFCC) method has high accuracy even with noise input. Then in classification, the Gaussian Mixture Model (GMM) method is most often used because it can work in noise. Recently, a hybrid classifier has been developed, which increases the accuracy value.

Keywords: Speaker Identification; MFCC; GMM; Hybrid Classifier

1. PENDAHULUAN

Metode konvensional yang masih sering digunakan untuk mengenali identitas seseorang biasanya menggunakan ID Card (KTP, SIM, paspor, dsb). Metode pengenalan konvensional memiliki keterbatasan, yaitu mudah rusak dan hilang. Sistem pengenalan *biometric* mampu mengatasi keterbatasan ini karena sistem identifikasi menggunakan bagian tubuh manusia [1]. *Biometric* adalah teknik yang mempelajari fisik atau tingkah laku manusia yang sering digunakan sebagai input pengenalan pola. Karakteristik tingkah laku manusia sering digunakan dalam teknik pengenalan pembicara. Setiap manusia mempunyai tingkah laku yang unik atau berbeda dengan yang lain, misalnya tanda tangan atau suara [2]. Informasi *biometric* yang paling mudah didapat dan sering digunakan dalam kehidupan sehari-hari adalah suara.

Manusia selalu berkomunikasi satu sama lain. Bahasa lisan yang digunakan untuk berkomunikasi memiliki sinyal-sinyal suara yang unik pada masing-masing individu. Dengan demikian, sinyal suara yang diucapkan manusia tidak hanya mencirikan apa yang diucapkan, namun memberikan karakteristik siapa yang berbicara. Pemrosesan suara digunakan untuk mengenali pola suara, terdapat dua pengenalan pola suara yaitu pengenalan pembicara dan pengenalan suara. Pengenalan suara bertujuan untuk mengenali kata atau kalimat yang diucapkan dari pembicara, sedangkan pengenalan pembicara bertujuan untuk mengenali siapa yang berbicara menggunakan kata atau kalimat tersebut. Salah satu bagian dari pengenalan pembicara adalah identifikasi pembicara [3].

Identifikasi pembicara mengidentifikasi suara tak dikenal dengan mencocokan suara dalam database yang menghasilkan identitas pengucap suara [3]. Sebuah sistem identifikasi pembicara dikatakan baik apabila dapat melakukan ektraksi ciri yang menjadi ciri khas seseorang, kemudian memproses fitur-fitur suara untuk diklasifikasikan ke dalam kelas-kelas tertentu untuk proses pengenalan. Terdapat dua tipe dalam penerapan identifikasi pembicara, yaitu *text dependent* dan *text independent*. *Text dependent* menggunakan input *testing* berupa kata atau frasa yang digunakan dalam data *training* sebelumnya. Sedangkan *text independent* lebih fleksibel karena input *testing* yang digunakan mungkin tidak terdapat dalam data *training*.

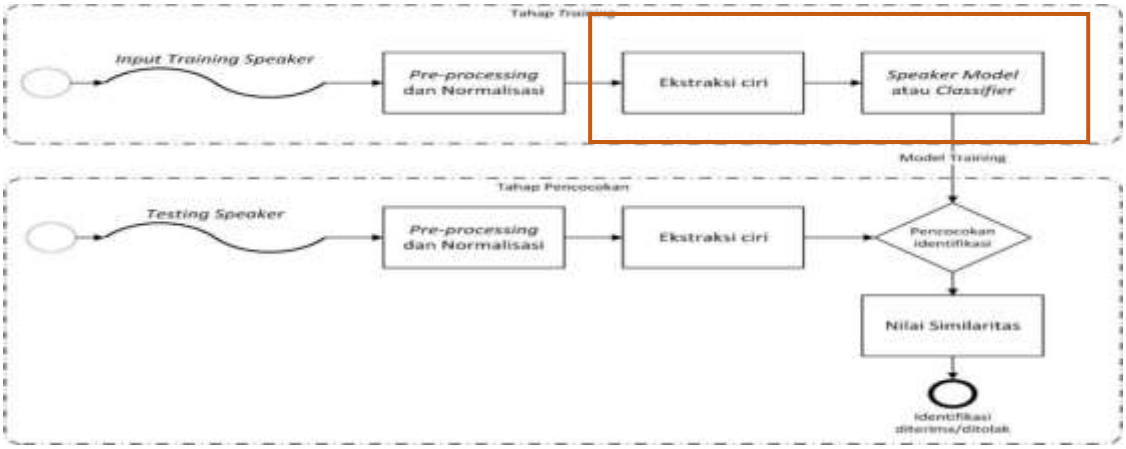
Secara umum terdapat dua tahapan dalam identifikasi pembicara. Tahap pertama adalah *training* dari data pembicara yang sudah dikenali dan terverifikasi. Tahap kedua adalah melakukan *testing* menggunakan data pembicara yang tidak dikenali untuk dicocokkan dengan data *training* yang dilakukan pada tahap sebelumnya.

Tahap training diawali dengan menerima input sinyal suara, kemudian dilakukan *pre-processing* dan normalisasi suara. *Pre-processing* adalah tahap menghilangkan bagian suara yang tidak digunakan dalam tahap ektraksi ciri. Tahap ini biasanya dilakukan dengan menghilangkan *noise* suara dan menghilangkan *silence-frame*

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

(*sample* suara yang tidak memiliki bunyi). Sedangkan normalisasi digunakan untuk menghilangkan variasi *sample* suara. Normalisasi dilakukan dengan menaikkan atau menurunkan amplitude atau volume dari *sample* suara supaya nilai *sample* berada pada rentang tertentu [4].



Gambar 1. Tahapan identifikasi pembicara

Langkah selanjutnya adalah ektraksi ciri yang menghasilkan parameter-parameter sinyal suara. Proses *training* dapat dilakukan *offline* (input suara berasal dari rekaman suara) maupun *online* (input suara menggunakan pengucapan langsung tanpa perekaman). Hasil dari tahap *training* disimpan untuk tahap selanjutnya. Tahap pencocokkan dilakukan dengan mencocokkan sinyal suara yang diperoleh dari ucapan tak dikenal dengan model suara yang disimpan pada tahap training. Tujuan pencocokkan adalah untuk mengidentifikasi siapa yang berbicara. Sama halnya dengan tahapan *training*, ucapan yang belum dikenali ini juga dilakukan *pre-processing* dan normalisasi untuk dimasukkan dalam tahapan ektraksi ciri [4]. Gambar 1 menjelaskan tahapan dalam identifikasi pembicara dengan fokus pembahasan pada tahap ekstraksi ciri dan *speaker model* atau klasifikasi.

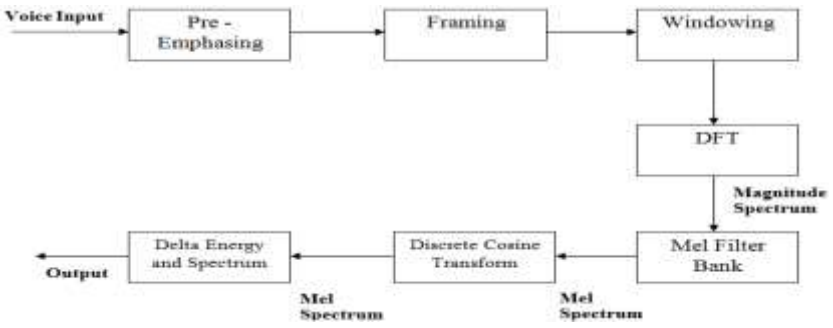
Metode yang dikembangkan dalam identifikasi pembicara biasanya untuk memperbaiki akurasi yang dihasilkan dari tahap ektraksi ciri atau model speaker yang digunakan sebagai klasifikasi. Beberapa metode yang digunakan dalam ektraksi ciri adalah Mel-Frequency Cepstral Coefficient (MFCC) [5][6][7][8][9], Frekuensi Formants [6], Linear Predictive Coding (LPC) [6][7], Linear Predictive Cepstral Coefficients (LPCC) [10], Discrete Wavelet Transform (DWT) [6]. Sedangkan metode yang sering digunakan untuk model speaker atau *classifier* adalah Support Vector Machine (SVM) [11], Hidden Markov Model (HMM) [12], Gaussian Mixture Model (GMM) [5] [8] [9] [13], i-Vector [9] [13], Dynamic Time Wrapping (DTW) [14].

Pada artikel ini memberikan penjelasan terkait identifikasi pembicara serta metode ekstraksi ciri dan klasifikasi yang sering digunakan dalam mengolah sinyal suara untuk mengidentifikasi siapa yang berbicara dari suara tak dikenal. Struktur paper ini pada bagian selanjutnya membahas tentang identifikasi pembicara beserta prosesnya. Pembahasan selanjutnya mengenai metode yang sering digunakan dalam ekstraksi ciri dan klasifikasi beserta hasil penerapan metode dalam beberapa tipe dataset identifikasi pembicara. Di bagian akhir paper berisi kesimpulan rekomendasi penelitian ke depan.

2. METODELOGI PENELITIAN

Suara yang dihasilkan dari seseorang yang berbicara terdapat sinyal suara yang mengandung informasi. Sinyal suara memberikan data informasi penting untuk ektraksi ciri, hal ini karena sinyal suara yang dihasilkan masing-masing orang berbeda-beda. Berbagai algoritma ektraksi ciri dibahas dalam artikel ini dengan tujuan menemukan metode terbaik untuk ektraksi ciri.

2.1 Mel Frequency Cepstral Coefficients (MFCC)



Submitted: 01/12/2021; Accepted: 27/12/2021; Published: 25/01/2022

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

Gambar 2. Tahapan metode MFCC

MFCC merupakan koefisien yang merepresentasikan audio. Metode ini diperkenalkan oleh Davis dan Mermelstein pada tahun 1980-an. Ekstraksi ciri dalam proses ini dimulai dengan mengubah data suara menjadi data citra yang berbentuk spektrum gelombang. Sistem pengenal ucapan saat ini kebanyakan menggunakan MFCC sebagai *feature* karena metode ini dapat diimplementasikan dalam berbagai kondisi [8]. Gambar 2 menunjukkan tahapan metode MFCC. Perkembangan metode MFCC sebagai berikut:

2.1.1 MFCC Dinamis

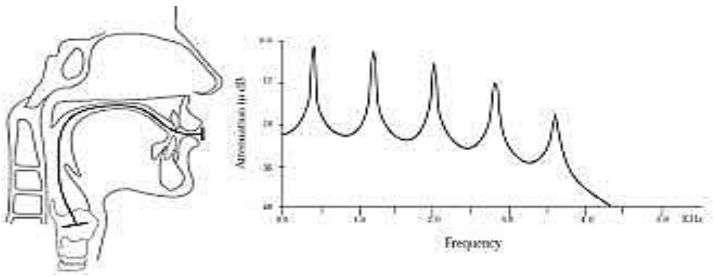
MFCC tradisional hanya bekerja pada fitur *voiceprint* pembicara, tetapi tidak mempertimbangkan karakteristik dinamis dari suara. Peningkatan kinerja dapat dilakukan dengan cara menggabungkan MFCC tradisional dengan *koefisien diferensial* orde pertama yang mencerminkan karakteristik dinamis dari suara [8]. Namun, dengan pendekatan ini dapat meningkatkan besarnya dimensi dari parameter dan kompleksitas komputasional.

2.1.2 Weighted Dynamic MFCC

Weighted Dynamic MFCC dapat memberikan fitur *voiceprint* pembicara dan karakteristik dinamis dari ucapan. Sehingga dengan menggunakan parameter fitur ini, kompleksitas komputasional dari sistem berkurang secara signifikan dengan tetap mempertahankan tingkat akurasi pengenalan yang tinggi [8]. Weighted Dynamic MFCC memiliki kinerja yang lebih baik dalam merefleksikan *voiceprint* dan karakteristik dinamis dari suara [8]. Metode MFCC tidak dapat mengatasi *noise* (gangguan) yang ada pada *background* suara [15].

2.2 Frekuensi Formants

Frekuensi *formants* didefinisikan sebagai puncak dalam *envelope* spektrum pada bunyi suara. Frekuensi *formant* dikeluarkan dari rongga bidang suara manusia, sebagaimana ditunjukkan dalam Gambar 3. Umumnya, suara manusia mempunyai tiga klasifikasi *formant* yaitu F1, F2, dan F3. Sedangkan F0 adalah frekuensi dasar (*pitch*) yang merupakan pengulangan unit terkecil yang mampu mengintegrasikan dua atau lebih periode dari suatu sinyal yang merepresentasikan secara subjektif bagaimana sifat suatu sinyal, khususnya pada sinyal periodik [16]. *Formants* merupakan puncak spektral bunyi yang secara menyeluruh dipengaruhi oleh saluran vokal (*vocal tract*). Oleh karena itu, frekuensi *formant* yang dihasilkan oleh setiap orang berbeda satu dengan yang lainnya karena setiap orang memiliki organ resonansi yang berbeda pula.



Gambar 3. Frekuensi *formant* saat terjadi bunyi

Metode *formants* jika dikombinasikan dengan metode Wavelet Entropy dan Neural Networks lebih baik dibandingkan dengan algoritma klasik yang sering digunakan untuk identifikasi pembicara [6]. *Formants* berdasarkan LPC memiliki kinerja identifikasi yang lebih baik di lingkungan yang bersih dan ber-*noise* [17]. Penambahan *formant* memberikan *threshold* yang lebih ketat untuk tujuan perbandingan dan membantu mengurangi *False Accept Rates* (FAR) [18]. *Formant* dapat digunakan sebagai fitur unik untuk pembicara. *Pitch* mungkin cukup untuk identifikasi pembicara, tetapi biasanya dikombinasikan dengan *formant* untuk pengenalan pembicara yang lebih baik [6] serta meningkatkan kinerja identifikasi pembicara [17]. *Formant* vokal dengan strategi berbasis skor tidak hanya lebih akurat dalam identifikasi tetapi juga lebih akurat [18].

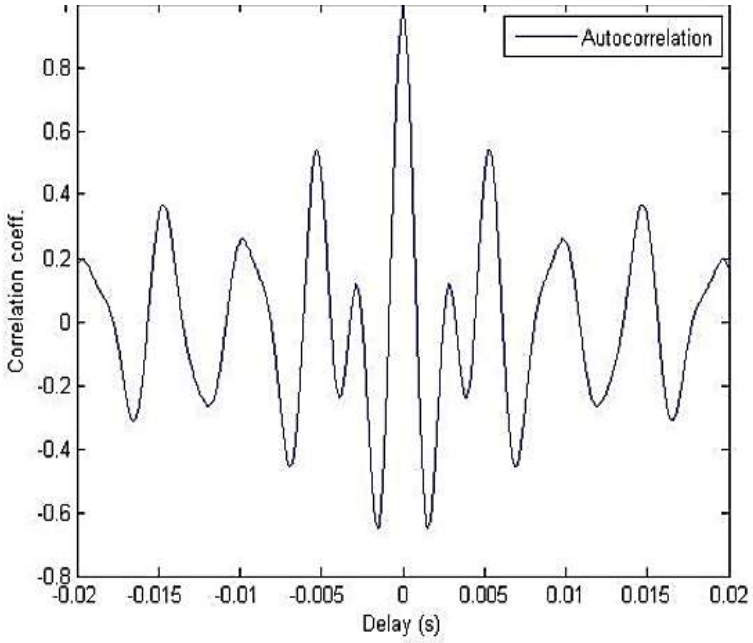
Kekurangan metode ini jika terlalu banyak menggunakan *formant*, dampaknya akurasi tidak selalu meningkat [18]. Penambahan *formant* juga menyebabkan *False Reject Rates* (FRR) [18]. Selain itu, tidak selalu menghasilkan akurasi yang baik karena vokal *formant* untuk vokal yang sama sering tumpang tindih dalam pembicara yang berbeda. Kadang-kadang hanya satu dari tiga nilai *formant* yang tumpang tindih, dan kadang-kadang dua nilai tumpang tindih dengan satu-satunya perbedaan berada di frekuensi ketiga [18].

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

2.3 Linear Predictive Coding (LPC)

LPC bertujuan untuk memisahkan frekuensi *formant* dengan frekuensi dasar (*pitch*) dari suara manusia. LPC juga digunakan untuk mendapatkan spektrum suara [16]. Koefisien korelasi dengan pendekatan LPC ditunjukkan pada Gambar 4.



Gambar 4. Koefisien korelasi dengan pendekatan LPC

Metode LPC dimulai dengan mengasumsikan sinyal suara dihasilkan oleh dengungan pada ujung bibir. Dengan memperkirakan *formants*, LPC menganalisa sinyal suara. Metode ini menghilangkan efek dari *formant* dari sinyal suara, dan memperkirakan intensitas serta frekuensi dengung (*buzz*) yang tersisa. Prosedur yang digunakan untuk menghapus *formant* disebut penyaringan terbalik, dan sinyal yang tersisa disebut residu [19].

Metode LPC *reliable* [18], akurat & kuat [18], kecepatan tinggi [18]. LPC dapat diimplementasikan untuk pemrosesan *real time* dari bentuk gelombang suara [17]. LPC lebih efisien daripada MFCC dan membutuhkan waktu komputasi yang jauh lebih sedikit dalam ekstraksi *formant* vokal [18]. Dapat mengekstraksi frekuensi *formant* secara efektif dengan mencari akar dari polinomial prediksi atau dengan puncak hasil dari *linear prediction spectrum* [17]. Disisi lain, metode ini mengalami penurunan kinerja karena adanya *noise* [20] dan pada bagian yang tidak bersuara dari sinyal suara [20] serta tidak dapat membedakan huruf vokal serupa [18].

2.4 Linear Predictive Cepstral Coefficients (LPCC)

LPCC termasuk kelompok *cepstral coefficient* yang memodelkan suara manusia pada lingkungan yang tidak ber-*noise* untuk menangkap suara vokal. Salah satu modifikasi LPCC adalah LPCC berbobot. Berdasarkan urutan, LPCC berbobot masing-masing dimensi dapat menghasilkan parameter fitur berbobot. Prinsip LPCC berbobot adalah menghasilkan urutan kontribusi rata-rata untuk mengidentifikasi kinerja pengenalan yang berbeda antar komponen [10], dan untuk menemukan pola yang paling jelas dalam peningkatan pengenalan melalui penentuan bobot yang berbeda untuk setiap komponen. Dibandingkan dengan LPCC tradisional, LPCC berbobot mampu meningkatkan akurasi sistem pengenalan pembicara [10]. Namun, koefisien bobot tidak mudah ditentukan dan memerlukan banyak percobaan [10].

2.5 Discrete Wavelet Transform (DWT)

Transformasi Wavelet adalah metode untuk menganalisis sinyal *non-stationer* seperti suara. Metode ini mendistribusikan fitur-fitur spesifik dari sinyal ke dalam pita frekuensi yang berbeda. Dalam versi diskrit, *wavelet* menguraikan sinyal dengan variabel *frame* untuk melakukan analisis multi-resolusi (MRA) dalam bentuk diadik, yang dikenal sebagai transformasi *wavelet* diskrit /*discrete wavelet transform* (DWT) [21].

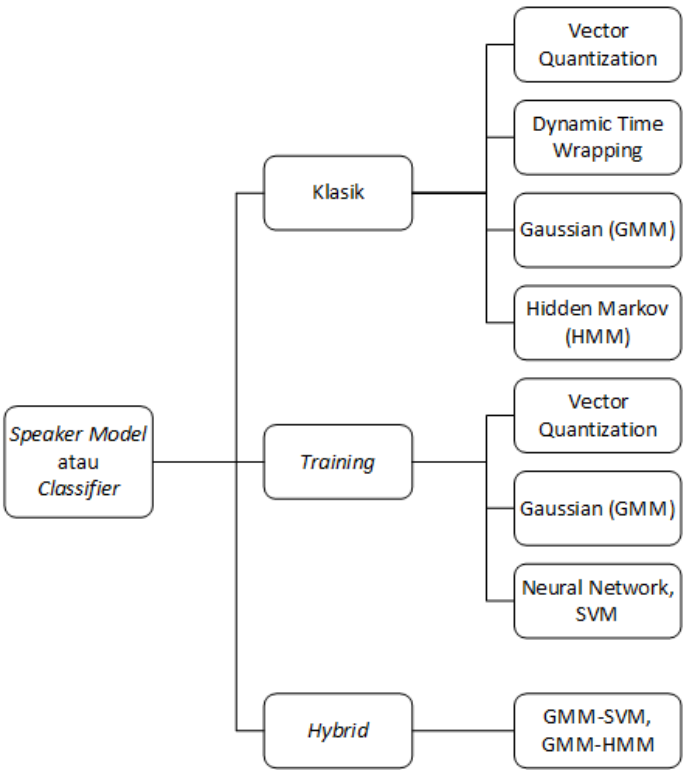
Transformasi wavelet diskrit dimulai dengan menerapkan *hi-pass filtering* yang menghasilkan sinyal frekuensi tinggi dan *lo-pass filtering* yang menghasilkan sinyal frekuensi rendah. Sinyal suara manusia merupakan representasi sinyal 1 dimensi, oleh karena itu filter *hi-pass* dan *lo-pass* masing-masing menggunakan 1 filter. Proses transformasi ini dikenal sebagai dekomposisi *wavelet* dan hasil dari dekomposisi *wavelet* disebut koefisien *wavelet*.

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

Penambahan level dekomposisi wavelet dapat meningkatkan akurasi pengenalan sampai ke tingkat tertentu. Apabila proses dekomposisi wavelet masih terus dilakukan, dapat menyebabkan penurunan tingkat akurasi karena informasi yang dihasilkan semakin sedikit dan menjadi umum [22]. DWT dapat mengurangi pengaruh *noise* pada input sinyal suara karena pada DWT terdapat fitur yang digunakan untuk pengenalan, yaitu koefisien DWT sinyal global yang tidak terpengaruh terhadap *noise* [22].

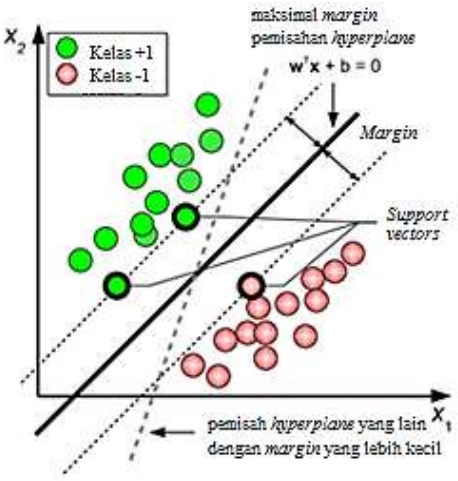
2.6 Metode klasifikasi



Gambar 5. Jenis-jenis metode klasifikasi

Klasifikasi adalah tahap identifikasi pembicara selanjutnya setelah tahap training. Tingkat keberhasilan klasifikasi diukur dengan akurasi identifikasi pembicara dan mempunyai tingkat kesalahan minimal. Terdapat tiga jenis metode klasifikasi, yaitu klasik, *training*, dan *hybrid* sebagaimana ditampilkan dalam Gambar 5.

2.7 Support Vector Machine (SVM)



Gambar 6. Ilustrasi metode SVM

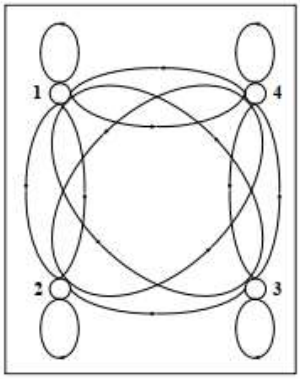
Support Vector Machine (SVM) menggunakan garis pemisah *hyperplane* untuk mengklasifikasikan dataset. Ilustrasi metode SVM sebagaimana ditunjukkan pada Gambar 6. Desain utama dalam SVM adalah kernel yang digunakan dalam fitur *space*. Tujuan desain kernel SVM adalah untuk menemukan metrik yang sesuai di fitur *space* SVM yang cocok untuk proses klasifikasi [23]. SVM merupakan *supervised* algoritma.

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

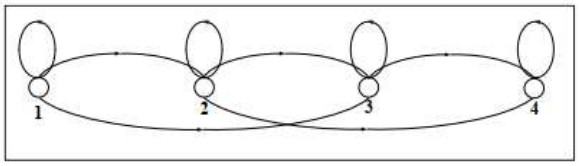
Metode SVM sangat baik ketika digunakan untuk mengklasifikasikan data *binary*. Namun dalam identifikasi pembicara yang mempunyai beberapa parameter hasil ekstraksi ciri, metode ini tidak cocok untuk digunakan [11]. Untuk meningkatkan kinerja SVM, terdapat metode *hybrid* [23]. Metode GMM-SVM meningkatkan kecepatan identifikasi pembicara. Classifier SVM meningkatkan kecepatan, sedangkan GMM membantu proses identifikasi menjadi lebih efisien. Penggunaan *hybrid* GMM-SVM pada *classifier* lebih baik dibanding hanya menggunakan metode SVM atau GMM. Bahkan metode GMM-SVM mencapai akurasi 100% yang hanya menggunakan 3 data training.

2.8 Hidden Markov Model (HMM)



Gambar 7. HMM tipe ergodic

Hidden Markov Model (HMM) merupakan model penerapan rantai Markov yang statusnya tidak teramati secara langsung (tersembunyi), tetapi hanya dapat dikaji melalui suatu himpunan pengamatan pada variable lain. HMM memiliki model *ergodic* sebagaimana ditunjukkan dalam Gambar 7 dan model kiri ke kanan sebagaimana ditunjukkan dalam Gambar 8 [24].



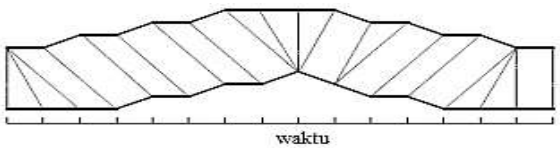
Gambar 8. HMM tipe kiri ke kanan

Model ergodik merupakan tipe model HMM yang setiap statusnya terhubung, sedangkan model kiri ke kanan merupakan tipe HMM yang urutan statusnya terhubung dengan dirinya dan terhubung dari kiri ke kanan. Jenis kiri ke kanan untuk model yang sifatnya berubah dari waktu ke waktu dan tidak dapat kembali ke status sebelumnya seperti pembicaraan (speech). HMM merupakan unsupervised algoritma.

Kelemahan metode HMM adalah komputasi HMM sangat kompleks, memerlukan banyak memori, serta membutuhkan banyak data *training* [12]. Pada artikel [25] menjelaskan metode GMM memiliki komputasi yang lebih efisien dibanding HMM.

2.9 Dynamic Time Warping (DTW)

DTW adalah algoritma pencocokan pola dengan yang menggunakan normalisasi waktu *non-linier* [26]. DTW merupakan algoritma untuk menghitung *optimal warping path* antara dua waktu. Pada Gambar 9 ditunjukkan *warping* antara dua seri waktu. Prinsip DTW adalah membandingkan dua pola dinamis dan mengukur kesamaannya dengan menghitung jarak minimum di antara keduanya [26].



Gambar 9. Warping antara dua seri waktu

Algoritma DTW mampu mendeteksi pola yang sangat lambat maupun sangat cepat dikarenakan pembacaan panjang pendeknya gelombang frekuensi serta mampu menghitung jarak dari dua *vector* data dengan panjang yang berbeda [27]. Ketika metode DTW diimplementasikan pada FPGA dengan menggunakan dataset kata-kata yang diterima oleh microfon dalam lingkungan VHDL, metode ini memperoleh hasil kinerja yang memuaskan. Sistem

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

berbasis DTW meningkatkan rata-rata tingkat pengenalan [14]. Sedangkan kelemahan algoritma DTW adalah sulit untuk membandingkan dua sekuens dari *channel* yang berbeda karena mungkin memiliki fitur yang berbeda [27].

2.10 Gaussian Mixture Model (GMM)

GMM pada umumnya digunakan sebagai fitur dalam sistem *biometric*, seperti fitur spektrum vokal dalam sistem identifikasi pembicara. Parameter GMM didapat dari data *training* yang menggunakan algoritma iteratif Ekspektasi-Maksimalisasi (EM) atau estimasi Maximum A Posteriori (MAP) dari model sebelumnya yang sudah di-*training* [18].

GMM mampu mengidentifikasi suara yang sulit diidentifikasi, misalnya suara yang rusak. Metode ini secara komputasi tidak rumit dan mudah diimplementasikan pada *platform real-time* serta memungkinkan terintegrasi langsung dengan sistem pengenalan suara [28]. Disisi lain, metode GMM mengharuskan pengguna untuk mengatur jumlah *mixture model* yang akan dicoba dan mencocokkan dengan dataset *training* [18]. Penggunaan GMM tidak maksimal dalam dimensi suara tinggi dan komputasi GMM akan meningkat saat jumlah data suara meningkat [18].

2.11 i-Vector

i-Vector adalah model sederhana untuk identifikasi pembicara dengan menghilangkan perbedaan antara ucapan dan saluran variabilitas subruang [29]. Sistem i-vector terbagi menjadi dua bagian utama yaitu *front-end* dan *back-end*. Front-end terdiri dari ekstraksi fitur *cepstral* dan *training* Universal Background Model (UBM), sedangkan *back-end* berisi perhitungan statistik, yaitu training T-matrix, ekstraksi i-vector, pengurangan dimensi dan *scoring*.

i-Vector efektif untuk identifikasi pembicara dengan dimensi-rendah yang disebut ruang variabilitas total [30]. Disisi lain i-vector tidak efektif pada ruang dimensi tinggi (ruang variabilitas total) seperti pada metode GMM kombinasi JFA [10].

3 HASIL DAN PEMBAHASAN

Terdapat beberapa tipe *dataset* yang sering digunakan dalam proses identifikasi pembicara, diantaranya adalah *text-dependent* dan *text-independent*.

3.1 *Text-dependent*

Dataset dengan tipe *text-dependent* menggunakan input *testing* berupa kata atau frasa yang digunakan dalam data *training* sebelumnya.

3.1.1 Metode Ekstraksi Ciri

Beberapa contoh penggunaan dataset untuk perbandingan metode ekstraksi ciri diantaranya menggunakan huruf vokal Arab (A-E-O) dari 80 suara manusia [6] untuk identifikasi pembicara dengan Neural Network sebagai *classifier*. Tabel 1 menunjukkan perbedaan hasil akurasi pada metode untuk ekstraksi ciri. *Text-dependent* bisa juga menggunakan kata, misalnya Hello, Turn On, Turn Off, Up, Down, Good bye [31]. Hasil penggunaan modifikasi MFCC untuk *text-dependent* berupa kata ditunjukkan pada tabel 2. Penerapan *text-dependent* juga dapat dilakukan pada kata berpasangan [32]. Hasil akurasi dataset kata berpasangan ditunjukkan pada Tabel 3.

Tabel 1. Hasil Akurasi vokal-*dependent* [6]

Metode ekstraksi ciri	*Recognition rate* [%]
FWE	90.09
DWT	81.44
MFCC	79.66
LPC	66.63

Tabel 2. Hasil akurasi modifikasi MFCC pada kata [31]

Kata	MFCC	Adaptive MFCC
Hello	96.7 %	98 %
Turn On	96.2 %	96 %
Turn Off	96.5 %	97.2 %
Up	96.4 %	97.6 %
Down	96.3 %	98 %
Good Bye	96 %	97.4 %

Tabel 3. Hasil akurasi pada kata berpasangan [32]

Kata	Metode Ekstraksi Ciri	*Recognition Rate* (%)
HUM-TUM	LPCC	97.0
	MFCC	100

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

Kata	Metode Ekstraksi Ciri	*Recognition Rate* (%)
YANHA-WANHA	LPCC	93.1
	MFCC	99.9
JINA-MARNA	LPCC	99.0
	MFCC	99.9
KHANA-PINA	LPCC	98.0
	MFCC	99.7
DIN-RAAT	LPCC	98.0
	MFCC	99.9

3.1.2 Metode Klasifikasi

Metode klasifikasi dalam proses identifikasi pembicara digunakan untuk mengukur kesamaan antara input ekstraksi dengan model ekstraksi dalam database. Penggunaan metode klasifikasi diantaranya dataset kata pada i-Vector [33], DTW [26], GMM [34], HMM [24], SVM [23], *hybrid* GMM-SVM [23] sebagaimana ditunjukkan dalam Tabel 4.

Tabel 4. Perbandingan hasil akurasi i-Vector, DTW, GMM, HMM, SVM, dan *hybrid* GMM

Classifier	*Accuracy* (%)
GMM-SVM [23]	100
DTW [26]	96
GMM [34]	90.6
HMM [24]	90
SVM [23]	81.25
i-Vector [33]	75.02

3.2 *Text-independent*

Dataset dengan tipe *text-independent* lebih fleksibel karena input *testing* yang digunakan mungkin tidak terdapat dalam data *training*.

3.2.1 Metode Ekstraksi Ciri

Tabel 5. Hasil Akurasi vokal-*independent* [6]

Metode ekstraksi ciri	*Recognition rate* [%]
FWE	82.50
DWT	79.32
MFCC	74.03
LPC	59.45

Beberapa contoh penggunaan dataset untuk perbandingan metode ekstraksi ciri diantaranya menggunakan 80 suara manusia yang direkam dengan *sound card* dalam Bahasa Arab [6]. Hasil akurasi metode ekstraksi ciri ini ditunjukkan pada Tabel 5. Penggunaan database TIMIT dengan 89 suara manusia menggunakan modifikasi MFCC menunjukkan hasil yang sangat signifikan dibanding MFCC tradisional [8]. Hasil akurasi medifikasi MFCC dengan menggunakan GMM sebagai *classifier* disajikan dalam Tabel 6. Perbandingan MFCC dengan LPCC dalam lingkungan ber-*noise* [35] menghasilkan MFCC masih lebih unggul seperti yang disajikan pada tabel 7.

Tabel 6. Hasil Akurasi modifikasi MFCC [8]

GMM	Parameter	*Recognition Rate* (%)
2	MFCC	76.4
	MFCC + ΔMFCC	78.7
	Weighted Dynamic MFCC	79.9
4	MFCC	88.7
	MFCC + ΔMFCC	88.7
	Weighted Dynamic MFCC	91.5
8	MFCC	92.2
	MFCC + ΔMFCC	94.4
	Weighted Dynamic MFCC	95.5

Tabel 7. Perbandingan metode LPCC dan MFCC dalam lingkungan ber-*noise* [35]

SNR	Metode ekstraksi ciri	*Recognition rate* [%]
20 dB	LPCC	73.27
	MFCC	97.03

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

SNR	Metode ekstraksi ciri	*Recognition rate* [%]
15 dB	LPCC	59.41
	MFCC	85.15
10 dB	LPCC	47.52
	MFCC	68.32

3.2.2 Metode Klasifikasi

Beberapa contoh perbandingan metode klasifikasi seperti penggunaan database TIMIT untuk membandingkan metode SVM, GMM, DTW [11]. Percobaan menggunakan 10 dan 50 dataset sebagaimana ditunjukkan pada tabel 8. Penggunaan database TIMIT juga dilakukan untuk membandingkan metode GMM dan i-Vector [13], sebagaimana ditunjukkan pada tabel 9. Kemudian penggunaan dataset tipe *text-independent* juga digunakan untuk membandingkan metode SVM dan HMM [36] sebagaimana ditunjukkan pada table 10.

Tabel 8. Perbandingan akurasi 10 dan 50 dataset pada GMM, DTW, dan SVM [11]

Classifier	Akurasi pada 10 dataset	Akurasi pada 50 dataset
GMM	98%	83%
DTW	92%	80%
SVM	72%	60%

Tabel 9. Perbandingan akurasi GMM dan i-vector [13]

Classifier	Percakapan pendek	Percakapan panjang
GMM	94.7%	97%
i-Vector	85%	95%

Tabel 10. Perbandingan akurasi HMM dan SVM [36]

Classifier	*Recognition rate* (%)
SVM	95%
HMM	30%

4 KESIMPULAN

Paper ini memberikan *review* singkat metode-metode ekstraksi ciri dan klasifikasi untuk proses identifikasi pembicara. Kelebihan beserta kelemahan metode ekstraksi ciri dan klasifikasi telah dijabarkan sehingga dapat menjadi pedoman untuk penelitian selanjutnya. Hasil akurasi metode ekstraksi ciri dan klasifikasi menjadi penentu dalam memberikan rekomendasi penelitian ke depan. Metode MFCC memiliki nilai akurasi paling tinggi dibanding metode ektraksi ciri lainnya. Akhir-akhir ini metode MFCC terus dikembangkan sehingga dapat digunakan untuk lingkungan ber-*noise*, modifikasi MFCC diantaranya terdapat Adaptive MFCC dan Weighted Dynamic MFCC. Kemudian pada metode klasifikasi, kinerja GMM lebih unggul dibanding metode klasifikasi lainnya. Sekarang mulai dikembangkan model *hybrid* sebagai *classifier* sehingga meningkatkan nilai akurasi *matching similarity* data.

REFERENCES

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011, doi: 10.1109/MCAS.2011.941079.
[2] A. H. Rasmussen and D. B. Mikalski, "Speaker Identification," Technical University of Denmark, 2007.
[3] L. Feng, "Speaker Recognition," Technical University of Denmark, 2004.
[4] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017, doi: 10.1016/j.eswa.2017.08.015.
[5] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," *Procedia Computer Science*, vol. 125, pp. 880–887, 2018, doi: 10.1016/j.procs.2017.12.112.
[6] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing Journal*, vol. 27, pp. 231–239, 2015, doi: 10.1016/j.asoc.2014.11.016.
[7] S. V. Chougule and M. S. Chavan, "Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition," *Procedia Computer Science*, vol. 58, pp. 272–279, 2015, doi: 10.1016/j.procs.2015.08.021.
[8] Z. Weng, L. Li, and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," *Proceedings - 2010 International Conference on Anti-Counterfeiting, Security and Identification, 2010 ASID*, pp. 285–288, 2010, doi: 10.1109/ICASID.2010.5551341.
[9] A. Shahab and D. Lestari, "An investigation of Indonesian speaker identification for channel dependent modeling using I-vector," *2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA 2016*, no. October, pp. 151–155, 2017, doi:

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**

Volume 6, Nomor 1, Januari 2022, Page 303-312
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v6i1.3469

10.1109/ICSDA.2016.7919002.
[10] L. Zhu and Q. Yang, "Speaker Recognition System Based on weighted feature parameter," *Physics Procedia*, vol. 25, pp. 1515–1522, 2012, doi: 10.1016/j.phpro.2012.03.270.
[11] L. M. Yee and A. M. Ahmad, "Comparative Study of Speaker Recognition Methods :DTW,GMM and SVM," 2008.
[12] N. Mohan, "GMM-UBM for Text-Dependent Speaker Recognition," *IEEE*, pp. 432–435, 2012, doi: 10.1109/ICALIP.2012.6376656.
[13] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods," *Procedia Computer Science*, vol. 115, pp. 47–54, 2017, doi: 10.1016/j.procs.2017.09.075.
[14] D. Pandey, "Implementation of DTW Algorithm for Voice Recognition using VHDL," pp. 1–4, 2017.
[15] S. B. Magre, R. R. Deshmukh, and P. P. Shrishrimal, "A comparative study on feature extraction techniques in speech recognition," no. June, 2013, doi: 10.1007/s40012-015-0063-y.
[16] M. Subali, M. Andriansyah, and C. Sinambela, "Analisis Frekuensi Dasar dan Frekuensi Formant dari Fonem Huruf Hijaiyah untuk Pengucapan Makhraj dengan Metode DTW," *Prosiding PESAT (Psikologi, Ekonomi, Sastra, Arsitektur &Teknik Sipil)*, vol. 6, pp. 60–73, 2015.
[17] S. Srivastava, P. Nandi, G. Sahoo, and M. Chandra, "Formant Based Linear Prediction Coefficients for Speaker Identification," *International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 685–688, 2014.
[18] N. Almaadeed, A. Aggoun, and A. Amira, "Text-Independent Speaker Identification Using Vowel Formants," *Journal of Signal Processing Systems*, vol. 82, no. 3, pp. 345–356, 2016, doi: 10.1007/s11265-015-1005-5.
[19] P. J. Chaudhary and K. M. Vagadia, "A Review Article on Speaker Recognition with Feature Extraction," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 2, pp. 94–97, 2015.
[20] K. Kaur and N. Jain, "Feature Extraction and Classification for Automatic Speaker Recognition System: A Review," *International Journal of Advances Research in Computer Science and Software Engineering*, vol. 5, no. 1, pp. 1–6, 2015.
[21] J. D. Wu and B. F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," *Expert Systems with Applications*, vol. 36, no. 2 PART 2, pp. 3136–3143, 2009, doi: 10.1016/j.eswa.2008.01.038.
[22] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam, and F. E. Abd El-samie, "A wavelet based approach for speaker identification from degraded speech," *International Journal of Communication Networks and Information Security*, vol. 1, no. 3, pp. 52–58, 2009.
[23] R. Chakroun, L. B. Zouari, M. Frikha, and A. Ben Hamida, "A hybrid system based on GMM-SVM for speaker identification," *International Conference on Intelligent Systems Design and Applications, ISDA*, pp. 654–658, 2016, doi: 10.1109/ISDA.2015.7489195.
[24] D. Handaya, H. Fakhruroja, E. M. I. Hidayat, and C. Machbub, "Comparison of Indonesian speaker recognition using vector quantization and Hidden Markov Model for unclear pronunciation problem," *Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016*, pp. 39–45, 2017, doi: 10.1109/FIT.2016.7857535.
[25] W. C. Chen, C. T. Hsieh, and C. H. Hsu, "Robust speaker identification system based on two-stage vector quantization," *Tamkang Journal of Science and Engineering*, vol. 11, no. 4, pp. 357–366, 2008.
[26] A. H. Mansour, G. Zen, A. Salh, and K. A. Mohammed, "Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms," *International Journal of Computer Applications*, vol. 116, no. 2, pp. 975–8887, 2015, doi: 10.5120/20312-2362.
[27] T. F. FURTUNA, "Dynamic Programming Algorithms in Speech Recognition," *Informatica Economica*, vol. XII, no. March, pp. 94–98, 2008, [Online]. Available: http://econpapers.repec.org/RePEc:aes:infoec:v:xii:y:2008:i:2:p:94-98.
[28] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated Models of Signal and Background with Application to Sneaker Identification in Noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994, doi: 10.1109/89.279273.
[29] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011, doi: 10.1109/TASL.2010.2064307.
[30] N. S. Ibrahim and D. A. Ramli, "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction," *Procedia Computer Science*, vol. 126, pp. 1534–1540, 2018, doi: 10.1016/j.procs.2018.08.126.
[31] H. S. Bae, H. J. Lee, and S. G. Lee, "Voice recognition based on adaptive MFCC and deep learning," *Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications, ICIEA 2016*, pp. 1542–1546, 2016, doi: 10.1109/ICIEA.2016.7603830.
[32] T. Gulzar, A. Singh, and S. Sharma, "Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks," *International Journal of Computer Applications*, vol. 101, no. 12, pp. 22–27, 2014, [Online]. Available: https://pdfs.semanticscholar.org/a9d5/3dce0ef368d9bb0e461ad73a4519319e79a6.pdf.
[33] C. Li, X. Ma, B. Jiang, and X. Li, "Deep Speaker : an End-to-End Neural Speaker Embedding System," *arXiv*, pp. 1–8, 2017.
[34] G. R. Dhinesh, G. R. Jagadeesh, and T. Srikanthan, "A low-complexity speaker-and-word recognition application for resource-constrained devices," *Proceedings - 2011 International Symposium on Electronic System Design, ISED 2011*, pp. 335–340, 2011, doi: 10.1109/ISED.2011.30.
[35] U. Bhattacharjee, "A Comparative Study Of LPCC And MFCC Features For The Recognition Of Assamese Phonemes," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 1, pp. 1–7, 2013.
[36] B. Srinivas and P. Subhashini, "Text Independent Speaker Identification using SVM with MFCC," *Global Journal of Advanced Engineering Technologies*, vol. 5, no. 2, pp. 255–266, 2016.

Submitted: **01/12/2021**; Accepted: **27/12/2021**; Published: **25/01/2022**