Jurnal Ilmiah Ilmu Bahasa & Komunikasi (MURABAH)
Volume. 2 Nomor. 3 Juli 2026
e-ISSN : 3123-738X; p-ISSN : 3123-5603, Hal 53-67
DOI: https://doi.org/10.64365/murabah.v2i3.464
Available online at: *https://jurnal.muaraedukasi.id/index.php/murabah*

Audio Deepfake dan Interseksi Sosiolinguistik: Peran Fitur Prosodi, Jeda, dan Kualitas Ujaran dalam Deteksi Konten Palsu Berbasis AI

Fajar Aditya [1*] ,Genta Prakoso Utomo[2]
[1,2] Universitas Advent, Indonesia
Alamat: Jl. Kolonel Masturi No. 288, Cihanjuang Rahayu, Kecamatan Parongpong, Kabupaten Bandung Barat, Jawa Barat
Email : Fajaraditya@unai.ac.id1, Gentaprakosoutomo@unai.ac.id2
Korespondensi penulis: Fajaraditya@unai.ac.id

Abstract*. The advancement of artificial intelligence (AI) technology has enabled the creation of audio deepfakes synthetic speech that is extremely difficult to distinguish from genuine human voices. This phenomenon poses serious threats across various domains, including disinformation, identity fraud, and public opinion manipulation. This study examines audio deepfakes from an intersectional perspective between technology and sociolinguistics, focusing on the role of three main prosodic features namely prosody (intonation, stress, pitch), pauses (hesitation patterns and silent durations), and speech quality (breathiness, hoarseness, articulation) in the detection process of AI-generated fake content. Employing a critical literature review approach and comparative analysis of state-of-the-art deepfake detection models, this study identifies that prosodic and paralinguistic features often represent the weakest points of AI synthesis due to generative algorithms' inability to replicate natural variations in speech rhythm, unpredictable breath group pauses, and voice quality characteristics dependent on the physiological condition of original speakers. The findings indicate that integrating sociolinguistic features into audio deepfake detection systems can significantly improve identification accuracy compared to approaches relying solely on spectral or phonetic features. This study concludes that the intersection between sociolinguistics and AI technology not only enriches our understanding of speech synthesis vulnerabilities but also opens opportunities for developing more robust detection systems by leveraging knowledge of prosodic variations that are inherently human and difficult for machines to imitate.*

Keywords: *audio deepfake, deepfake detection, prosody, speech pauses, voice quality, sociolinguistics, artificial intelligence.*

Abstrak. Perkembangan teknologi kecerdasan buatan (AI) telah memungkinkan pembuatan audio deepfake sintesis suara tiruan yang sangat sulit dibedakan dari suara asli manusia. Fenomena ini membawa ancaman serius di berbagai bidang, termasuk disinformasi, penipuan identitas, hingga manipulasi opini publik. Penelitian ini mengkaji audio deepfake dari perspektif interseksional antara teknologi dan sosiolinguistik, dengan fokus pada peran tiga fitur prosodik utama yakni prosodi (intonasi, tekanan, nada), jeda (pola hentian dan durasi diam), serta kualitas ujaran (speech quality) seperti nafas, serak, dan artikulasi dalam proses deteksi konten palsu berbasis AI. Menggunakan pendekatan studi pustaka kritis dan analisis komparatif terhadap berbagai model deteksi deepfake mutakhir, penelitian ini mengidentifikasi bahwa fitur-fitru prosodik dan paralinguistik sering kali menjadi titik lemah sintesis AI karena ketidakmampuan algoritma generatif mereplikasi variasi alami dalam ritme ujaran, jeda breath group yang tidak terduga, serta karakteristik kualitas suara yang bergantung pada kondisi fisiologis penutur asli. Hasil kajian menunjukkan bahwa integrasi fitur sosiolinguistik ke dalam sistem deteksi audio deepfake dapat meningkatkan akurasi identifikasi secara signifikan dibandingkan pendekatan yang hanya mengandalkan fitur spektral atau fonetik semata. Penelitian ini menyimpulkan bahwa interseksi antara sosiolinguistik dan teknologi AI tidak hanya memperkaya pemahaman tentang kelemahan sintesis suara, tetapi juga membuka peluang pengembangan sistem deteksi yang lebih robust dengan memanfaatkan pengetahuan tentang variasi prosodik yang secara inheren manusiawi dan sulit ditiru oleh mesin.

Kata kunci: audio deepfake, deteksi deepfake, prosodi, jeda ujaran, kualitas suara, sosiolinguistik, kecerdasan buatan.

1. LATAR BELAKANG

Perkembangan teknologi kecerdasan buatan (artificial intelligence/AI) dalam satu dekade terakhir telah mencapai kemajuan yang luar biasa, khususnya dalam bidang pemrosesan bahasa

Received: Mei 15, 2026; Revised: Mei 27, 2026; Accepted: Juni 15, 2026; Online Available: Juni 20, 2026; Published: Juli 8,2026

alami (natural language processing) dan sintesis suara. Teknologi seperti text-to-speech (TTS) generatif, voice cloning, dan generative adversarial networks (GANs) telah memungkinkan komputer untuk menghasilkan suara yang nyaris tidak dapat dibedakan dari suara manusia asli. Fenomena ini melahirkan apa yang dikenal sebagai audio deepfake rekaman suara palsu yang dihasilkan oleh AI dengan meniru suara seseorang secara sangat realistis.

Audio deepfake pertama kali menarik perhatian publik secara luas pada tahun 2019 ketika muncul rekaman palsu suara seorang CEO perusahaan energi yang memerintahkan transfer dana sebesar 243.000 dolar AS. Sejak saat itu, kasus penyalahgunaan audio deepfake terus bermunculan, mulai dari pembuatan konten pornografi nonsensual, penipuan identitas melalui panggilan telepon, hingga upaya manipulasi politik dengan menyebarkan rekaman palsu tokoh publik. Laporan dari Deeptrace Labs (2020) mencatat adanya peningkatan 900% konten deepfake secara global hanya dalam kurun waktu satu tahun, dengan sebagian besar berupa konten video dan audio palsu.

Ancaman audio deepfake tidak bersifat hipotetis; ia telah menjadi realitas berbahaya yang menggerogoti kepercayaan publik terhadap rekaman audio sebagai alat bukti. Di Indonesia sendiri, kekhawatiran tentang penggunaan deepfake untuk menyebarkan disinformasi dan hoaks politik semakin meningkat menjelang berbagai agenda pemilihan umum. Kemampuan untuk membuat seseorang "mengatakan" sesuatu yang tidak pernah ia ucapkan merupakan ancaman serius bagi demokrasi, stabilitas sosial, dan keamanan individu.

Menghadapi ancaman audio deepfake, berbagai upaya pengembangan sistem deteksi telah dilakukan. Sebagian besar pendekatan deteksi yang ada saat ini berfokus pada analisis spektral dan fitur akustik tingkat rendah, seperti koefisien cepstral frekuensi Mel (Mel-frequency cepstral coefficients/MFCC), fitur spektogram, dan representasi vektor suara berbasis jaringan saraf. Model-model deep learning seperti CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), dan transformer telah dilatih pada dataset besar berisi rekaman asli dan palsu untuk mengklasifikasikan keaslian suara.

Namun, pendekatan ini memiliki sejumlah kelemahan signifikan. Pertama, model deteksi yang dilatih pada dataset tertentu sering gagal menggeneralisasi ke jenis deepfake baru yang belum pernah dilihat sebelumnya sebuah masalah yang disebut sebagai generalization gap. Kedua, teknik adversarial attack dapat dengan mudah mengelabui sistem deteksi dengan menambahkan noise atau modifikasi minimal pada sinyal suara. Ketiga, pendekatan konvensional mengabaikan dimensi sosiolinguistik dan paralinguistik yang merupakan karakteristik inheren dari ujaran manusia alami.

Keterbatasan-keterbatasan ini menunjukkan bahwa pendekatan deteksi audio deepfake yang hanya mengandalkan fitur akustik teknis tidak akan pernah cukup untuk mengimbangi pesatnya perkembangan teknologi sintesis suara. Diperlukan perspektif baru yang melampaui analisis sinyal dan memasukkan pemahaman tentang hakikat komunikasi verbal manusia sebagai fenomena sosial dan biologis yang kompleks.

Salah satu kelemahan paling mendasar dari audio deepfake yang dihasilkan AI dan sekaligus celah kritis yang dapat dieksploitasi untuk deteksi terletak pada ketidakmampuan algoritma generatif mereplikasi sepenuhnya variabilitas alami yang melekat pada ujaran manusia. Ujaran manusia tidak pernah seragam sempurna; ia dipenuhi dengan variasi ritmis, perubahan intonasi yang dipengaruhi emosi, jeda yang tidak terduga untuk mengambil napas atau berpikir, serta karakteristik kualitas suara yang berubah tergantung kondisi fisiologis penutur.

Di sinilah sosiolinguistik cabang linguistik yang mempelajari hubungan antara bahasa dan masyarakat memainkan peran penting. Sosiolinguistik mengajarkan bahwa variasi dalam ujaran tidak bersifat acak, melainkan terstruktur secara sosial dan kontekstual. Prosodi (intonasi, tekanan, nada), jeda (pola hentian dan durasi diam), serta kualitas ujaran (breathiness, serak, kestabilan artikulasi) bukanlah fenomena aksesori, melainkan komponen integral dari makna dan identitas pembicara.

Fitur-fitur prosodik dan paralinguistik ini sering kali menjadi titik lemah sintesis AI karena beberapa alasan. Pertama, jaringan saraf generatif, meskipun sangat kuat, beroperasi berdasarkan pola statistik dan sering menghasilkan output yang too perfect terlalu halus, terlalu konsisten, tanpa variasi mikro yang merupakan ciri khas ujaran manusia alami. Kedua, jeda dan pola pernapasan alami dalam ujaran sangat dipengaruhi oleh konten semantik, kondisi mental pembicara, serta konteks sosial; algoritma saat ini belum mampu memodelkan dependensi kompleks ini secara memadai. Ketiga, kualitas suara seperti breathiness atau serak akibat kelelahan, sakit, atau emosi tertentu hampir tidak pernah tertangkap dengan baik dalam data pelatihan deepfake, sehingga menghasilkan suara yang terdengar "steril".

2. KAJIAN TEORITIS

2.1 Hakikat Audio Deepfake dan Teknologi Sintesis Suara

Audio deepfake merujuk pada rekaman suara palsu yang dihasilkan menggunakan teknologi kecerdasan buatan (AI) untuk meniru suara seseorang secara realistis. Teknologi ini telah mencapai tingkat keandalan yang membuatnya semakin sulit dibedakan dari suara

manusia asli, sehingga menimbulkan berbagai risiko seperti pencurian identitas dan penyebaran disinformasi.

Teknologi di balik audio deepfake mencakup beberapa pendekatan utama:

2.1.1 Text-to-Speech (TTS) Generatif

TTS generatif adalah teknologi yang mengubah teks tertulis menjadi ucapan yang terdengar alami. Model TTS modern seperti Tacotron, WaveNet, dan FastSpeech menggunakan arsitektur neural network deep learning untuk mempelajari pemetaan antara teks dan suara. Kualitas suara yang dihasilkan saat ini sangat realistis sehingga sulit dibedakan dari suara manusia asli, terutama dengan durasi pendek.

2.1.2 Voice Conversion (VC)

Voice conversion adalah teknologi yang mengubah karakteristik suara seseorang menjadi menyerupai suara orang lain sambil mempertahankan konten linguistik yang sama. Berbeda dengan TTS yang membangkitkan suara dari teks, VC bekerja pada sinyal suara yang sudah ada. Teknologi ini sering digunakan untuk membuat deepfake karena hanya memerlukan sedikit sampel suara target.

2.1.3 Generative Adversarial Networks (GANs)

GANs terdiri dari dua jaringan neural yang saling bersaing: generator yang menciptakan konten palsu dan diskriminator yang mencoba membedakan konten asli dari palsu. Proses persaingan ini menghasilkan suara sintetis yang semakin realistis seiring waktu.

Penelitian terbaru menunjukkan bahwa audio deepfake yang dihasilkan oleh berbagai teknologi ini termasuk GAN, diffusion models, TTS, dan voice conversion meninggalkan artefak akustik yang dapat dideteksi pada tingkat frekuensi, keselarasan fase, temporal, dan entropi sinyal. Artefak-artefak inilah yang menjadi dasar bagi upaya deteksi, terutama yang berbasis pada fitur prosodik dan paralinguistik.

2.2 Pendekatan Deteksi Audio Deepfake

Upaya deteksi audio deepfake secara garis besar dapat dikategorikan ke dalam tiga pendekatan utama:

2.2.1 Deteksi Berbasis Fitur Akustik (*Acoustic Feature-based Detection*)

Pendekatan ini menganalisis karakteristik sinyal audio tingkat rendah seperti koefisien cepstral frekuensi Mel (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Constant Q Cepstral Coefficients (CQCC), serta fitur spektogram. Penelitian Jędrasiak (2026) menunjukkan bahwa LFCC, CQCC, dan MFCC memiliki efektivitas diferensiasi tertinggi antara ucapan asli dan sintetis, dengan nilai Δp hingga 0,25 dan rasio PR sekitar 1,6–1,8.

2.2.2 Deteksi Berbasis Deep Learning

Model deep learning seperti Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), dan transformer telah banyak digunakan untuk mengklasifikasikan keaslian suara. Model-model ini dilatih pada dataset besar berisi rekaman asli dan palsu. Namun, pendekatan ini memiliki kelemahan dalam generalisasi terhadap jenis deepfake baru yang belum pernah dilihat sebelumnya (*generalization gap*) serta kerentanan terhadap adversarial attack.

2.2.3 Deteksi Berbasis Watermarking

Watermarking adalah strategi pelabelan konten yang dihasilkan AI pada sumbernya. Teknologi seperti AudioSeal dari Meta dan SynthID dari Google berusaha menanamkan tanda yang tidak terlihat oleh manusia namun dapat dideteksi secara algoritmik. Kelemahan pendekatan ini adalah ketergantungan pada adopsi standar oleh semua generator; jika generator tidak mengimplementasikan watermarking, maka metode ini menjadi tidak berguna.

2.3 Perspektif Sosiolinguistik dalam Deteksi Audio Deepfake

Sosiolinguistik cabang linguistik yang mempelajari hubungan antara bahasa dan masyarakat memberikan kerangka teoretis yang penting untuk memahami mengapa fitur prosodik, jeda, dan kualitas ujaran menjadi celah kritis dalam deteksi audio deepfake. Sebagaimana dinyatakan dalam penelitian Nwosu dkk. (2024), terdapat peluang signifikan untuk konvergensi antara (sosio)linguistik dan deteksi audio deepfake melalui pendekatan interdisipliner.

2.3.1 Prinsip Dasar Sosiolinguistik yang Relevan

Beberapa prinsip dasar sosiolinguistik yang relevan dengan kajian ini meliputi:

a. Variasi sebagai karakteristik inheren ujaran: Ujaran manusia tidak pernah seragam sempurna. Variasi dalam ritme, intonasi, jeda, dan kualitas suara merupakan ciri alami yang dipengaruhi oleh faktor sosial, kontekstual, dan fisiologis.
b. Konteks sosial sebagai penentu bentuk kebahasaan: Cara seseorang berbicara sangat ditentukan oleh konteks sosial siapa yang diajak bicara, di mana, dan dalam situasi apa. Fitur prosodik seperti nada dan intensitas suara berubah tergantung pada hubungan sosial antarpartisipan.
c. Identitas dan gaya berbahasa: Setiap penutur memiliki gaya berbahasa yang unik sebagai cerminan identitas sosialnya. Gaya ini tercermin dalam pola prosodik, kebiasaan jeda, dan karakteristik kualitas suara yang sulit ditiru secara sempurna oleh AI.

2.3.2 Interseksi Sosiolinguistik dan Teknologi Deteksi

Interseksi antara sosiolinguistik dan deteksi audio deepfake terjadi pada titik di mana fitur-fitur bahasa yang secara sosial terstruktur yang secara alami hadir dalam ujaran manusia menjadi "titik lemah" bagi algoritma generatif. Penelitian Nwosu dkk. (2024) mengidentifikasi bahwa fitur-fitur linguistik seperti jeda, napas, pitch, konsonan burst, dan kualitas audio merupakan indikator penting dalam pembedaan ucapan asli dari sintetis.

Kajian ini mengusulkan bahwa deteksi audio deepfake dapat ditingkatkan secara signifikan dengan mengintegrasikan pengetahuan sosiolinguistik tentang:

a. Pola variasi prosodik yang bergantung pada konteks
b. Karakteristik jeda alami dalam tuturan spontan versus tuturan baca
c. Variasi kualitas suara yang dipengaruhi kondisi fisiologis dan emosional

3. METODE PENELITIAN

3.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kualitatif dengan jenis penelitian studi pustaka (*library research*) yang diperkaya dengan analisis komparatif terhadap berbagai model deteksi audio deepfake yang ada. Pendekatan ini dipilih karena penelitian ini bertujuan untuk mengeksplorasi dan mensintesis pengetahuan dari berbagai disiplin ilmu khususnya sosiolinguistik, fonetik akustik, dan teknologi deteksi deepfake guna mengidentifikasi bagaimana fitur prosodi, jeda, dan kualitas ujaran dapat berperan dalam deteksi konten palsu berbasis AI .

Sebagai penelitian interdisipliner, metode ini mengintegrasikan:

a. Kajian teoretis sosiolinguistik tentang variasi prosodik dan paralinguistik dalam ujaran manusia
b. Analisis teknis tentang artefak akustik yang dihasilkan oleh teknologi sintesis suara AI
c. Evaluasi komparatif terhadap berbagai pendekatan deteksi deepfake yang telah dikembangkan

3.2 Desain Penelitian

Desain penelitian ini terdiri atas tiga tahap utama:

Tahap 1: Pengumpulan Literatur Sistematis

a. Pengumpulan literatur dari sumber-sumber akademik terpercaya yang membahas:
b. Teknologi pembangkitan dan deteksi audio deepfake
c. Fitur prosodik dan paralinguistik dalam ujaran manusia
d. Pendekatan deteksi berbasis linguistik dan sosiolinguistik

Tahap 2: Identifikasi dan Klasifikasi Fitur

a. Mengidentifikasi dan mengklasifikasikan fitur-fitur prosodik, jeda, dan kualitas ujaran yang relevan untuk deteksi deepfake berdasarkan temuan penelitian empiris terkini .

Tahap 3: Analisis Komparatif dan Sintesis

a. Membandingkan keefektifan berbagai pendekatan deteksi serta mensistesiskannya ke dalam kerangka deteksi terintegrasi yang menggabungkan wawasan sosiolinguistik.

3.3 Sumber Data dan Literatur

Penelitian ini menggunakan sumber data sekunder yang terdiri atas:

Tabel 3.3 Literatur Ilmiah Utama

Jenis Literatur	Fokus	Periode
Jurnal internasional terindeks (Scopus/WoS)	Deteksi audio deepfake, prosodi forensik	2020-2026
Prosiding konferensi (ICASSP, Interspeech, IEEE ISI)	Metode deteksi mutakhir	2020-2026
Artikel riset (arXiv, preprints)	Penelitian terbaru tentang prosodi dan deepfake	2024-2026

3.4 Prosedur Analisis Data

Prosedur analisis data dalam penelitian ini mengikuti kerangka yang sistematis.

Tabel 3.4 Ekstraksi dan Identifikasi Fitur

Sub-fitur	Deskripsi	Relevansi Deteksi
Pitch / F0	Frekuensi fundamental suara	Ujaran sintetis cenderung memiliki variasi pitch yang terbatas
Jitter	Variasi siklus ke siklus dalam frekuensi pitch	Nilai jitter yang terlalu rendah mengindikasikan suara terlalu sempurna
Shimmer	Variasi amplitudo siklus ke siklus	Smoothing prosodik pada suara sintetis
Intonasi	Pola nada dalam frasa	Deepfake gagal mereplikasi intonasi natural yang bergantung konteks

Sub-fitur	Deskripsi	Relevansi Deteksi
Ritme	Pola durasi dan tekanan	Ritme AI terlalu teratur dibanding ujaran manusia

3.5 Teknik Analisis Data

Teknik analisis data yang digunakan dalam penelitian ini adalah analisis tematik (*thematic analysis*) dan analisis komparatif konstan (*constant comparative analysis*).

3.5.1 Analisis Tematik

Analisis tematik digunakan untuk mengidentifikasi, menganalisis, dan melaporkan pola-pola (tema) dalam literatur tentang:

a. Artefak akustik yang membedakan suara asli dari suara sintetis
b. Kelemahan umum teknologi sintesis AI dalam mereplikasi fitur prosodik
c. Peluang integrasi pengetahuan sosiolinguistik dalam deteksi deepfake

3.5.2 Analisis Komparatif Konstan

Metode ini digunakan untuk membandingkan:

a. Keefektifan berbagai jenis fitur (MFCC, LFCC, CQCC, prosodik, modulasi)
b. Performa berbagai arsitektur model deteksi (*CNN, LSTM, transformer, capsule networks*)

4. HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

4.1.1 Gambaran Umum Temuan Penelitian

Berdasarkan kajian sistematis terhadap berbagai penelitian empiris tentang deteksi audio deepfake berbasis fitur prosodik dan paralinguistik, penelitian ini mengidentifikasi sejumlah temuan kunci yang dikelompokkan ke dalam tiga kategori utama: fitur prosodi, fitur jeda, dan fitur kualitas ujaran. Temuan-temuan ini disintesis dari berbagai sumber yang mencakup penelitian tentang deteksi deepfake menggunakan pendekatan linguistik, pengembangan model machine learning untuk deteksi, serta studi tentang persepsi pendengar manusia terhadap suara sintetis.

4.1.2 Efektivitas Fitur Prosodi dalam Deteksi Deepfake

Penelitian "*Pitch Imperfect*" yang dilakukan oleh tim peneliti (2025) mengembangkan detektor deepfake berbasis enam fitur prosodik klasik dan mencapai kinerja yang sebanding dengan model baseline komunitas dengan akurasi 93% dan Equal Error Rate (EER) sebesar

24,7% pada dataset standar . Temuan paling signifikan dari penelitian ini adalah identifikasi tiga fitur prosodik yang memiliki dampak tertinggi terhadap keputusan model:

a. Jitter: Variasi siklus-ke-siklus dalam frekuensi fundamental (F0)
b. Shimmer: Variasi siklus-ke-siklus dalam amplitudo
c. Mean Fundamental Frequency (F0): Rata-rata frekuensi dasar suara

Keunggulan utama pendekatan berbasis prosodi ini terletak pada robustness-nya terhadap serangan adversarial. Ketika model-model lain diuji dengan serangan L∞ norm (metode yang memodifikasi input secara minimal untuk mengelabui detektor), model tersebut mengalami degradasi akurasi relatif sebesar 99,3%. Sebaliknya, pendekatan berbasis prosodi menunjukkan ketahanan yang jauh lebih baik karena fitur-fitur prosodik mencerminkan properti fundamental dari ujaran manusia yang sulit dimanipulasi tanpa mengubah esensi sinyal secara signifikan.

4.1.3 Konvergensi (Sosio)Linguistik dan Deteksi Deepfake

Penelitian oleh Mallinson, Nwosu, dan kolega (2024) yang dipublikasikan di *Language and Linguistics Compass* secara eksplisit mengartikulasikan perlunya konvergensi antara (sosio)linguistik dan teknologi deteksi deepfake . Mereka mengidentifikasi bahwa pendekatan deteksi yang ada memiliki keterbatasan fundamental dan bahwa ahli (sosio)linguistik dapat berkontribusi secara langsung dalam tiga area utama:

a. *Incorporating expert knowledge*: Pengetahuan ahli tentang fitur-fitur linguistik seperti jeda, napas, pitch, konsonan burst, dan kualitas audio secara keseluruhan telah terbukti meningkatkan deteksi deepfake dengan algoritma AI
b. *Developing techniques for everyday listeners*: Mengembangkan teknik yang dapat digunakan oleh pendengar awam untuk menghindari penipuan
c. *Benevolent applications of generative AI*: Mengembangkan aplikasi positif dari teknologi AI generative

4.2 Pembahasan

4.2.1 Mengapa Fitur Prosodi, Jeda, dan Kualitas Ujaran Efektif untuk Deteksi Deepfake?

Temuan penelitian ini secara konsisten menunjukkan bahwa fitur prosodi, jeda, dan kualitas ujaran memiliki efektivitas tinggi dalam mendeteksi audio deepfake. Efektivitas ini dapat dijelaskan melalui beberapa mekanisme fundamental:

4.2.1.1 Ketidakmampuan AI Mereplikasi Variabilitas Alami

Ucapan manusia secara inheren bersifat variabel. Tidak ada dua ucapan dari orang yang sama yang persis identik dalam hal prosodi setiap ujaran mengandung variasi mikro dalam pitch, durasi, intensitas, dan kualitas vokal. Variasi ini bukanlah "noise" yang tidak bermakna,

melainkan cerminan dari kondisi fisiologis (tingkat kelelahan, kesehatan pita suara), keadaan emosional (kegembiraan, kesedihan, kemarahan), dan konteks sosial pembicara.

Algoritma generatif, terutama yang berbasis deep learning, cenderung memproduksi output yang too perfect terlalu halus, terlalu konsisten karena mereka dioptimalkan untuk meminimalkan fungsi kerugian (*loss function*) yang mendorong prediksi yang "rata-rata" dan stabil. Jitter yang terlalu rendah, shimmer yang terlalu kecil, dan durasi jeda yang terlalu seragam merupakan artefak dari optimasi ini .

4.2.1.2 Kompleksitas Multilevel dari Ujaran Manusia

Ujaran manusia dapat dianalisis pada setidaknya empat level :

a. Level sinyal: Fitur akustik dasar (*spektrum, harmonik, fase*)
b. Level prosodik: Pitch, intonasi, ritme, tekanan
c. Level temporal: Jeda, durasi, kecepatan bicara, pola respons
d. Level paralinguistik: Emosi, sikap, identitas sosial, kesehatan

Model deepfake saat ini, meskipun sangat canggih dalam mereplikasi level sinyal, masih kesulitan dengan level prosodik dan temporal yang lebih tinggi. Kesenjangan inilah yang dieksploitasi oleh pendekatan deteksi berbasis prosodi.

4.2.1.3 Variasi Idiosinkratik sebagai Sidik Jari Vokal

Setiap penutur memiliki karakteristik prosodik yang unik pola intonasi khas, kebiasaan jeda, kecenderungan pitch range tertentu yang berfungsi sebagai semacam "sidik jari vokal". Dalam komunikasi sehari-hari, pendengar menggunakan karakteristik ini (seringkali secara tidak sadar) untuk mengidentifikasi pembicara dan menilai keaslian ujaran.

Sistem deepfake, meskipun dapat meniru timbre dan aksen seseorang dengan cukup baik, hampir selalu gagal mereplikasi karakteristik prosodik idiosinkratik ini secara sempurna karena keterbatasan data pelatihan (hanya beberapa detik hingga menit suara target) dan keterbatasan model dalam menangkap dependensi jangka panjang dalam sinyal prosodik.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

5.1.1 Efektivitas Fitur Prosodi dalam Deteksi Audio Deepfake

Fitur prosodi terutama jitter (*variasi frekuensi fundamental siklus-ke-siklus*), shimmer (*variasi amplitudo siklus-ke-siklus*), dan mean fundamental frequency (F0) terbukti efektif dalam mendeteksi audio deepfake dengan tingkat akurasi mencapai 93% (EER 24,7%) pada dataset standar. Keunggulan utama pendekatan berbasis prosodi adalah ketahanannya terhadap serangan adversarial; ketika model lain mengalami degradasi akurasi hingga 99,3% akibat serangan L∞ norm, model berbasis prosodi menunjukkan robust yang jauh lebih baik karena

fitur-fitur prosodik mencerminkan properti fundamental ujaran manusia yang sulit dimanipulasi tanpa mengubah esensi sinyal secara signifikan.

5.1.2 Peran Fitur Patologis Ujaran dalam Deteksi

Fitur patologis ujaran yang terdiri atas sebelas fitur meliputi 3 jenis jitter, 4 jenis shimmer, Harmonics-to-Noise Ratio (HNR), *Cepstral Harmonics-to-Noise Ratio* (CHNR), *Normalized Noise Energy* (NNE), dan *Glottal-to-Noise Excitation Ratio* (GNE) mencapai kinerja deteksi yang sangat impresif pada dataset ASVspoof 2019 dengan akurasi 95,06%, recall 99,46%, dan F1-score 97,30%. Keberhasilan fitur-fitur ini terletak pada kemampuannya menangkap

kehalusan berlebihan (*over-smoothness*) yang menjadi ciri khas suara sintetis, yaitu kecenderungan algoritma generatif untuk memproduksi suara yang terlalu halus dan konsisten tanpa variasi mikro alami yang melekat pada ujaran manusia.

5.1.3 Signifikansi Fitur Jeda sebagai Biomarker Vokal

Pola jeda dalam ujaran meliputi panjang dan variasi ujaran, tingkat micropause (jeda mikro antar suku kata), tingkat macropause (jeda lebih panjang antar frasa), serta proporsi waktu berbicara versus jeda terbukti mampu membedakan audio asli dari deepfake dengan akurasi sekitar 80%. Jeda napas (*breath pause*) yang secara alami terjadi pada batas intonational phrase, variasi durasi jeda yang bergantung pada konteks, serta karakteristik jeda idiosinkratik setiap penutur merupakan fitur yang sangat sulit direplikasi oleh AI. Sistem sintesis suara cenderung menempatkan jeda pada posisi yang tidak sesuai dengan struktur gramatikal atau kebutuhan napas alami, serta menghasilkan durasi jeda yang terlalu seragam.

5.1.4 Representasi Prosodik untuk Deteksi Ucapan Ekspresif dan Emosional

Sistem ProSDD (*Learning Prosodic Representations for Speech Deepfake Detection*) berhasil mengatasi kelemahan signifikan dari sistem deteksi konvensional yang gagal menangani ucapan yang mengandung variasi emosional seperti tertawa, tangisan, kemarahan, atau sarkasme. Dengan mempelajari variabilitas prosodik dari ucapan asli berdasarkan tiga dimensi (*pitch, voice activity detection, dan energy*) kemudian mengoptimalkannya secara bersama dengan klasifikasi spoofing, ProSDD mencapai pengurangan Equal Error Rate (EER) yang dramatis: dari 39,62% menjadi 7,38% pada ASVspoof 2024 (peningkatan 81,4%), serta pengurangan relatif 50% pada dataset EmoFake dan EmoSpoof-TTS.

5.2 Saran

5.2.1 Saran untuk Peneliti Selanjutnya

a. Pengembangan Dataset Multidimensi: Peneliti selanjutnya disarankan untuk mengembangkan dataset audio deepfake yang tidak hanya bervariasi dari segi teknologi pembangkitan (GAN, diffusion models, TTS, voice conversion), tetapi juga dari segi

variasi prosodik alami termasuk ucapan dalam berbagai emosi, kondisi fisiologis (lelah, sakit, sehat), serta konteks sosial yang berbeda. Dataset semacam ini akan memungkinkan evaluasi yang lebih komprehensif terhadap kemampuan deteksi sistem.

b. Integrasi Fitur Prosodik ke dalam Arsitektur End-to-End: Meskipun fitur prosodik telah terbukti efektif, sebagian besar implementasi masih bersifat sebagai ekstraksi fitur terpisah sebelum diklasifikasikan. Penelitian selanjutnya dapat mengembangkan arsitektur end-to-end yang secara implisit mempelajari representasi prosodik secara hierarkis, mirip dengan pendekatan ProSDD namun dengan mekanisme attention yang lebih canggih.
c. Studi Lintas Bahasa dan Lintas Budaya: Karakteristik prosodik sangat dipengaruhi oleh bahasa dan budaya. Penelitian tentang deteksi deepfake perlu dilakukan secara lintas bahasa untuk mengidentifikasi fitur prosodik yang bersifat universal versus yang bergantung pada bahasa tertentu. Penelitian juga perlu melibatkan lebih banyak partisipan dari latar belakang aksen yang beragam.
d. Pengembangan Metode untuk Pendengar Awam: Sebagian besar penelitian fokus pada pengembangan sistem deteksi otomatis. Penelitian selanjutnya perlu mengembangkan teknik yang dapat digunakan oleh pendengar awam untuk melindungi diri dari penipuan berbasis deepfake, misalnya melalui pelatihan sederhana tentang isyarat prosodik mana yang paling mudah dikenali sebagai indikator kepalsuan.
e. Studi Longitudinal tentang Evolusi Deepfake: Teknologi deepfake terus berkembang. diperlukan studi longitudinal yang melacak bagaimana kualitas prosodik deepfake meningkat dari waktu ke waktu dan fitur mana yang tetap menjadi titik lemah meskipun terjadi peningkatan teknologi.

5.2.2 Saran untuk Pengembang Sistem Deteksi

a. Adopsi Pendekatan Ensemble Fitur: Berdasarkan temuan bahwa berbagai jenis fitur (prosodi, patologis ujaran, biomarker vokal) memiliki keunggulan komplementer, pengembang disarankan untuk mengadopsi pendekatan ensemble yang menggabungkan multiple feature streams. Misalnya, menggabungkan fitur prosodi klasik (jitter, shimmer) dengan fitur patologis ujaran (HNR, NNE, GNE) dan fitur temporal (pola jeda).
b. Implementasi ProSDD untuk Skenario Dunia Nyata: Sistem deteksi yang hanya diuji pada dataset standar sering gagal di dunia nyata karena deepfake dalam praktiknya sering mengandung variasi emosional dan ekspresif. Pengembang disarankan untuk

mengimplementasikan pendekatan ProSDD atau variannya yang secara eksplisit mempelajari variabilitas prosodik.

c. Pengembangan Sistem Real-Time dengan Latensi Rendah: Implementasi industri seperti Synthetic Audio Detection dari FaceOff Technologies menunjukkan bahwa deteksi berbasis prosodi dapat beroperasi secara real-time. Pengembang perlu terus mengoptimalkan efisiensi komputasi agar deteksi dapat dilakukan tanpa mengorbankan akurasi, terutama untuk aplikasi keamanan dan forensik.
d. Pelabelan Konten Sintetis (Watermarking Progressif): Meskipun deteksi berbasis prosodi efektif, pendekatan proaktif seperti watermarking perlu terus dikembangkan. Pengembang disarankan untuk mengadopsi pendekatan hybrid: watermarking untuk pelabelan awal dan deteksi berbasis prosodi sebagai lapisan verifikasi kedua.

5.2.3 Saran bagi Pembuat Kebijakan dan Regulator

a. Standarisasi Evaluasi Deteksi Deepfake: Pemerintah dan badan standardisasi internasional perlu mengembangkan protokol evaluasi standar untuk sistem deteksi audio deepfake yang mencakup tidak hanya akurasi pada dataset benchmark tetapi juga robustness terhadap serangan adversarial, kemampuan generalisasi lintas teknologi generatif, dan kinerja pada ucapan emosional.
b. Regulasi Transparansi Konten Sintetis: Berdasarkan tingkat akurasi deteksi yang belum sempurna (terbaik saat ini sekitar 95%), regulator perlu mempertimbangkan kewajiban transparansi bagi generator deepfake misalnya melalui watermarking wajib atau metadata terbuka daripada hanya mengandalkan deteksi ex post facto.
c. Pendanaan Riset Interseksi Lingusitik-AI: Temuan penelitian ini menunjukkan bahwa investasi dalam riset interdisipliner antara linguistik dan AI sangat produktif. Regulator disarankan untuk mengalokasikan dana riset khusus untuk penelitian yang menjembatani (sosio)linguistik dan deteksi konten sintetis.

5.2.4 Saran untuk Masyarakat Umum dan Pengguna Media

a. Meningkatkan Literasi Audio Digital: Masyarakat perlu diedukasi bahwa suara bukanlah bukti mutlak. Keanehan dalam prosodi suara yang terdengar terlalu mulus (too smooth), tidak ada variasi napas, ritme yang terlalu teratur, atau jeda yang tidak natural dapat menjadi indikator awal bahwa suatu rekaman mungkin adalah deepfake.
b. Verifikasi Lintas Saluran: Untuk panggilan atau pesan suara yang mencurigakan (terutama yang meminta transfer dana atau informasi sensitif), masyarakat disarankan untuk melakukan verifikasi lintas saluran misalnya menelepon kembali melalui nomor yang diketahui asli atau menggunakan aplikasi pesan teks untuk konfirmasi.

c. Pelaporan Konten Curiga: Masyarakat perlu didorong untuk melaporkan konten audio yang mencurigakan ke platform media sosial atau otoritas terkait. Laporan massal dapat membantu platform mengidentifikasi dan menandai deepfake lebih awal.

5.2.5 Saran untuk Platform Media Sosial

a. Integrasi Deteksi Deepfake Otomatis: TikTok, X (Twitter), Facebook, dan platform lain yang memungkinkan unggahan audio perlu mengintegrasikan sistem deteksi deepfake berbasis prosodi sebagai bagian dari pipeline moderasi konten. Sistem ini dapat menandai atau memblokir deepfake berkualitas rendah secara otomatis.
b. Peringatan Pengguna (User Warning): Untuk deepfake yang tidak dapat dikonfirmasi sepenuhnya (misalnya hanya melewati ambang deteksi 50-80%), platform dapat memberikan peringatan kepada pengguna seperti: "Konten audio ini memiliki karakteristik yang mirip dengan deepfake. Harap verifikasi secara independen sebelum mempercayai isinya."
c. Public Dataset Anonim dan Berkualitas: Platform disarankan untuk berkontribusi pada ekosistem deteksi deepfake dengan menyediakan dataset anonim dari deepfake yang berhasil diidentifikasi (dengan izin pengguna dan perlindungan privasi) untuk melatih model deteksi generasi berikutnya.

DAFTAR REFERENSI

Cohen, A., Shyrman, D., Solonskyi, A., Frenkel, R., Krishtul, A., & Gal, O. (2025). Robust prosody modeling for synthetic speech detection. Speech Communication, 174, 103283. https://doi.org/10.1016/j.specom.2025.103283

Mahapatra, A., Ulgen, I. R., Lee, K. A., Andrews, N., & Sisman, B. (2026). ProSDD: Learning prosodic representations for speech deepfake detection against expressive and emotional attacks. arXiv preprint, arXiv:2604.13229.

Warren, K., et al. (2025). Pitch imperfect: Detecting audio deepfakes through acoustic prosodic analysis. arXiv preprint, arXiv:2502.14726.

Khanjani, Z., Ale, T., Wang, J., Davis, L., Mallinson, C., & Janeja, V. P. (2024). Investigating causal cues: Strengthening spoofed audio detection with human-discernible linguistic features. arXiv preprint, arXiv:2409.06033.

Nwosu, L., et al. (2023). Auto-annotation of expert-defined linguistic features for deepfake audio detection. Dalam Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI). (Catatan: Disebutkan dalam Khanjani dkk. 2024 sebagai penelitian tentang EDLFs)

Fossat, Y., et al. (2024). Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation. JMIR Biomedical Engineering.

Mallinson, C., Nwosu, L., & Janeja, V. P. (2024). The intersection of (socio)linguistics and deepfake detection: Opportunities for convergence. Language and Linguistics Compass.

Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., & Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint, arXiv:1904.05441.

Delgado, H., Evans, N., Kinnunen, T., Lee, K. A., Liu, X., Nautsch, A., Patino, J., Sahidullah, M., Todisco, M., & Wang, X. (2021). ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. arXiv preprint, arXiv:2109.00535.

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., & Nautsch, A. (2022). ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. arXiv preprint, arXiv:2210.02437.

Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., Kinnunen, T., Lee, K. A., Vestman, V., & Nautsch, A. (2019). ASVspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database.

Sahidullah, M., et al. (2024). Publicly available datasets analysis and spectrogram-ResNet41 based improved features extraction for audio spoof attack detection. International Journal of System Assurance Engineering and Management, 15, 5611–5636. https://doi.org/10.1007/s13198-024-02550-1

Kinnunen, T., et al. (2020). t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. Dalam Proceedings of Odyssey 2020. (Catatan: Disebutkan dalam literatur evaluasi ASVspoof)

Todisco, M., et al. (2019). Integrated presentation attack detection and automatic speaker verification: Common evaluation framework. IEEE/ACM Transactions on Audio, Speech, and Language Processing. (Catatan: Disebutkan dalam literatur ASVspoof)

Müller, N. M., et al. (2022). Human perception of synthetic speech: A systematic review. Computer Speech & Language. (Catatan: Relevan untuk memahami isyarat prosodik yang dideteksi manusia)

Burkhardt, F., & Sager, S. (2021). Detection of synthetic speech: A human listening study. Dalam Proceedings of Interspeech 2021.

Jin, Z., et al. (2021). Deepfake voice generation: A survey of technologies and detection methods. ACM Computing Surveys.

Shen, J., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. Proceedings of ICASSP 2018. (Catatan: Referensi tentang teknologi generasi TTS seperti Tacotron)