BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi generatif suara dalam beberapa tahun terakhir telah berkembang sangat cepat dan hasilnya makin realistis, tidak hanya untuk audiobook, podcast, dubbing, dan asisten virtual, tetapi juga untuk meniru identitas suara melalui kloning suara [1]. Kemajuan yang menonjol terlihat pada kemampuan meniru suara dengan data yang sangat sedikit, di mana kloning suara dapat bekerja dengan kisaran sekitar 5 sampai 10 detik sampel suara target untuk menghasilkan suara sintetik yang mengikuti karakter penutur [2]. Salah satu bentuk kloning suara dilakukan melalui *text-to-speech*, yaitu dengan menggunakan audio referensi untuk membawa karakteristik identitas vokal penutur, kemudian menghasilkan ujaran berdasarkan teks yang ditentukan secara terpisah [3]. Teknik generasi suara modern juga terus meningkat dari sisi natural dan kemiripan penutur, serta semakin mudah diintegrasikan ke produk digital karena *pipeline text-to-speech* semakin efisien [4]. Ketersediaan layanan generasi suara membuat teknologi tersebut semakin mudah diakses dan diterapkan di luar lingkungan penelitian. Namun, kemudahan akses tersebut menimbulkan risiko penyalahgunaan, terutama pada skenario kloning suara berbasis *text-to-speech* yang meniru identitas vokal seseorang untuk memanipulasi korban.

Peningkatan kemampuan kloning suara berbasis *text-to-speech* memperbesar potensi penyalahgunaan dalam kejahatan berbasis manipulasi audio karena suara sering dipakai sebagai sinyal kepercayaan dalam komunikasi personal maupun bisnis. Jenis kloning suara ini dapat digunakan untuk membuat suara hasil kloning mengucapkan kalimat baru yang tidak pernah diucapkan oleh penutur asli, sehingga teknologi ini digunakan pada skenario impersonasi ketika identitas suara seseorang digunakan untuk menyampaikan instruksi atau informasi palsu [5]. Laporan Regula Forensics tahun 2024 menunjukkan bahwa persentase insiden *deepfake* berbasis audio yang dilaporkan meningkat dari 37% pada tahun 2022 menjadi 49% pada

tahun 2024 [6]. Peningkatan ini relevan terhadap ancaman kloning suara karena teknologi kloning suara merupakan salah satu bentuk audio *deepfake* yang dapat digunakan untuk meniru identitas vokal seseorang. Salah satu contoh kasus penipuan impersonasi terjadi di Inggris, di mana pelaku mengkloning suara direktur perusahaan menggunakan teknologi *Artificial Intelligence* (AI) dan menghubungi seorang eksekutif pada perusahaan anak untuk menginstruksikan transfer sebesar 220 ribu Euro atau sekitar 4,53 miliar Rupiah kepada pemasok di Hungaria [7]. Di Indonesia, penyalahgunaan kloning suara juga muncul dalam penyebaran disinformasi politik. Menjelang Pemilu 2024, beredar rekaman suara yang dibuat seolah-olah Ketua Umum Partai NasDem Surya Paloh sedang memarahi Anies Baswedan. Hasil pemeriksaan menunjukkan bahwa rekaman tersebut merupakan hasil generasi AI *text-to-speech*, dan Partai NasDem serta Kominfo menyatakan percakapan itu sebagai hoaks. Kasus ini menunjukkan bahwa kloning suara dapat digunakan untuk mempengaruhi opini publik dan berpotensi merusak reputasi pihak yang identitas suaranya ditiru [8]. Data global dan nasional menunjukkan bahwa kloning suara sudah menjadi alat penipuan yang canggih, sehingga urgensi membangun sistem deteksi kloning suara semakin tinggi.

Sistem deteksi kloning suara umumnya membedakan audio asli dan kloning berdasarkan artefak yang tertinggal dari proses pembentukan suara. Kemampuan deteksi dipengaruhi oleh bentuk representasi masukan dan arsitektur yang digunakan. Penelitian sistem berbasis fitur manual CQCC-GMM dan LFCC-GMM menghasilkan EER sebesar 15,62% dan 19,30%, sedangkan sistem berbasis representasi *(self-supervised)* yang mencapai EER sebesar 2,89% [9]. Namun model *attention* yang bukan merupakan berbasis representasi juga dapat mencapai EER 0,83%, lebih rendah dibandingkan sistem berbasis representasi pada benchmark yang sama [10]. Temuan ini menunjukkan bahwa performa sistem deteksi tidak dapat hanya dijelaskan dari kategori fitur manual atau *self-supervised*. Selain perbedaan representasi, keterbatasan lain terdapat pada ketersediaan evaluasi yang secara khusus menggunakan data Bahasa Indonesia. Keterbatasan kajian Bahasa Indonesia terlihat dari perbedaan performa arsitektur yang sama pada *benchmark* berbeda. Pada InaSpoof-v1, model AASIST-L menghasilkan EER

sebesar 4,19%, sedangkan pada data ASVspoof 2019 LA, model tersebut mencapai EER 0,99% [10], [11]. Hasil ini menunjukkan bahwa performa pada *benchmark* umum belum tentu sama pada data Bahasa Indonesia. Oleh karena itu, diperlukan evaluasi menggunakan data Bahasa Indonesia sekaligus kajian terhadap bentuk representasi yang sesuai untuk membedakan audio asli dan kloning.

Perkembangan *deep learning* dalam deteksi kloning suara pada data bahasa Inggris menunjukkan bahwa performa sistem deteksi ditentukan oleh jenis fitur masukan dan kesesuaian arsitektur dalam memproses representasi tersebut. Model berbasis *Convolutional Neural Network* (CNN) banyak digunakan karena mampu mempelajari pola lokal pada representasi spektral, dengan laporan akurasi 98% dengan EER yang berkisar antara 1,6% hingga 2,95% [12], [13]. Namun, artefak sintesis dapat muncul pada perubahan temporal maupun struktur frekuensi sehingga membutuhkan pemodelan hubungan waktu dan frekuensi yang lebih eksplisit. Salah satu model yang relevan adalah *Dual-Path Time-Frequency Attention Network* (DPTFAN), yang memproses informasi waktu dan frekuensi melalui dua jalur perhatian terpisah sebelum hasilnya digabungkan [14]. Model ini mencapai EER sebesar 0,35%, lebih unggul dibandingkan model CNN *baseline* yang menghasilkan EER 10,54% pada dataset pengujian dalam penelitian asalnya, sehingga DPTFAN dipilih sebagai cabang yang mewakili pendekatan berbasis representasi spektral dan akustik. Hasil tersebut menunjukkan bahwa pemodelan waktu-frekuensi berbasis *attention* memiliki kinerja kuat, tetapi representasi akustik belum selalu menjadi satu-satunya sumber informasi yang optimal untuk deteksi audio kloning. Artefak sintesis juga dapat tercermin pada susunan unit ujaran dan hubungan temporal antarbunyi.

Representasi lain yang perlu dipertimbangkan adalah representasi fonetik dan kontekstual, karena representasi akustik menggambarkan pola energi, frekuensi, dan perubahan temporal sinyal, sedangkan informasi fonetik berkaitan dengan unit bunyi dan hubungan ujaran yang terbentuk sepanjang sinyal. Representasi ini dapat diperoleh melalui pendekatan *self-supervised speech representation*, yang dilaporkan mampu mengodekan informasi kelas fonetik, identitas kata, pelafalan, fitur sintaksis, dan fitur semantik dari sinyal ujaran [15]. Dalam deteksi audio

kloning suara, model *self-supervised* seperti Wav2Vec 2.0 dan WavLM banyak digunakan sebagai front-end representasi. Dari sisi performa, beberapa sistem deteksi kloning suara berbasis Wav2Vec 2.0 menunjukkan nilai EER pada rentang 0,40% hingga 1,25%, sedangkan sistem berbasis WavLM berada pada rentang 0,42% hingga 2,47% [4]. Perbandingan tersebut menunjukkan bahwa Wav2Vec 2.0 merupakan salah satu model *self-supervised* yang sangat kompetitif untuk deteksi audio kloning. Wav2Vec 2.0 juga dipilih dalam penelitian ini karena model tersebut membentuk representasi ujaran kontekstual dari sinyal audio, sehingga berbeda dari DPTFAN yang menggunakan representasi waktu-frekuensi. Perbedaan cara pembentukan representasi tersebut membuat DPTFAN dan Wav2Vec 2.0 tidak hanya diperlakukan sebagai dua model tunggal, tetapi sebagai dua sumber informasi yang perlu digabungkan melalui strategi hibrida yang sesuai.

Pendekatan hibrida menjadi relevan karena keluaran dari dua model dengan representasi berbeda dapat saling melengkapi dalam proses keputusan deteksi. Pada penggabungan fitur yang dirancang secara khusus, MAFF-Net memadukan fitur multi-spektrogram dan fitur Wav2Vec 2.0 melalui *cross-attention* untuk menggabungkan karakteristik akustik tingkat rendah dan representasi ujaran tingkat tinggi [16]. Sementara itu, penelitian dengan metode fusi skor menunjukkan bahwa penggabungan keluaran beberapa detektor dapat memanfaatkan informasi yang saling melengkapi tanpa menyatukan fitur internal setiap model. Pada studi tersebut, model tunggal terbaik memperoleh EER 8,90%, sedangkan gabungan model dengan fusi skor menurunkan EER menjadi 5,30% [17]. Hasil tersebut menunjukkan bahwa fusi dapat memperbaiki keputusan model, tetapi titik dan mekanisme fusi tetap perlu disesuaikan dengan karakter model yang digabungkan. Dalam penelitian ini, fusi skor dipilih karena DPTFAN dan Wav2Vec 2.0 membentuk representasi dari jalur yang berbeda. DPTFAN bekerja pada log-Mel dan fitur akustik pendamping, sedangkan Wav2Vec 2.0 menghasilkan representasi *self-supervised* dari sinyal ujaran. Jika kedua keluaran internal tersebut digabung langsung pada level fitur atau *embedding*, sistem perlu membentuk ruang representasi bersama melalui penyesuaian dimensi, proyeksi, dan pelatihan tambahan. Risiko ini terlihat pada studi fusi HuBERT dan WavLM, ketika

penggabungan *embedding* secara langsung menghasilkan EER 0,89%, lebih buruk dibanding model individual HuBERT sebesar 0,63% dan WavLM sebesar 0,64% [18]. Temuan tersebut menunjukkan bahwa penggabungan representasi berbeda tidak otomatis memperbaiki performa apabila mekanisme fusinya tidak dirancang dengan tepat. Oleh karena itu, penelitian ini menggunakan metode fusi skor agar DPTFAN dan Wav2Vec 2.0 tetap bekerja pada ruang representasinya masing-masing. Pemilihan fusi skor tidak dimaksudkan untuk menyatakan bahwa fusi skor selalu lebih baik daripada fusi fitur atau *embedding*, tetapi untuk menyesuaikan metode fusi dengan desain penelitian yang menggunakan dua model berbeda dan tidak melatih ulang keduanya sebagai satu arsitektur *end-to-end*.

Penelitian ini diarahkan untuk mendeteksi audio asli dan audio hasil kloning suara bahasa Indonesia berbasis *text-to-speech* yang memadukan *Dual-Path Time-Frequency Attention Network* dan Wav2Vec 2.0 melalui fusi skor. Penelitian ini diharapkan dapat menghasilkan model yang mampu meningkatkan kinerja deteksi kloning suara pada dataset yang digunakan melalui penggabungan dua jenis representasi yang berbeda. Model berbasis *attention* digunakan untuk membentuk representasi akustik pada domain waktu dan frekuensi, sedangkan model *self-supervised transformer* digunakan untuk membentuk representasi ujaran kontekstual yang dapat memuat informasi fonetik dan temporal dari sinyal ujaran. Kontribusi penelitian ini adalah menyusun kerangka evaluasi deteksi kloning suara yang spesifik untuk bahasa Indonesia, perbandingan komparatif antara model tunggal dan model fusi dengan metrik EER sebagai acuan utama, serta pembuktian bahwa pendekatan hibrida berbasis fusi skor mampu memberikan peningkatan performa dibandingkan arsitektur tunggal pada data yang diuji.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini diantaranya:

1. Bagaimana performa kinerja sistem deteksi kloning suara berbahasa Indonesia berbasis *Dual-Path Time-Frequency Attention Network* mampu menangkap artefak akustik pada domain waktu dan frekuensi secara efektif pada dataset penelitian?

2. Bagaimana performa kinerja model Wav2Vec 2.0 sebagai representasi *self-supervised transformer* mampu menangkap ciri fonetik dan linguistik yang muncul pada audio kloning dalam dataset penelitian?
3. Bagaimana performa kinerja model hibrida *Dual-Path Time-Frequency Attention Network* dan Wav2Vec 2.0 mampu meningkatkan kinerja deteksi dibandingkan model tunggal berdasarkan metrik evaluasi *Equal Error Rate*, AUC, dan akurasi pada dataset penelitian?
4. Bagaimana mengimplementasikan model hibrida *Dual-Path Time-Frequency Attention Network* dan Wav2Vec 2.0 pada deteksi kloning suara berbahasa Indonesia berbasis situs web menggunakan Streamlit?

1.3 Batasan Masalah

Batasan masalah dari penelitian ini diantaranya:

1. Penelitian difokuskan pada deteksi kloning suara berbasis Bahasa Indonesia, tidak mencakup deteksi *deepfake* video atau manipulasi multimodal lainnya.
2. Dataset yang digunakan berasal dari dataset Fake-or-Real sebagai dataset *benchmark*, serta dataset Common Voice Indonesia sebagai dataset penelitian.
3. Jenis serangan yang ditangani terbatas pada kloning suara berbasis *text-to-speech*, tidak mencakup konversi suara dan *replay attack*.
4. Penelitian tidak melakukan pengelompokan maupun evaluasi performa model berdasarkan aksen atau dialek penutur karena dataset tidak menyediakan metadata dialek yang terstruktur dan dapat diverifikasi.
5. Penelitian tidak mencakup pengujian lintas dataset maupun evaluasi terhadap generator kloning suara lain yang tidak digunakan dalam pembentukan dataset.
6. Mekanisme fusi dilakukan pada level skor tanpa eksplorasi fusi multimodal di luar domain audio.
7. Evaluasi performa difokuskan pada metrik EER sebagai metrik utama, dengan metrik pendukung seperti AUC dan akurasi.

8. Implementasi sistem dilakukan pada lingkungan komputasi terbatas dan tidak mencakup optimasi *deployment* skala industri secara penuh.
9. Penelitian tidak membahas aspek hukum atau kebijakan terkait penyalahgunaan kloning suara, melainkan berfokus pada pengembangan model deteksi.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, terdapat tujuan dari penelitian ini, yaitu:

1. Mengetahui performa kinerja model deteksi kloning suara berbahasa Indonesia berbasis *Dual-Path Time-Frequency Attention Network* yang mampu mempelajari representasi akustik pada domain waktu dan frekuensi secara simultan pada dataset penelitian.
2. Mengetahui performa kinerja model Wav2Vec 2.0 sebagai representasi *self-supervised transformer* untuk menangkap pola fonetik dan linguistik pada audio kloning dalam dataset penelitian.
3. Mengukur dan membandingkan performa kinerja model tunggal dan model hibrida *Dual-Path Time-Frequency Attention Network* dan Wav2Vec 2.0 dengan metrik evaluasi *Equal Error Rate*, AUC, dan akurasi pada dataset penelitian.
4. Mengimplementasikan model hibrida *Dual-Path Time-Frequency Attention Network* dan Wav2Vec 2.0 pada deteksi kloning suara berbahasa Indonesia berbasis situs web menggunakan Streamlit.

1.4.2 Manfaat Penelitian

Penelitian ini memiliki manfaat yang diharapkan, antara lain:

1. Manfaat Teoritis

   Manfaat teoritis yang diharapkan dalam penelitian ini adalah:

   a. Memberikan kontribusi pada pengembangan metode deteksi kloning suara berbasis *attention* yang secara khusus dirancang untuk karakteristik fonetik Bahasa Indonesia.

b. Menyediakan analisis empiris mengenai efektivitas fusi antara model CNN berbasis *attention* dan *self-supervised transformer* dalam sistem deteksi kloning suara.
c. Menambah literatur akademik terkait evaluasi deteksi kloning suara menggunakan metrik *Equal Error Rate* pada bahasa non-Inggris yang masih relatif terbatas dibandingkan penelitian berbahasa Inggris.
d. Memberikan dasar konseptual mengenai integrasi representasi akustik dan linguistik dalam satu kerangka sistem deteksi terpadu.

2. Manfaat Praktis

Manfaat praktis yang diharapkan dalam penelitian ini adalah:

a. Menyediakan hasil evaluasi empiris mengenai model yang diuji dalam membedakan audio asli dan audio kloning pada dataset yang digunakan dalam penelitian.
b. Menjadi referensi implementasi penggabungan dua model dengan representasi berbeda melalui fusi skor, tanpa menggabungkan fitur internal atau melatih ulang kedua model sebagai satu arsitektur *end-to-end*.
c. Memberikan gambaran awal mengenai perubahan performa model ketika audio mengalami degradasisebagai bahan pertimbangan dalam pengembangan sistem selanjutnya.
d. Menyediakan prototipe demonstrasi sederhana (unggah atau rekam audio) sebagai uji coba.

1.5 Sistematika Penulisan

Penulisan karya ilmiah ini didasarkan pedoman tugas akhir mahasiswa Sistem Informasi Universitas Multimedia Nusantara, yang bertujuan untuk mempermudah para pembaca dalam memahami isi dari karya ilmiah ini. Isi dari karya ilmiah ini terbagi menjadi lima bab secara terstruktur, yaitu:

BAB I LATAR BELAKANG

Bab ini berisi latar belakang penelitian, termasuk penjelasan tentang pentingnya topik yang diangkat, permasalahan yang menjadi dasar penelitian, tujuan penelitian,

dan batasan masalah. Bab ini juga menjelaskan manfaat penelitian baik dari segi teoritis maupun praktis, serta rumusan masalah yang ingin dijawab melalui penelitian ini.

BAB II LANDASAN TEORI

Bab ini menguraikan teori-teori yang mendukung penelitian, serta tinjauan terhadap penelitian-penelitian terdahulu yang relevan. Penelitian terdahulu akan merangkum studi terkait dan teori yang dibahas mencakup tentang algoritma dan pendekatan yang digunakan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan rencana dan tahapan penelitian, mencakup pengumpulan *dataset*, proses pelatihan model, dan evaluasi kinerja model. Metodologi disusun secara sistematis untuk memberikan gambaran jelas mengenai proses penelitian tanpa memuat hasil atau pembahasan.

BAB IV ANALISIS DAN HASIL PENELITIAN

Bab ini memaparkan hasil penelitian, termasuk hasil penerapan metode yang diusulkan, hasil eksperimen, serta pengujian dan validasi model. Bab ini juga mencakup analisis mendalam terhadap hasil penelitian, termasuk analisis peningkatan akurasi, efektivitas model, dan bagaimana model berhasil mengatasi tantangan yang dihadapi.

BAB V KESIMPULAN DAN SARAN

Bab ini merangkum hasil penelitian dalam bentuk kesimpulan yang singkat dan jelas, menjawab rumusan masalah yang diajukan, serta memberikan saran untuk pengembangan penelitian lebih lanjut.