JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i1.9064

PENGEMBANGAN SISTEM SPEECH EMOTION RECOGNITION BERBASIS DEEP LEARNING WAV2VEC2.0 UNTUK RESPONS EMOSIONAL KARAKTER KUCING INTERAKTIF DI GAME UNITY

Farrel Reyhan Putra[1*], Hafidz Muhammad Dzaky[2], Maheswara Putratama[3], Mochammad Rabee Fathi Al Fikri[4], Vitri Tundjungsari[5]

[1,2,3,4,5]Universitas Esa Unggul; Jl Arjuna Utara No.9, Kebon Jeruk, Jakarta, Indonesia; (021) 39529950

Keywords:
Artificial Intelligence, Speech Emotion Recognition, Wav2Vec2.0, Deep Learning, Game Unity.

Corespondent Email:
sunmgaming@student.esaunggul.ac.id

Abstrak. *Penelitian ini bertujuan untuk mengembangkan sistem Speech Emotion Recognition (SER) berbasis deep learning yang mampu mengenali emosi suara dan mengimplementasikannya pada sebuah game interaktif menggunakan Unity. Model SER dibangun dengan memanfaatkan arsitektur Wav2Vec 2.0 yang telah dipra-latih (pretrained) dan dilakukan fine-tuning menggunakan dataset CREMA-D dengan empat kelas emosi, yaitu angry, happy, neutral, dan sad. Data dibagi dengan rasio 80% untuk pelatihan dan 20% untuk validasi. Hasil pengujian menunjukkan bahwa model mampu mencapai nilai akurasi validasi maksimum sebesar 78–79% dengan weighted F1-score sebesar 0,79. Analisis confusion matrix memperlihatkan bahwa kelas angry memiliki tingkat pengenalan tertinggi, sementara kesalahan klasifikasi paling sering terjadi antara kelas neutral dan sad akibat kemiripan karakteristik prosodi. Model yang telah dilatih kemudian diekspor ke format ONNX dan berhasil diintegrasikan ke dalam game Unity untuk mendeteksi emosi suara pemain secara real-time. Hasil implementasi menunjukkan bahwa sistem mampu memberikan respons karakter yang adaptif berdasarkan emosi suara pengguna, sehingga meningkatkan interaksi dalam permainan*

Abstract. *This study aims to develop a deep learning-based Speech Emotion Recognition (SER) system capable of recognizing voice emotions and implementing it in an interactive game using Unity. The SER model was built using the pretrained Wav2Vec 2.0 architecture and fine-tuned using the CREMA-D dataset with four emotion classes, namely angry, happy, neutral, and sad. The data was divided with a ratio of 80% for training and 20% for validation. The test results showed that the model was able to achieve a maximum validation accuracy value of 78–79% with a weighted F1-score of 0.79. The confusion matrix analysis showed that the angry class had the highest recognition rate, while classification errors most often occurred between the neutral and sad classes due to similarities in prosodic characteristics. The trained model was then exported to ONNX format and successfully integrated into the Unity game to detect players' voice emotions in real-time. The implementation results show that the system is capable of providing adaptive character responses based on the user's voice emotions, thereby enhancing interaction in the game.*

1. PENDAHULUAN

Sinyal suara merupakan salah satu sarana komunikasi alami manusia yang tidak

hanya menyampaikan informasi verbal, tetapi juga mengandung informasi emosional yang tercermin melalui parameter vokal seperti *pitch*, intensitas, kecepatan berbicara, dan kualitas suara [1]. Meskipun demikian, pengenalan emosi dari suara secara otomatis masih menjadi tantangan bagi sistem berbasis mesin.

Speech Emotion Recognition (SER) merupakan metode untuk mengidentifikasi emosi pembicara melalui analisis karakteristik sinyal suara, seperti intonasi, energi, dan dinamika temporal. Dengan memanfaatkan pendekatan *deep learning*, SER mampu mempelajari pola emosional secara otomatis dari data audio dan menghasilkan performa yang lebih akurat serta adaptif dibandingkan metode tradisional [2]. Studi lain juga telah menelaah kemajuan riset SER berdasarkan *deep learning*, yang mencakup berbagai arsitektur jaringan dan strategi pelatihan untuk menangani kompleksitas emosi dalam suara [3].

Dalam pengembangan *game* interaktif, respons karakter yang bersifat statis dapat mengurangi tingkat realisme interaksi. Penerapan SER memungkinkan karakter dalam *game* menyesuaikan perilaku dan respons secara langsung berdasarkan emosi suara pemain, sehingga interaksi menjadi lebih natural. Oleh karena itu, penelitian ini mengembangkan sistem SER berbasis *deep learning* menggunakan *Wav2Vec 2.0* untuk menentukan respons emosional karakter kucing interaktif dalam *game Unity* dengan memanfaatkan dataset CREMA-D sebagai data pelatihan dan evaluasi.

2. TINJAUAN PUSTAKA

2.1. *Machine Learning*

Machine Learning merupakan cabang dari *Artificial Intelligence* (AI) yang mempelajari cara membuat komputer mampu belajar dari data, mengenali pola, serta meningkatkan kemampuannya secara mandiri dengan meniru, bahkan melampaui, kemampuan belajar manusia, sehingga sistem dapat melakukan prediksi atau pengelompokan secara otomatis dan meningkatkan kinerja serta keakuratan melalui proses pelatihan dan paparan data dalam jumlah besar [4][5][6].

2.2. *Deep Learning*

Deep learning merupakan pendekatan dalam *machine learning* yang berada dalam ranah *Artificial Intelligence* dan terinspirasi dari struktur jaringan saraf tiruan (*Artificial Neural Networks*) pada otak manusia [5], [7], [8].

Pendekatan ini memungkinkan sistem mempelajari representasi data secara otomatis dari data mentah tanpa memerlukan perancangan fitur manual, berbeda dengan metode *machine learning* konvensional yang umumnya bergantung pada ekstraksi fitur berbasis statistic [9].

Dalam pengolahan suara, *deep learning* banyak digunakan karena kemampuannya dalam mempelajari pola kompleks seperti intonasi, frekuensi, dan dinamika temporal yang berkaitan dengan emosi. Model *end-to-end* berbasis *deep learning* juga memanfaatkan komputasi tensor yang dapat dioptimalkan menggunakan GPU, sehingga menghasilkan kinerja dan efisiensi yang lebih baik dibandingkan metode konvensional [10]. Oleh karena itu, pendekatan *deep learning* menjadi pilihan utama dalam penelitian *speech recognition* dan *speech emotion recognition*.

2.3. Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) adalah proses untuk mengidentifikasi emosi seseorang berdasarkan sinyal suara yang diucapkan [11]. Emosi yang umum dikenali antara lain senang, sedih, marah, takut, dan netral. Informasi emosi ini diperoleh dari karakteristik suara seperti tinggi nada, intensitas, tempo, dan pola intonasi. Tujuan dari SER adalah memanfaatkan informasi yang berkaitan dengan kondisi emosional manusia agar sistem komputer atau mesin dapat merespons secara lebih tepat, sehingga interaksi dan komunikasi antara manusia dan mesin menjadi lebih alami, efektif, dan adaptif terhadap keadaan emosional pengguna [12].

Penerapan SER banyak digunakan dalam berbagai bidang, seperti *virtual assistance*, sistem interaksi manusia dan komputer, serta pengembangan karakter interaktif dalam *game*. Dengan adanya SER, sistem dapat memberikan respons yang lebih natural dan sesuai dengan kondisi emosional pengguna.

2.4. *Wav2Vec 2.0*

Wav2Vec 2.0 merupakan model *deep learning* berbasis arsitektur *transformer* yang dikembangkan untuk pengolahan data suara. Metode *Wav2Vec 2.0* menggunakan konsep "*self-supervised learning*" yang memungkinkan model untuk memahami data tanpa memerlukan label (data tanpa transkripsi teks) [13]. Model ini bekerja langsung pada sinyal audio mentah (*raw audio*) dan mampu mempelajari representasi fitur suara tanpa memerlukan ekstraksi fitur secara manual.

Model *Wav2Vec 2.0* terdiri dari beberapa tahap pemrosesan. Pada tahap awal, sinyal *raw audio* diproses oleh *local encoder* yang menggunakan *convolutional neural network* untuk mengekstraksi ciri-ciri dasar suara. Tahap berikutnya adalah *contextualized encoder* berbasis *transformer* yang berfungsi untuk memahami hubungan antar potongan suara, sehingga model dapat menangkap konteks dan makna ucapan secara keseluruhan [14].

Dalam penelitian ini, *Wav2Vec 2.0* digunakan untuk mengekstraksi fitur suara yang kemudian dimanfaatkan dalam proses pengenalan emosi suara. Model *Wav2vec 2.0* dipilih karena representasi ucapan berbasis *self-supervised learning* terbukti efektif pada berbagai tugas pemrosesan suara, termasuk pengenalan emosi [15].

Keunggulan utama *Wav2Vec 2.0* adalah kemampuannya dalam mempelajari pola suara secara *self-supervised*, sehingga dapat memberikan performa yang baik meskipun jumlah data berlabel terbatas. Sebagai salah satu model yang dianggap efektif dalam pengenalan ucapan, ada juga kelemahan dari model ini, *Wav2Vec 2.0* memiliki beberapa keterbatasan, terutama saat menghadapi kondisi audio dengan tingkat kebisingan yang tinggi [16].

2.6. SER pada Sistem Interaktif/Game

Dalam sistem interaktif dan *game*, konsep karakter adaptif mengacu pada kemampuan karakter virtual untuk menyesuaikan perilaku atau respons berdasarkan kondisi pengguna. Emosi pengguna berperan penting dalam menciptakan pengalaman bermain yang lebih imersif, karena respons karakter yang sesuai dengan emosi dapat meningkatkan rasa keterlibatan dan realisme interaksi.

Penerapan *Speech Emotion Recognition* memungkinkan sistem *game* mengenali emosi pemain melalui suara dan menggunakannya sebagai dasar pengambilan keputusan karakter. Secara konseptual, integrasi SER dengan *game engine* seperti *Unity* mendukung pengembangan interaksi yang lebih natural dan dinamis, di mana karakter dapat merespons kondisi emosional pemain secara *real-time*, sehingga memperkaya pengalaman bermain dan meningkatkan kualitas interaksi manusia komputer.

3. METODE PENELITIAN

3.1. Rancangan Penelitian

Penelitian ini merupakan penelitian eksperimental yang bertujuan untuk mengembangkan dan menguji sistem *Speech Emotion Recognition* (SER) berbasis *deep learning* serta mengintegrasikannya ke dalam *game* interaktif. Eksperimen dilakukan untuk mengevaluasi kemampuan sistem dalam mengenali emosi suara pengguna dan menghasilkan respons karakter yang adaptif.

Secara umum, penelitian ini mencakup proses pengolahan data suara, pengenalan emosi menggunakan model SER, serta pemanfaatan hasil prediksi emosi sebagai dasar respons karakter dalam *game* berbasis Unity.

3.2. Sumber Data

Data suara yang digunakan dalam penelitian ini berasal dari dataset publik CREMA-D (*Crowd-sourced Emotional Multimodal Actors Dataset*) yang diperoleh dari sumber *open-source* pada platform Kaggle. Dataset CREMA-D menyediakan rekaman suara berbahasa Inggris dengan label emosi yang telah terdefinisi secara jelas dan banyak digunakan dalam penelitian pengenalan emosi suara.

*Sumber:*https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en

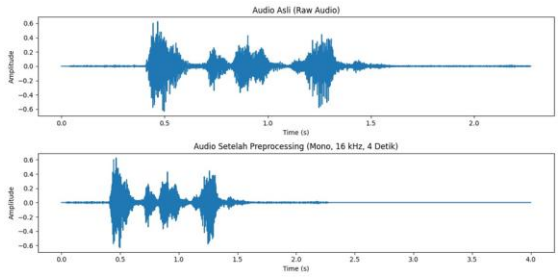
Tabel 1. CREMA-D Dataset *Speech Emotion Recognition* Kaggle

Dataset	Bahasa	Kelas Emosi	Jumlah Sampel
CREMA-D	Inggris	*SAD - sadness* *ANG - angry* *DIS - disgust* *FEA - fear* *HAP - happy* *NEU - neutral*	7.442 Samples

Dalam penelitian ini, hanya empat kelas emosi yang digunakan, yaitu *angry*, *happy*, *neutral*, dan *sad*, menyesuaikan dengan kebutuhan sistem dan ketersediaan label pada dataset.

3.3. *Preprocessing* dan Augmentasi Data

Tahap *preprocessing* dilakukan untuk menyeragamkan format data audio sebelum diproses oleh model. Setiap *file audio* dikonversi menjadi sinyal *mono*, disesuaikan ke *sampling rate* 16 kHz, serta direpresentasikan dalam format *floating-point*. Selain itu, durasi *audio* diseragamkan menjadi maksimum 3 detik menggunakan teknik pemotongan (*truncation*) atau *zero-padding*. Contoh tahapan dari *pre-processing audio* ditunjukkan pada gambar 1 di bawah ini.



Gambar 1. Tahapan Konversi *Mono*, *Resampling*, dan Penyeragaman Durasi Audio (*Pre-processing Audio)*

Untuk meningkatkan variasi data dan mencegah *overfitting*, dilakukan teknik augmentasi data berupa injeksi *Additive White Gaussian Noise* (AWGN). Gangguan derau (noise) ditambahkan pada sinyal asli menggunakan distribusi normal standar ($\mu=0$, $\sigma=1$) dengan faktor intensitas konstan sebesar 0.003. Augmentasi ini diterapkan secara stokastik dengan probabilitas $p=0.3$ pada setiap sampel data latih, sehingga model dipaksa untuk mempelajari fitur emosi yang robust terhadap gangguan akustik minor.

3.4. Arsitektur dan Pelatihan Model

Model yang digunakan dalam penelitian ini adalah *Wav2Vec 2.0* base yang diadaptasi untuk tugas klasifikasi emosi suara menggunakan *sequence classification head*. *Wav2Vec 2.0* berfungsi sebagai ekstraktor fitur dari sinyal *audio* mentah, sementara lapisan klasifikasi digunakan untuk memprediksi kelas emosi.

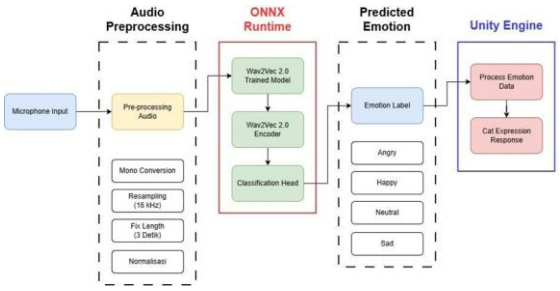
Proses pelatihan dilakukan dengan metode *fine-tuning* parsial, di mana hanya tiga lapisan *encoder* terakhir dari *Wav2Vec 2.0* yang dilatih ulang, sedangkan lapisan lainnya dibekukan. Pendekatan ini bertujuan untuk mempertahankan representasi dasar suara sekaligus menyesuaikan model dengan tugas pengenalan emosi. Optimasi dilakukan menggunakan algoritma *AdamW* dengan *learning rate* berbeda untuk *encoder* dan *classifier*. Untuk mengatasi ketidak seimbangan kelas, digunakan *class-weighted cross entropy loss*. Data dibagi menjadi data latih dan data validasi dengan rasio 80% untuk pelatihan dan 20% untuk evaluasi. Proses pelatihan dilakukan selama maksimal 40 *epoch* dengan mekanisme *early stopping*.

3.5. Metode Evaluasi

Evaluasi performa model dilakukan menggunakan data validasi yang tidak digunakan dalam proses pelatihan. Metode evaluasi yang digunakan meliputi akurasi dan *F1-score* berbobot (*weighted F1-score*) untuk mengukur kemampuan model dalam mengklasifikasikan emosi suara secara keseluruhan.

Selain metrik kuantitatif, performa model juga dianalisis menggunakan *confusion matrix* untuk melihat distribusi prediksi pada setiap kelas emosi. Model dengan nilai akurasi terbaik pada data validasi disimpan sebagai model akhir. Hasil evaluasi ini digunakan sebagai dasar untuk menilai efektivitas sistem *Speech Emotion Recognition* sebelum diimplementasikan ke dalam *game Unity* sebagai pengendali respons emosional karakter kucing interaktif.

3.6. Integrasi Model SER ke Unity



Gambar 2. Rancangan Integrasi Model SER pada *Unity* menggunakan *ONNX Runtime*

Model *Speech Emotion Recognition* (SER) yang telah dilatih diekspor ke format *ONNX* (*Open Neural Network Exchange*), yaitu format pertukaran model *deep learning* yang dikembangkan oleh *Microsoft* untuk mendukung interoperabilitas antar *platform.* Model tersebut dijalankan pada *Unity* menggunakan *ONNX Runtime* sebagai engine inferensi.

Pada sisi *Unity*, suara pemain direkam melalui mikrofon dan diproses melalui tahap audio *preprocessing* sesuai dengan data pelatihan. *Audio* kemudian diberikan sebagai input ke model SER untuk menghasilkan prediksi emosi berupa label *angry, happy, neutral,* atau *sad*. Hasil prediksi emosi digunakan untuk mengatur ekspresi dan animasi karakter kucing dalam *game* secara *real-time.*

3.7 Tools dan Lingkungan Pengembangan

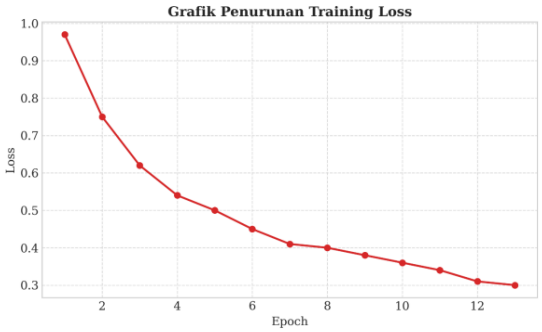
Dalam proses pengembangan, digunakan beberapa perangkat lunak sebagai berikut:

1. *Unity Engine* : Sebagai *platform* utama pengembangan *game*.
2. C# : Bahasa pemrograman untuk penerapan model *Wav2Vec 2.0* dan respon interaktif kucing.
3. *Visual Studio Code* : Sebagai lingkungan pengembangan (*code editor*) untuk menulis dan menguji *script* C#.
4. *Python* : Bahasa pemrograman untuk melakukan *training* model AI.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Training

Proses pelatihan model dilakukan dengan pembagian data sebesar 80% untuk data pelatihan dan 20% untuk data validasi.



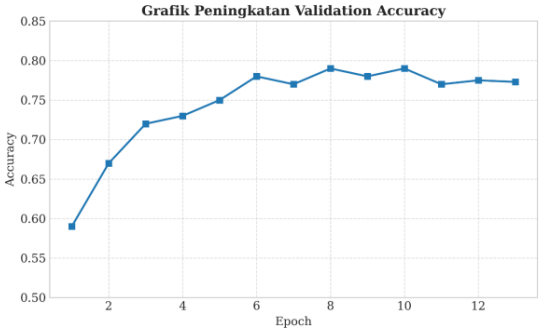
Gambar 3. Hasil Grafik *Training Loss*

Gambar 3 menunjukkan perubahan nilai training loss selama proses pelatihan model. Terlihat bahwa nilai *loss* mengalami penurunan yang signifikan pada *epoch* awal, dari sekitar 0,97 pada *epoch* pertama hingga mendekati 0,30 pada *epoch* ke-12.

Penurunan *loss* yang berlangsung secara bertahap dan konsisten mengindikasikan bahwa model mampu mempelajari representasi fitur emosi suara dengan baik. Selain itu, tidak terlihat lonjakan nilai loss pada *epoch* akhir, yang menunjukkan bahwa proses pelatihan berjalan stabil dan tidak mengalami ketidakstabilan atau kegagalan konvergensi.

4.2. Hasil Evaluasi

Evaluasi performa model dilakukan menggunakan data validasi yang tidak digunakan selama proses pelatihan.

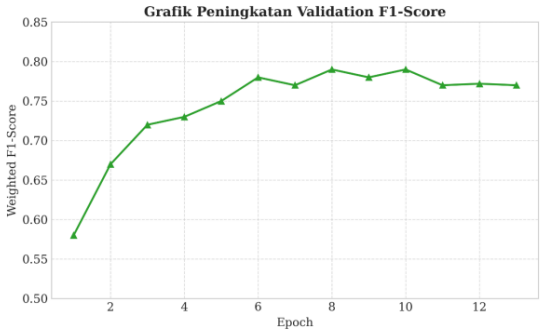


Gambar 4. Hasil Grafik *Validation Accuracy*

Gambar 4 menunjukkan perkembangan nilai akurasi pada data validasi selama proses pelatihan. Nilai akurasi meningkat secara

signifikan pada beberapa *epoch* awal, dari sekitar 59% hingga mencapai nilai maksimum mendekati 79%. Akurasi validasi juga mengalami peningkatan pesat di awal dan kemudian stabil di kisaran 77% - 79%.

Setelah mencapai nilai tersebut, akurasi validasi cenderung berada pada kondisi stabil dengan fluktuasi kecil pada *epoch* selanjutnya. Akurasi validasi yang stabil pada kisaran nilai tinggi menunjukkan bahwa model mampu melakukan generalisasi dengan baik [17]. Hal ini menunjukkan bahwa model telah mencapai kondisi konvergen dan mampu mempertahankan performa yang konsisten pada data yang tidak digunakan selama proses pelatihan.

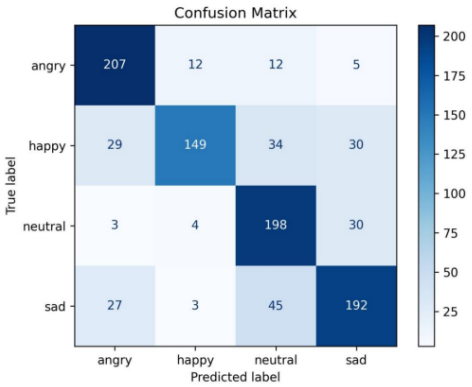


Gambar 5. Hasil Grafik *Validation F1-Score*

Pada Gambar 5 menunjukkan perubahan nilai *F1-score* pada data validasi selama proses pelatihan. Pola peningkatan *F1-score* terlihat sejalan dengan grafik akurasi, dengan nilai maksimum mendekati 0,79.

Keselarasan antara nilai akurasi dan *F1-score* menunjukkan bahwa performa model tidak didominasi oleh satu kelas emosi tertentu, melainkan relatif seimbang pada seluruh kelas. Hal ini menandakan bahwa model mampu melakukan klasifikasi emosi suara secara proporsional meskipun terdapat variasi jumlah data pada setiap kelas.

4.3. Analisis Performa



Gambar 6. *Confusion Matrix* Hasil Klasifikasi Emosi

Analisis performa model dilakukan menggunakan *confusion matrix* untuk mengetahui pola klasifikasi dan kesalahan prediksi pada setiap kelas emosi. Berdasarkan Gambar 6, *confusion matrix* menunjukkan bahwa model mampu mengklasifikasikan emosi dengan baik, ditandai oleh dominasi prediksi benar pada diagonal matriks. Kelas *angry* memiliki tingkat pengenalan tertinggi dengan 207 prediksi benar, menunjukkan bahwa karakteristik akustik emosi *angry* yang berintensitas tinggi dapat dikenali secara efektif oleh model.

Pada kelas *happy*, model menghasilkan 149 prediksi benar, namun masih terjadi kesalahan klasifikasi ke kelas *neutral* (34) dan *sad* (30) akibat kemiripan karakteristik suara. Kesalahan paling menonjol terjadi antara kelas *neutral* dan *sad*, dengan 30 data *neutral* diprediksi sebagai *sad* dan 45 data *sad* diprediksi sebagai *neutral*, yang mencerminkan kemiripan prosodi kedua emosi tersebut. Secara keseluruhan, pola kesalahan yang terjadi masih tergolong wajar dan menunjukkan bahwa model memiliki performa yang stabil serta kemampuan generalisasi yang baik.

4.4. Implementasi Game *Unity*

Gambar 7. Tampilan utama *visual game*

Gambar 7 menunjukkan tampilan antarmuka utama *game* yang dikembangkan menggunakan *Unity*. Pada tampilan ini, pemain dapat berinteraksi dengan karakter virtual yang berperan sebagai agen responsif terhadap emosi suara pemain.

Gambar 8. *Mic Recorder*

Pada Gambar 8 terdapat modul perekaman suara yang dilengkapi dengan indikator *Voice Activity Detection* (VAD) yang berfungsi untuk mendeteksi keberadaan suara sebelum dilakukan analisis emosi. Modul ini menampilkan visualisasi gelombang suara serta status proses analisis emosi secara *real-time*.

Gambar 9. Aktor kucing beserta panel *Affection Meter* dan tombol *Action*

Gambar 9 menampilkan karakter utama sebagai representasi visual sistem yang dilengkapi *Affection Meter* untuk menunjukkan tingkat kedekatan emosional. Perubahan nilai afeksi diatur menggunakan logika *rule-based* berdasarkan hasil deteksi emosi suara dengan ambang kepercayaan 50%.

Emosi *happy* dengan *confidence* di atas ambang batas akan meningkatkan afeksi (+5 poin) dan memicu animasi respons positif, sedangkan emosi *Angry* akan menurunkan afeksi (–5 poin) dan memicu reaksi ketakutan. Selain respons berbasis suara, *game* juga menyediakan interaksi manual seperti elus, main, dan makan untuk meningkatkan afeksi, sehingga menciptakan pengalaman bermain yang adaptif dan interaktif.

Gambar 10. *Panel UI* informasi deteksi emosi

Pada Gambar 10 ditampilkan panel informasi hasil deteksi emosi suara. Panel ini menunjukkan emosi yang terdeteksi beserta nilai tingkat kepercayaan (*confidence score*) untuk setiap kelas emosi, sehingga pemain dapat mengetahui hasil analisis emosi secara langsung.

4.5. Studi Kasus Pengujian Sistem

Pengujian sistem dilakukan menggunakan aktor manusia (satu laki-laki dan satu perempuan) sebagai sumber suara untuk mensimulasikan penggunaan sistem dalam kondisi nyata. Aktor mengucapkan kalimat dengan emosi tertentu yang telah ditentukan. Proses perekaman suara dilakukan menggunakan mikrofon BM-800 Condenser, sedangkan sistem dijalankan pada perangkat komputer dengan spesifikasi Intel Core i5 Gen 12, RAM 32 GB, dan GPU NVIDIA RTX 4060. Pengujian ini bertujuan untuk mengamati kinerja sistem *Speech Emotion Recognition* terhadap variasi karakteristik suara (*gender*) di luar lingkungan data pelatihan.

Tabel 2. Tahap Pengujian Halaman Beranda

No	Kelamin Kator	Kalimat	Emosi Aktor	Emosi Terdeteksi	Confidence (%)	Emosi Lain (%)
1.	laki-laki	"Kucing Nakal"	*Angry*	*Angry*	58,9	*Happy* : 12,2 *Neutral* : 16,3 *Sad* : 12,6
2.	laki-laki	"Aku capek dan sedih"	*Sad*	*Sad*	78,6	*Angry* : 1,5 *Happy* : 11,7 *Neutral* : 8,2
3.	laki-laki	"Wah hebat sekali"	*Happy*	*Happy*	54,9	*Angry* : 8,6 *Neutral* : 17,0 *Sad* : 19,5
4.	laki-laki	"Yaudah deh"	*Neutral*	*Happy*	84,8	*Angry* : 2,6 *Neutral* : 6,0 *Sad* : 6,7
5.	perempuan	"Kucing Nakal"	*Angry*	*Happy*	78,1	*Angry* : 8,9 *Happy* : 10,9 *Neutral* : 2,1
6.	perempuan	"Aku capek dan sedih"	*Sad*	*Sad*	91,8	*Angry* : 0,4 *Happy* : 4,0 *Neutral* : 3,8
7.	perempuan	"Wah hebat sekali"	*Happy*	*Happy*	69,1	*Angry* : 4,1 *Neutral* : 2,4 *Sad* : 24,5
8.	perempuan	"Yaudah deh"	*Neutral*	*Happy*	87,3	*Angry* : 2,1 *Neutral* : 7,2 *Sad* : 3,3

Berdasarkan Tabel 2, pengujian pada subjek laki-laki dan perempuan menunjukkan bahwa sistem mampu mengenali emosi *Angry*, *Sad,* dan *Happy* dengan cukup baik dalam kondisi penggunaan nyata. Emosi *Sad* terdeteksi paling konsisten pada kedua *gender* dengan tingkat kepercayaan tertinggi, yaitu 78,6% pada subjek laki-laki dan 91,8% pada subjek perempuan, yang menunjukkan kemampuan model dalam menangkap karakteristik prosodi kesedihan.

Namun, masih ditemukan kesalahan klasifikasi, terutama pada emosi *Neutral* yang terdeteksi sebagai *Happy* dengan *confidence* tinggi (84,8% pada laki-laki dan 87,3% pada perempuan), serta pada emosi *Angry* pada suara perempuan yang terklasifikasi sebagai *Happy* dengan confidence 78,1%. Kesalahan ini diduga disebabkan oleh kemiripan karakteristik akustik emosi berenergi tinggi (*high arousal)* dan variasi *pitch* suara berdasarkan *gender*. Secara keseluruhan, studi kasus ini menunjukkan bahwa model memiliki performa yang baik, namun masih memerlukan peningkatan untuk mengurangi bias prediksi dan meningkatkan *robustnes* terhadap variasi karakteristik suara.

5. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan beberapa poin utama terkait pengembangan dan implementasi sistem *Speech Emotion Recognition* (SER) berbasis *Wav2Vec 2.0* pada game *Unity* sebagai berikut:

1. Penelitian ini berhasil mengembangkan sistem *Speech Emotion Recognition* (SER) berbasis *Wav2Vec 2.0* yang terintegrasi dengan *game Unity* dan mampu mendeteksi emosi suara pemain secara *real-time* untuk menghasilkan respons emosional karakter yang adaptif.
2. Model SER yang dilatih menggunakan dataset CREMA-D dengan empat kelas emosi (*angry, happy, neutral,* dan *sad*) menunjukkan performa yang baik dengan akurasi validasi mendekati 79% serta *weighted F1-score* sebesar 0,79, yang menandakan kemampuan klasifikasi emosi suara yang cukup stabil.
3. Implementasi sistem SER pada g*ame Unity* memungkinkan karakter kucing memberikan respons emosional yang lebih natural dan interaktif, sehingga mampu meningkatkan kualitas interaksi dan imersi pemain dalam permainan.
4. Kelebihan dari sistem yang dikembangkan adalah pemanfaatan *Wav2Vec 2.0* berbasis *self-supervised learning,* yang mampu mengekstraksi representasi suara secara efektif tanpa memerlukan ekstraksi fitur manual, serta kemudahan integrasi model ke dalam *Unity* melalui format *ONNX*.
5. Keterbatasan penelitian ini terletak pada bias deteksi emosi tertentu, khususnya pada kesalahan klasifikasi antara emosi *neutral* dan *sad*, serta sensitivitas model terhadap variasi karakteristik suara seperti perbedaan *gender* dan intonasi.
6. Pengembangan selanjutnya dapat dilakukan dengan menambah jumlah kelas emosi, menggunakan dataset yang lebih beragam, meningkatkan teknik *augmentasi* dan normalisasi *audio*, serta mengoptimalkan model agar lebih robust terhadap variasi suara

dan kondisi lingkungan permainan yang lebih kompleks.

DAFTAR PUSTAKA

[1] S. Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intell. Syst. with Appl.*, vol. 20, no. September 2022, p. 200266, 2023, doi: 10.1016/j.iswa.2023.200266.

[2] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors (Switzerland)*, vol. 21, no. 4, pp. 1–27, 2021, doi: 10.3390/s21041249.

[3] Z. He, "Research Advanced in Speech Emotion Recognition based on Deep Learning," *Theor. Nat. Sci.*, vol. 86, no. 1, pp. 45–52, 2025, doi: 10.54254/2753-8818/2025.20333.

[4] A. Agustani, H. Setiawan, and T. Tasmi, "Analisis Perilaku Pengguna Terhadap Akses Internet Di Pt Chiyoda International Indonesia Menggunakan Machine Learning," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, pp. 1807–1814, 2025, doi: 10.23960/jitet.v13i3.6594.

[5] V. Tundjungsari, "Dasar Machine Learning_v.3.0_FULL ISBN," 2024, *Deepublish.*

[6] M. Ichsan, "Machine Learning Deteksi Penyakit Pada Kucing Menggunakan," vol. 13, no. 3, 2000.

[7] S. Sunitha, "An overview of deep learning," *Informatics Med. Unlocked*, vol. 26, no. 5, pp. 300–303, 2021, doi: 10.1016/j.imu.2021.100723.

[8] M. R. S. Alfarizi, M. Z. Al-farish, M. Taufiqurrahman, G. Ardiansah, and M. Elgar, "Use of Python as a Programming Language for Machine Learning and Deep Learning," *Sci. Work Students Belief Monoth. (KARIMAH TAUHID)*, vol. 2, no. 1, pp. 1–6, 2023.

[9] S. SONI and D. NANDAN, "A Comprehensive Review of Machine Learning and Deep Learning Methods and Applications," *Int. J. Multidiscip. Res.*, vol. 7, no. 4, 2025, doi: 10.36948/ijfmr.2025.v07i04.56249.

[10] S. Rahmadani, Cicih Sri Rahayu, Agus Salim, and Karno Nur Cahyo, "Deteksi Emosi Berdasarkan Wicara Menggunakan Deep Learning Model," *J. Inform. Teknol. dan Sains*, vol. 4, no. 3, pp. 220–224, 2022, doi: 10.51401/jinteks.v4i3.1952.

[11] P. K. S. Raja and P. D. D. Sanghani, "Speech Emotion Recognition Using Machine Learning," *Educ. Adm. Theory Pract.*, vol. 30, no. 6, pp. 118–124, 2024, doi: 10.53555/kuey.v30i6(s).5333.

[12] D. Sriharsha, C. A. Reddy, P. K. Kumar, and R. S. Kumar, "Speech Emotion Recognition," *IET Conf. Proc.*, vol. 2021, no. 11, pp. 55–59, 2021, doi: 10.1049/icp.2022.0313.

[13] D. Ferdiansyah, C. Sri, and K. Aditya, "Implementasi Automatic Speech Recognition Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper," vol. 11, no. 1, pp. 0–5, 2024.

[14] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 551–555, 2021, doi: 10.21437/Interspeech.2021-703.

[15] M. Kodali, S. R. Kadiri, and P. Alku, "Classification of Vocal Intensity Category from Speech using the Wav2vec2 and Whisper Embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-Augus, pp. 4134–4138, 2023, doi: 10.21437/Interspeech.2023-2038.

[16] A. NOERCHOLIS, T. DWIANDINI, and F. S. MUKTI, "Optimasi Teknologi WAV2Vec 2.0 menggunakan Spectral Masking untuk meningkatkan Kualitas Transkripsi Teks Video bagi Tuna Rungu," *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, vol. 12, no. 4, p. 877, 2024, doi: 10.26760/elkomika.v12i4.877.

[17] N. H. Muttaqin and A. M. Widodo, "Evaluation of Transfer Learning-Based Convolutional Neural Networks (InceptionV3 and MobileNetV2) for Facial Skin-Type Classification," *J. Ilmu Komput. dan Inform.*, vol. 5, no. 1, pp. 11–32, 2025, doi: 10.54082/jiki.264.