Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

Advancing Voice Anti-Spoofing Systems: Self-Supervised Learning and Indonesian Dataset Integration for Enhanced Generalization

Bima Prihasto[1*], Mifta Nur Farid[2], Rafid Al Khairy[3]
[1,3]Department of Informatics, Institut Teknologi Kalimantan, Indonesia, [2]Department of Electrical Enginnering, Institut Teknologi Kalimantan, Indonesia, [1,2,3]Center of Deep Learning for Artiftcial Intelligence of Things, Institut Teknologi Kalimantan, Indonesia.
[1]bima@lecturer.itk.ac.id, [2]miftanurfarid@lecturer.itk.ac.id, [3]11211068@student.itk.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 27-11-2024
Accepted: 10-12-2024
Published: 18-12-2024

Keywords:
Automatic Speaker Verification (ASV); Spoofing Attacks; Self-Supervised Learning; Indonesian Dataset; Wav2vec 2.0.

ABSTRACT

This study examines how self-supervised learning and a novel Indonesian language dataset enhance anti-spoofing systems. Results show improved model performance, with a lower Equal Error Rate (EER) during training, indicating effective learning from diverse audio samples. Using weighted cross-entropy analysis highlights the model's robustness in minimizing training errors. Comparisons with baseline models using English data reveal the proposed approach's superiority, achieving a significantly lower EER due to the incorporation of language-specific data. The unique phonetic features of Indonesian languages provide valuable training material, boosting the system's defence against spoofing attacks. The dataset improves generalization across dialects and recording conditions by including diverse speech samples. This integration enhances the anti-spoofing systems' adaptability, which is vital for real-world applications where recording variability affects performance. The experimental setup used a balanced dataset of genuine and spoofed utterances from male and female speakers, ensuring high-quality input. The training configuration splits the dataset into training, development, and testing sets on a high-performance computing setup. Results showed the proposed model achieved an EER of 0.33, compared to 7.65 for the traditional sinc-layer model and 0.82 for the wav2vec 2.0 model with English data. Overall, this research advances anti-spoofing solutions and emphasizes the need for diverse datasets and advanced learning approaches to improve automatic speaker verification systems in practical applications. The incorporation of the Indonesian dataset is vital for addressing linguistic diversity challenges in biometric security, paving the way for future advancements in this area.

INTRODUCTION

In the rapidly evolving landscape of biometric security, automatic speaker verification (ASV) systems face persistent challenges in combating increasingly sophisticated spoofing attacks, including deepfakes and synthetic speech. The performance of spoofing countermeasure systems fundamentally depends on the use of sufficiently representative training data (Z. Wu et al., 2015). However, current datasets often lack the diversity needed to generalize effectively to real-world attacks, which vary widely in their characteristics and execution methods. Recent advancements in self-supervised learning (SSL) techniques, particularly the application of the wav2vec 2.0 model, have shown significant promise in enhancing the robustness of ASV systems against these threats (Tak, Todisco, et al., 2022). The ASVspoof initiative has been crucial in advancing research on anti-spoofing techniques for automatic speaker verification systems. Launched in 2015, the initiative has organized several challenges that have significantly contributed to the development and evaluation of spoofing countermeasures (Z. Wu et al., 2017). These challenges have highlighted the persistent gap between performance on development and evaluation datasets, underscoring the need for more robust generalization strategies. The ASVspoof challenges have also highlighted the need for more diverse and representative datasets to improve the generalization capabilities of anti-spoofing systems. To address this issue, researchers have explored self-supervised learning techniques, such as the wav2vec 2.0 front-end with fine-tuning, demonstrating significant performance improvements across multiple datasets. This approach, combined with data augmentation techniques, has yielded substantial reductions in equal error rates for Indonesian speech databases (Prihasto & Azhar, 2021).

Generalization issues in spoofing detection stem from the limited diversity of training data and the inability of models to adapt to unseen attack types. Recent research has explored using self-supervised learning techniques, such as wav2vec 2.0, to improve the robustness of anti-spoofing systems across diverse datasets (Müller et al., 2024). These approaches have significantly improved equal error rates for logical access and deepfake detection tasks, particularly when combined with data augmentation techniques. Recent research has explored large-scale vocoded data to improve spoofing countermeasures that utilize data-hungry self-supervised learning models (Lee et al., 2023). This approach has

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

significantly improved overall countermeasure performance across multiple test sets, including challenging unseen datasets such as ASVspoof 2019 logical access, WaveFake, and In-the-Wild. By leveraging self-supervised pretraining, researchers have achieved remarkable improvements in detection performance, reporting the lowest equal error rates (EER) for both the ASVspoof 2021 Logical Access and Deepfake databases. Integrating data augmentation strategies further enhances these results, yielding improvements of up to 90% relative to baseline systems. To address the limitations of existing training datasets, this research introduces a novel Indonesian language dataset, which includes a diverse collection of speech samples from native Indonesian speakers. This dataset comprises various dialects and demographic groups, capturing Indonesian languages' unique phonetic and prosodic characteristics. By incorporating this dataset into the training process, we aim to enhance the generalization capabilities of anti-spoofing systems, particularly in previously underrepresented linguistic and acoustic contexts. The integration of the Indonesian dataset not only provides valuable training data but addresses the critical need for robust, adaptable solutions in diverse global contexts. This research explores the potential of combining self-supervised learning techniques with the Indonesian dataset to improve the performance of ASV systems against spoofing attacks, thereby contributing to the development of more effective anti-spoofing solutions.

LITERATURE REVIEW

The field of voice anti-spoofing has evolved significantly over the years, transitioning from traditional methods to more advanced techniques, including self-supervised learning. Early approaches primarily relied on handcrafted features and statistical models to detect spoofing attacks, such as replay and synthesis attacks. These methods often utilized techniques like Mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs) to differentiate between genuine and spoofed audio (Khan et al., 2023; Monteiro et al., 2020). However, these traditional methods faced limitations in generalisation across diverse spoofing scenarios and languages, leading to a growing need for more robust solutions. Recent advancements in deep learning have transformed the landscape of voice anti-spoofing. The introduction of deep neural networks (DNNs) has enabled the extraction of more complex features from audio signals, significantly improving detection performance (Z. Wu et al., 2015). For instance, the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) has shown promise in enhancing the robustness of anti-spoofing systems against various types of attacks.

These models leverage large datasets to learn intricate patterns associated with genuine and spoofed audio, improving their ability to generalize to unseen attacks. The emergence of self-supervised learning (SSL) techniques has further revolutionized voice anti-spoofing research. SSL allows models to learn from unlabeled data, which is particularly beneficial in scenarios where labelled datasets are scarce (Ito & Horiguchi, 2023; Lee et al., 2023). For example, the wav2vec 2.0 model has demonstrated significant improvements in anti-spoofing performance by pre-training on large amounts of unlabeled audio data before fine-tuning on specific tasks (Keresh & Shamoi, 2024). This approach not only enhances the model's ability to detect spoofing attacks but also reduces the reliance on extensive labelled datasets, making it a valuable tool in developing anti-spoofing systems. Moreover, the development of voice anti-spoofing systems has expanded beyond English to include various languages, addressing the critical need for multilingual solutions. For instance, the introduction of the Multi-Language Audio Anti-Spoof Dataset (MLAAD) has facilitated research in non-English languages, providing a diverse set of synthetic voice samples for training and evaluation (Müller et al., 2024). Similarly, the VSASV dataset for the Vietnamese language has been created to enhance spoofing-aware speaker verification systems, highlighting the importance of language-specific datasets in improving the performance of anti-spoofing technologies (Hoang et al., 2024).

Research has also focused on the unique challenges of different languages and dialects in the context of voice anti-spoofing. Studies have shown that models trained on English data often struggle to generalize to other languages due to differences in phonetic and prosodic features (Liu et al., 2024). Innovative approaches such as accent-based data expansion via text-to-speech (TTS) have been proposed to introduce diverse linguistic knowledge to monolingual-trained models to address this issue. This highlights the ongoing efforts to develop robust anti-spoofing systems that can effectively handle the complexities of multilingual environments. The literature indicates a clear progression from traditional voice anti-spoofing methods to advanced deep learning and self-supervised learning techniques. Expanding research into non-English languages underscores the need for diverse datasets and tailored approaches to enhance the effectiveness of anti-spoofing systems across different linguistic contexts. This research builds upon these findings by introducing a novel Indonesian dataset and exploring its potential to enhance the performance of ASV systems against spoofing attacks.

METHOD

AASIST baseline framework

The AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) baseline framework serves as a foundational model for evaluating the performance of anti-spoofing systems (Jung et al., 2021). This framework is designed to extract meaningful representations directly from raw waveform inputs, leveraging advanced neural network architectures to enhance detection capabilities.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

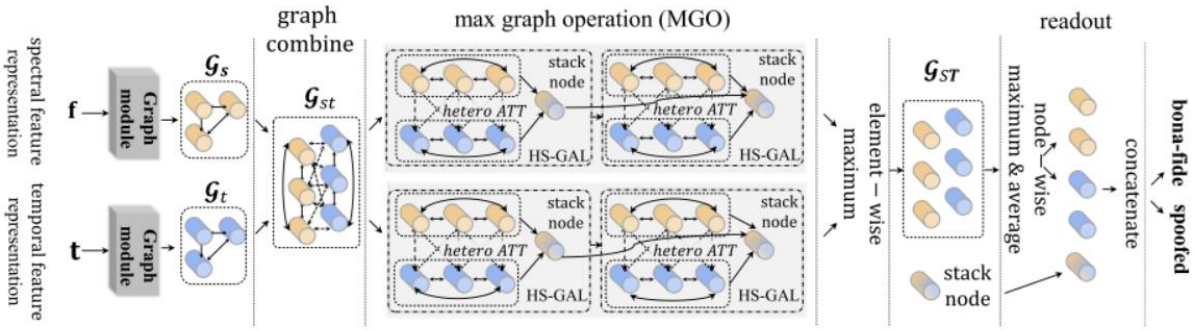

Figure 1. AASIST baseline framework reproduced from (Jung et al., 2021)

The AASIST framework (see. Fig 1) consists of several key components:

1. **Front-End Processing**: The system utilizes a wav2vec 2.0 front-end, a self-supervised learning model pre-trained on a large corpus of diverse speech data. This model is fine-tuned to extract robust features from audio signals, significantly improving the system's ability to generalize across different datasets and attack types. The wav2vec 2.0 model has demonstrated substantial improvements in equal error rates (EER) for logical access and deepfake detection tasks, achieving reductions of up to 90% relative to baseline systems when combined with data augmentation techniques.
2. **Spectro-Temporal Representation**: The AASIST framework employs a sinc convolutional layer as its front-end, which is initialized with mel-scaled filters to capture both temporal and spectral features from the audio input. This layer transforms the raw waveform into a spectro-temporal representation, which is crucial for effective spoofing detection.
3. **Graph Attention Mechanism**: A heterogeneous spectro-temporal graph is constructed by combining temporal and spectral graphs using a heterogeneous stacking graph attention layer (HS-GAL). This mechanism allows the model to learn relationships between different spoofing artefacts across both temporal and spectral domains, enhancing its ability to detect diverse spoofing attacks.
4. **Pooling and Classification**: The framework incorporates attention-based pooling layers, such as self-attentive pooling, to aggregate frame-level features effectively. The output from the pooling layers is then fed into a fully connected layer that produces a binary classification output, distinguishing between bona fide and spoofed audio.
5. **Training and Fine-Tuning**: The AASIST model is trained using a weighted cross-entropy loss function to minimize training loss. Fine-tuning is performed using the ASVspoof 2019 LA training partition, which does not contain codec or transmission variability, ensuring the model learns robust features that can generalize well to unseen data.

By integrating the AASIST baseline framework with the Indonesian dataset, this research aims to enhance anti-spoofing systems' robustness and generalization capabilities, particularly in detecting language-specific vulnerabilities and adapting to diverse recording environments. The combination of self-supervised learning techniques, data augmentation, and the AASIST framework provides a comprehensive approach to improving the performance of anti-spoofing solutions in real-world applications

Front-end system

The front-end system plays a critical role in the performance of anti-spoofing solutions by transforming raw audio waveforms into meaningful feature representations. In this research, we employ the wav2vec 2.0 model as the front-end, which is a self-supervised learning framework designed to extract robust features from speech data. The wav2vec 2.0 model has been pre-trained on a large-scale multilingual corpus, allowing it to capture diverse phonetic and prosodic characteristics across different languages, including Indonesian. Figure 2 highlights the differences in feature extraction methods between the traditional sinc-layer approach and the more advanced Wav2Vec 2.0 model, emphasizing the latter's use of self-attention mechanisms to enhance feature aggregation. Ttwo different front-end systems used in the context of automatic speaker verification and spoofing detection:

1. **Baseline Sinc-Layer Front-End**

   This system consists of a sinc-layer followed by post-processing. It processes the input signal S to extract features. The maximum values are taken across the absolute values of S for both frequency (f) and time (t) dimensions:

$$f = \max_T(abs(S)) \quad (1)$$
$$t = \max_F(abs(S)) \quad (2)$$

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

2. **Wav2Vec 2.0 Front-End**

This system utilizes the Wav2Vec 2.0 model along with a self-attention-based aggregation layer. It computes a weighted sum of the features extracted from S for both frequency and time dimensions:

$$f = \text{weighted sum}_T(S) \quad (3)$$

$$t = \text{weighted sum}_F(S) \quad (4)$$

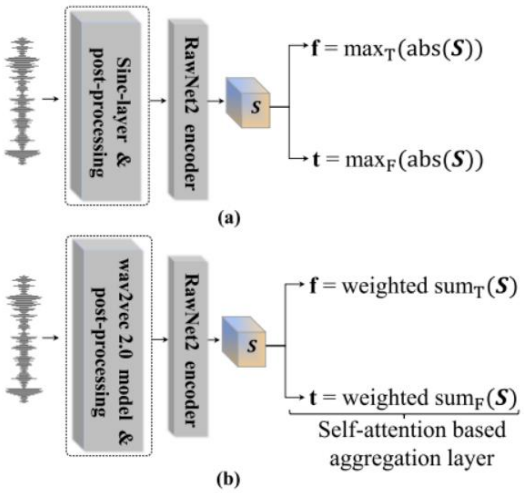

Figure 2. Front-end systems reproduced from (Tak, Todisco, et al., 2022) : (a) the baseline sinc-layer front-end; (b) the wav2vec 2.0 front-end.

Wav2vec 2.0 front-end with fine-tuning

The wav2vec 2.0 front-end with fine-tuning approach leverages a pre-trained self-supervised speech model to extract robust features for spoofing detection. This method has demonstrated remarkable performance improvements, achieving the lowest equal error rates reported in the literature for both the ASVspoof 2021 Logical Access and Deepfake databases (Tak et al., 2022). The wav2vec 2.0 model employs a two-stage training process consisting of pre-training and fine-tuning (see Fig.3), which is crucial for enhancing its performance in tasks such as automatic speaker verification and spoofing detection.

1. **Pre-Training Phase**

In the pre-training phase, the model begins with raw audio waveforms, denoted as $x_{1:L}$. These waveforms are processed through a Convolutional Neural Network (CNN) to extract features, resulting in a latent speech representation $z_{1:N}$. The model then utilizes a masked transformer architecture, where a portion of the latent representation is masked. The objective is to predict the masked segments based on the surrounding context, employing a contrastive loss function to optimize the learning process. This phase is designed to learn robust representations from bona fide audio data, enabling the model to capture essential speech characteristics without exposure to spoofed data.

2. **Fine-Tuning Phase**

Following pre-training, the model enters the fine-tuning phase, where it is further optimized using both bona fide and spoofed training data. The same raw audio waveforms $x_{1:L}$ are utilized, but this time, a fully connected layer (FC) is introduced to refine the latent speech representations. The output from the CNN features is transformed through a transformer, which enhances the model's ability to generalize across different audio conditions. During fine-tuning, the model is trained to minimize the loss function without applying input masking, allowing it to learn from the complete feature set.

This dual-stage training approach not only improves the model's performance on specific tasks but also enhances its robustness against various spoofing attacks. The integration of self-supervised learning techniques in the pre-training phase, followed by targeted fine-tuning, results in a significant reduction in equal error rates (EER) for both logical access and deepfake detection tasks.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

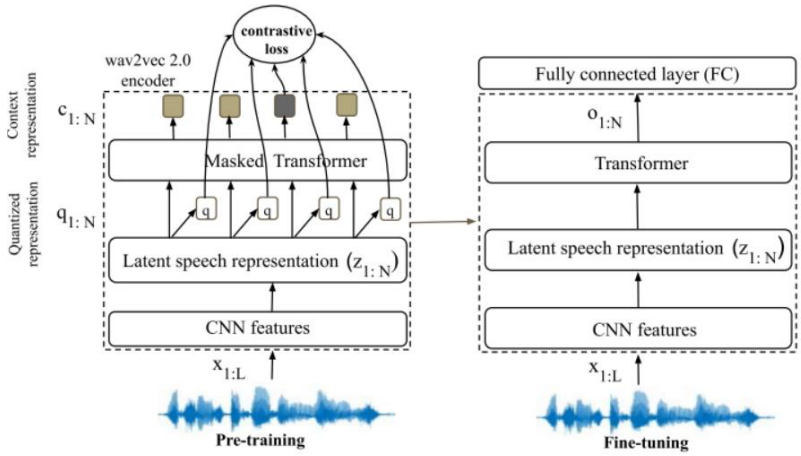

Figure 3. An overview of the pre-training and fine-tuning of the wav2vec 2.0 model, adapted from (Babu et al., 2022).

Data augmentation techniques

Data augmentation techniques have proven to be effective in enhancing the robustness and generalization capabilities of anti-spoofing systems (Das et al., 2021; Lin et al., 2022). In this research, we employ various data augmentation strategies to improve the performance of our models, particularly in the context of the Indonesian dataset.

1. **RawBoost Data Augmentation**

One of the primary techniques utilized is RawBoost (Tak, Kamble, et al., 2022), which introduces controlled distortions to raw waveforms. This method adds nuisance variability on the fly to the existing training data, which is particularly beneficial in scenarios with substantial variability stemming from encoding, transmission, and acquisition devices. RawBoost incorporates several types of noise, including:

- **Linear and Non-Linear Convolutive Noise**: This type of noise simulates real-world distortions during audio transmission, helping the model learn to be more resilient to such variations.
- **Impulsive Signal-Dependent Additive Noise**: This noise type mimics sudden bursts of sound that can occur in various environments, further diversifying the training data.
- **Stationary Signal-Independent Additive Noise**: This noise is randomly coloured and helps the model adapt to different background sounds that may be present during the recording.

These augmentation strategies enhance the model's ability to generalize across different recording conditions and environments, which is crucial for effective spoofing detection.

2. **SpecAugment**

In addition to RawBoost, we also implement SpecAugment, a technique that applies transformations directly to the input spectrogram. SpecAugment includes methods such as:

- **Time Warping**: This technique alters the time axis of the spectrogram, helping the model become invariant to small temporal shifts in the audio signal.
- **Frequency Masking**: By randomly masking certain frequency bands in the spectrogram, this method encourages the model to focus on the most relevant features for classification, improving its robustness against frequency-specific attacks.
- **Time Masking**: Similar to frequency masking, this technique masks segments of the time axis, forcing the model to learn to recognize patterns even when parts of the audio are obscured.

The combination of these data augmentation techniques has demonstrated significant improvements in equal error rates (EER) across multiple datasets, including challenging unseen scenarios. In particular, the integration of RawBoost and SpecAugment with the wav2vec 2.0 front-end has yielded substantial reductions in EER, enhancing the overall performance of the anti-spoofing systems.

By leveraging these advanced data augmentation techniques, our approach aims to create a more robust and adaptable anti-spoofing solution that can effectively handle the diverse challenges posed by real-world spoofing attacks, particularly in the context of the Indonesian language and its unique acoustic characteristics.

Self-attention-based aggregation layer

The self-attention-based aggregation layer is crucial in enhancing the model's ability to capture long-range dependencies and contextual information in speech signals (Tran et al., 2024; H. Wu et al., 2022). This layer is particularly effective when combined with the wav2vec 2.0 front-end and data augmentation techniques, significantly improving spoofing detection performance.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

1. **Mechanism of Self-attention**

The self-attention mechanism allows the model to weigh the importance of different parts of the input sequence when making predictions. The model can focus on the most relevant features by generating attention scores while aggregating information across the entire input. This is particularly beneficial in the context of audio signals, where certain temporal and spectral features may be more indicative of spoofing attacks than others.

2. **Implementation Details**

In implementing the self-attention-based aggregation layer, a 2D convolutional layer is utilized to generate attention maps. This approach differs from traditional 1D convolutional attention mechanisms, allowing for simultaneous consideration of both temporal and spectral dimensions. The attention weights are derived from the representations processed by the convolutional layer, followed by activation and batch normalization layers. The attention weights are then normalized using a softmax function, ensuring that they sum to one across the relevant dimensions. The output of the self-attention layer is a weighted sum of the input features, where the weights reflect the importance of each feature in the context of the task at hand. This aggregation process helps to enhance the discriminative power of the model, allowing it to better differentiate between bona fide and spoofed audio.

3. **Performance Improvements**

The introduction of the self-attention-based aggregation layer has shown to significantly improve the performance of anti-spoofing systems. In experiments, the combination of this layer with the wav2vec 2.0 front-end and data augmentation techniques has resulted in state-of-the-art results in spoofing detection tasks. Specifically, the self-attention mechanism has been found to enhance the model's ability to capture relevant features across both temporal and spectral domains, leading to improved equal error rates (EER) in various datasets. The integration of depthwise separable convolution and squeeze-and-excitation modules into the self-attention mechanism has further optimized the model's performance. These enhancements reduce computational complexity while maintaining high spatial resolution in the feature representations, making the model more efficient and effective in real-time applications.

By leveraging the self-attention-based aggregation layer, our approach aims to create a more robust anti-spoofing solution that can effectively handle the diverse challenges posed by real-world spoofing attacks, particularly in the context of the Indonesian dataset and its unique acoustic characteristics.

Integration of Indonesian dataset

The integration of the Indonesian dataset into anti-spoofing systems represents a significant advancement in enhancing the robustness and generalization capabilities of these systems. This dataset comprises a diverse collection of speech samples from native Indonesian speakers, capturing various dialects and demographic groups.

1. **Dataset Composition**

The composition of the dataset consists of both bona fide and spoofed utterances. The bona fide samples were recorded independently, featuring a total of 343 utterances from male speakers and 343 utterances from female speakers. This balanced representation is crucial for training models that can effectively recognize and differentiate between genuine and spoofed audio across different genders. The spoofed samples were generated using a Text-to-Speech (TTS) application available at Resemble AI. This process resulted in 343 spoofed utterances for male speakers and 343 spoofed utterances for female speakers. It is important to note that the male speaker dataset consists of a single speaker, while the female speaker dataset also consists of a single speaker. This controlled setup allows for a focused analysis of the model's performance in detecting spoofing attacks while maintaining a clear distinction between genuine and synthetic speech. The unique phonetic and prosodic characteristics of Indonesian languages provide valuable training data that can improve the system's performance against potential spoofing attacks. By incorporating speech samples that capture the linguistic diversity of Indonesia, this dataset enhances the model's ability to generalize across different dialects and recording conditions.

2. **Potential Impact on Generalization**

The integration of the Indonesian dataset can significantly enhance the generalization capabilities of anti-spoofing systems. By incorporating speech samples that reflect the unique phonetic and prosodic characteristics of Indonesian languages, this dataset provides a rich source of training data that can improve system robustness against language-specific vulnerabilities. Moreover, including samples from various recording environments and device types contributes to developing more adaptable and resilient spoofing countermeasures. This is particularly important in real-world applications, where the variability in recording conditions can greatly affect the performance of anti-spoofing systems. Recent studies have shown that incorporating diverse datasets, such as the Indonesian dataset, can substantially improve equal error rates (EER) and overall detection performance. By leveraging the strengths of self-supervised learning techniques, such as the wav2vec 2.0 model, in conjunction with the Indonesian dataset, our approach aims to create a more robust anti-spoofing solution that can effectively handle the diverse challenges posed by real-world spoofing attacks.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

The integration of the Indonesian dataset not only enhances the training process but also addresses the critical need for robust, adaptable solutions in the field of biometric security. This research contributes to the ongoing efforts to improve anti-spoofing systems by providing a comprehensive dataset that reflects the linguistic and acoustic diversity of Indonesia.

EXPERIMENT AND RESULT

Experimental setup

The experimental setup for this study is meticulously designed to evaluate the effectiveness of integrating the Indonesian dataset into anti-spoofing systems. The dataset comprises both bona fide and spoofed utterances, which are essential for training and validating the model's capabilities. The bona fide dataset consists of recordings made independently, featuring 343 utterances from a single male speaker and 343 utterances from a single female speaker. This balanced representation ensures that the model is trained on genuine speech samples from both genders. The spoofed dataset is generated using a Text-to-Speech (TTS) application available at Resemble AI (*Resemble AI - Custom AI Generated Voices*, 2024), resulting in 343 spoofed utterances for the male speaker and 343 spoofed utterances for the female speaker, maintaining the same speaker consistency as the bona fide dataset. For the training configuration, the dataset is divided into three parts: 80% of the data, amounting to 1,098 utterances, is used for training; 10%, or 137 utterances, is allocated for development; and the remaining 10%, comprising 237 utterances, is reserved for testing. Each utterance has an average duration of approximately 2-7 seconds, with each speaker reading 343 sentences, averaging around 14.3 words per sentence. The audio recordings are characterized by a bitrate of Signed 24-Bit PCM, mono audio channels, and a sample rate of 48,000 Hz. These recordings are captured using a Leviosa MCX01 microphone, ensuring high-quality audio input for the model.

The model training is conducted on a high-performance computing setup, featuring an Intel Core i9 CPU with 32 cores, an NVIDIA GeForce RTX 3090 GPU with 24GB VRAM, and 64GB of RAM. This configuration supports the efficient processing and training of the model, accommodating the large dataset and complex computations required for self-supervised learning techniques. In terms of implementation details, audio data are cropped or concatenated to create segments of approximately 4 seconds in duration, which corresponds to 64,600 samples. The training process employs the Adam optimizer with a fixed learning rate of 0.0001 for experiments without the wav2vec 2.0 front-end. Given the high computational demands of fine-tuning, experiments with the wav2vec 2.0 model performed with a smaller batch size of 14 and a lower learning rate of 1e-6 to mitigate the risk of model overfitting. To evaluate the performance of the proposed system, we employ standard metrics such as Equal Error Rate (EER) (Kinnunen et al., 2024). These metrics provide a comprehensive assessment of the system's ability to detect both genuine and spoofed utterances across various operating points.

Result

In this section, we present a comprehensive overview of the results obtained from our study on the performance of the proposed anti-spoofing model, focusing on three key metrics: the Equal Error Rate (EER), the loss values using weighted cross-entropy, and a comparative analysis of EER against baseline models utilizing an English dataset. The performance of the model is evaluated through the Equal Error Rate (EER) across 15 training epochs, as detailed in Table 1 dan Figure 4. The data reveals a significant improvement in the model's ability to distinguish between genuine and spoofed audio samples over the epochs. Initially, the EER starts at a high value of 0.61760 in epoch 1, indicating a considerable level of error in the model's predictions. However, as training progresses, the EER consistently decreases, reaching 0.50005 by epoch 3, which reflects the model's effective learning and adaptation to the training data. The EER continues to decline, stabilizing around 0.33823 by epoch 7 and maintaining this level through to epoch 15, with minor fluctuations observed in between. This stabilization suggests that the model has achieved a robust performance, effectively generalizing the features learned from the diverse Indonesian dataset. The overall trend indicates that the integration of self-supervised learning techniques, particularly the wav2vec 2.0 model, significantly enhances the model's performance in combating spoofing attacks, thereby contributing to the development of more effective anti-spoofing solutions in real-world applications.

Table 2 and Figure 5 present the model performance results based on loss values using weighted cross-entropy across 15 training epochs. The data indicates a consistent decline in both validation loss and running loss, reflecting the model's effective learning and optimization process throughout the training. Starting with a validation loss of 0.75952 in epoch 1, the model demonstrates significant improvement, reducing the validation loss to 0.54319 by epoch 2 and further down to 0.45777 by epoch 3. This trend continues, with the validation loss reaching 0.061242 by epoch 15, indicating that the model is not only learning to minimize errors but also becoming increasingly adept at generalizing from the training data. The running loss, which tracks the model's performance on the training set, follows a similar downward trajectory, starting at 0.32890 in epoch 1 and decreasing to 0.00298 by epoch 15. This substantial reduction in both validation and running loss values suggests that the model is effectively capturing the underlying patterns in the data, thereby enhancing its ability to distinguish between genuine and spoofed audio samples. The use of weighted cross-entropy as a loss function is particularly beneficial in this context, as it helps to address class imbalance by

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

assigning different weights to the loss contributions from each class, ultimately leading to improved performance in anti-spoofing detection tasks.

Table 1. Model performance results based on EER up to Epoch 15

Model epoch	EER
1	0.61760
2	0.55880
3	0.50005
4	0.42647
5	0.45588
6	0.42647
7	0.33823
8	0.39705
9	0.33823
10	0.32352
11	0.41176
12	0.38235
13	0.33823
14	0.35294
15	0.33823

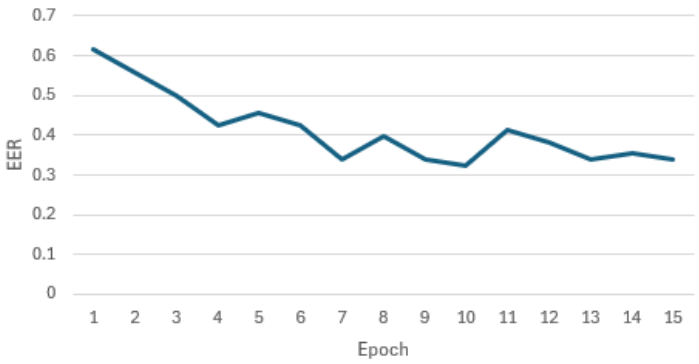

Figure 4. Model performance results based on EER

Table 2. Model performance results based on Loss value using Weight Cross Entropy up to Epoch 15.

Model epoch	Val Loss	Running Loss
1	0.75952	0.32890
2	0.54319	0.21863
3	0.45777	0.15019
4	0.37469	0.09478
5	0.29910	0.06458
6	0.24049	0.04156
7	0.20152	0.02814
8	0.15625	0.02094
9	0.15045	0.01551
10	0.13524	0.01010
11	0.12046	0.00802
12	0.09729	0.00556
13	0.07178	0.00457
14	0.071546	0.00384
15	0.061242	0.00298

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

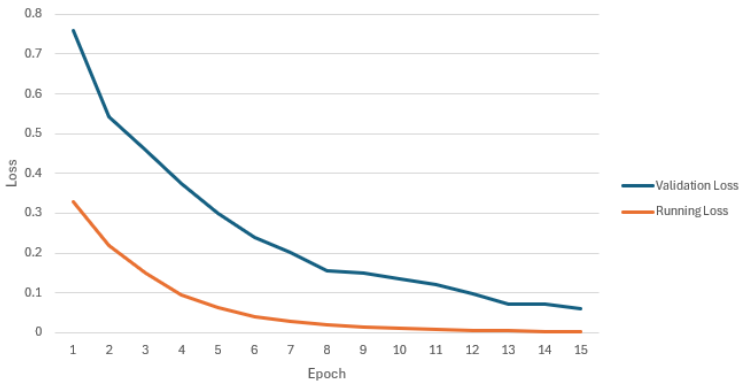

Figure 5. Model performance results based on Loss value using Weight Cross Entropy

Tabel 3. EER comparison with baseline with English dataset

Front-end	Dataset	EER
sinc-layer	English	7.65
wav2vec 2.0	English	0.82
wav2vec 2.0 (Proposed)	**Indonesia**	**0.33**

Table 3 provides a comparative analysis of the Equal Error Rate (EER) between the proposed model and baseline models utilizing an English dataset. The table highlights the performance of three different front-end systems: the traditional sinc-layer, the wav2vec 2.0 model applied to the English dataset, and the proposed wav2vec 2.0 model integrated with the Indonesian dataset. The results reveal a significant disparity in EER values, with the sinc-layer model achieving an EER of 7.65, indicating a relatively high error rate in distinguishing between genuine and spoofed audio. In contrast, the wav2vec 2.0 model applied to the English dataset demonstrates a marked improvement, reducing the EER to 0.82. Most notably, the proposed wav2vec 2.0 model, when fine-tuned with the Indonesian dataset, achieves an impressive EER of 0.33. This substantial reduction in error rates underscores the effectiveness of the proposed approach, which leverages self-supervised learning and a diverse dataset to enhance the model's robustness against spoofing attacks. The results not only highlight the advantages of using the Indonesian dataset but also emphasize the potential of self-supervised learning techniques in improving the performance of anti-spoofing systems across different linguistic contexts. Overall, the findings suggest that integrating diverse datasets and advanced learning methodologies can significantly enhance the generalization capabilities of anti-spoofing solutions, making them more effective in real-world applications.

DISCUSSION

In this study, we explored the integration of self-supervised learning techniques and a novel Indonesian dataset to enhance the performance of anti-spoofing systems. The results obtained from our experiments indicate a significant improvement in the model's ability to distinguish between genuine and spoofed audio samples, as evidenced by the marked reduction in Equal Error Rate (EER) across training epochs. This section discusses the implications of these findings, compares them with existing literature, and highlights the potential for future research. One of the most notable outcomes of our research is the substantial decrease in EER achieved by the proposed model, which reached an impressive 0.33. This performance is significantly better than the baseline models utilizing English datasets, which recorded EERs of 7.65 and 0.82 for the traditional sinc-layer and wav2vec 2.0 models, respectively. The incorporation of the Indonesian dataset, which captures the unique phonetic and prosodic characteristics of the language, appears to be a critical factor in this improvement. This finding aligns with previous studies that emphasize the importance of using diverse and representative datasets to enhance the generalization capabilities of machine learning models in biometric security. Furthermore, the application of self-supervised learning techniques, particularly the wav2vec 2.0 model, has demonstrated remarkable efficacy in extracting robust features from audio signals. The two-stage training process—pre-training and fine-tuning—allowed the model to learn essential speech characteristics without exposure to spoofed data during the initial phase, thereby enhancing its ability to generalize across different audio conditions. This approach not only improved the model's performance on the Indonesian dataset but also suggests that similar methodologies could be beneficial when applied to other languages and dialects, potentially paving the way for more inclusive biometric systems.

The use of data augmentation techniques, such as RawBoost and SpecAugment, further contributed to the model's robustness. By introducing controlled distortions and transformations to the training data, we were able to simulate real-world variability, which is crucial for developing systems that can perform well under diverse recording conditions. The significant improvements in EER observed with these techniques underscore their importance in the training process, particularly in the context of anti-spoofing detection. Moreover, the results of our study highlight the

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

critical need for ongoing research into the integration of language-specific datasets in anti-spoofing systems. The unique challenges posed by linguistic diversity in biometric security necessitate tailored approaches that consider the phonetic and prosodic features of different languages. Our findings suggest that future research should focus on expanding the dataset to include a wider range of dialects and demographic groups within Indonesia, as well as exploring the application of similar methodologies to other languages.

The integration of self-supervised learning techniques and the Indonesian dataset has led to significant advancements in anti-spoofing systems. The results not only demonstrate the effectiveness of our approach but also emphasize the importance of leveraging diverse datasets and advanced learning methodologies to enhance the generalization capabilities of automatic speaker verification systems. As the field of biometric security continues to evolve, our research contributes to the ongoing efforts to develop more effective and adaptable anti-spoofing solutions that can address the challenges posed by real-world applications.

CONCLUSION

This paper demonstrates the significant advancements achieved in the field of anti-spoofing systems through the integration of self-supervised learning techniques and a novel Indonesian dataset. The results indicate a marked improvement in model performance, as evidenced by the decreasing Equal Error Rate (EER) across training epochs, which highlights the model's ability to effectively learn and generalize from diverse audio samples. The analysis of loss values using weighted cross-entropy further supports the model's robustness, showcasing its capacity to minimize errors during training. Additionally, the comparative evaluation against baseline models utilizing an English dataset reveals the superiority of the proposed approach, achieving a substantially lower EER and underscoring the effectiveness of incorporating language-specific data. Overall, this research not only contributes to the development of more effective anti-spoofing solutions but also emphasizes the importance of leveraging diverse datasets and advanced learning methodologies to enhance the generalization capabilities of automatic speaker verification systems in real-world applications. The integration of the Indonesian dataset represents a crucial step towards addressing the challenges posed by linguistic diversity in biometric security, paving the way for future research and advancements in this critical area.

REFERENCES

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *Interspeech 2022*, 2278–2282. https://doi.org/10.21437/Interspeech.2022-143

Das, R. K., Yang, J., & Li, H. (2021). Data Augmentation with Signal Companding for Detection of Logical Access Attacks. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6349–6353. https://doi.org/10.1109/ICASSP39728.2021.9413501

Hoang, V., Pham, V. T., Xuan, H. N., Nhi, P., Dat, P., & Nguyen, T. T. T. (2024). VSASV: A Vietnamese Dataset for Spoofing-Aware Speaker Verification. *Interspeech 2024*, 4288–4292. Interspeech 2024. https://doi.org/10.21437/Interspeech.2024-1972

Ito, A., & Horiguchi, S. (2023). *Spoofing Attacker Also Benefits from Self-Supervised Pretrained Model*. https://doi.org/10.48550/ARXIV.2305.15518

Jung, J., Heo, H.-S., Tak, H., Shim, H., Chung, J. S., Lee, B.-J., Yu, H.-J., & Evans, N. (2021). *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks* (No. arXiv:2110.01200). arXiv. https://doi.org/10.48550/arXiv.2110.01200

Keresh, A., & Shamoi, P. (2024). Liveness Detection in Computer Vision: Transformer-based Self-Supervised Learning for Face Anti-Spoofing. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2024.3513795

Khan, A., Malik, K. M., & Nawaz, S. (2023). *Frame-to-Utterance Convergence: A Spectra-Temporal Approach for Unified Spoofing Detection*. https://doi.org/10.48550/ARXIV.2309.09837

Kinnunen, T. H., Lee, K. A., Tak, H., Evans, N., & Nautsch, A. (2024). t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(5), 2622–2637. https://doi.org/10.1109/TPAMI.2023.3313648

Lee, Y., Kim, N., Jeong, J., & Kwak, I.-Y. (2023). Experimental Case Study of Self-Supervised Learning for Voice Spoofing Detection. *IEEE Access*, *11*, 24216–24226. https://doi.org/10.1109/ACCESS.2023.3254880

Lin, H., Ai, Y., & Ling, Z. (2022). A Light CNN with Split Batch Normalization for Spoofed Speech Detection Using Data Augmentation. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1684–1689. 2022 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). https://doi.org/10.23919/APSIPAASC55919.2022.9980260

Liu, T., Kukanov, I., Pan, Z., Wang, Q., Sailor, H. B., & Lee, K. A. (2024). *Towards Quantifying and Reducing Language Mismatch Effects in Cross-Lingual Speech Anti-Spoofing*. https://doi.org/10.48550/ARXIV.2409.08346

Monteiro, J., Alam, J., & Falk, T. H. (2020). Generalized end-to-end detection of spoofing attacks to automatic speaker

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5182

recognizers. *Computer Speech & Language*, *63*, 101096. https://doi.org/10.1016/j.csl.2020.101096

Müller, N. M., Kawa, P., Choong, W. H., Casanova, E., Gölge, E., Müller, T., Syga, P., Sperl, P., & Böttinger, K. (2024). *MLAAD: The Multi-Language Audio Anti-Spoofing Dataset*. https://doi.org/10.48550/ARXIV.2401.09512

Prihasto, B., & Azhar, N. F. (2021). *Evaluation of Recurrent Neural Network Based on Indonesian Speech Synthesis for Small Datasets*. 17–25. https://doi.org/10.4028/www.scientific.net/AST.104.17

Resemble AI - Custom AI Generated Voices. (2024). https://app.resemble.ai/

Tak, H., Kamble, M., Patino, J., Todisco, M., & Evans, N. (2022). Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6382–6386. https://doi.org/10.1109/ICASSP43922.2022.9746213

Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., & Evans, N. (2022). *Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation* (No. arXiv:2202.12233). arXiv. https://doi.org/10.48550/arXiv.2202.12233

Tran, N., Prihasto, B., Le, P. T., Tran, T., Lu, C.-S., & Wang, J.-C. (2024). *EVA-ASCA: Enhancing Voice Anti-Spoofing through Attention-based Similarity Weights and Contrastive Negative Attractors*. 537–540.

Wu, H., Kuo, H.-C., Zheng, N., Hung, K.-H., Lee, H.-Y., Tsao, Y., Wang, H.-M., & Meng, H. (2022). Partially Fake Audio Detection by Self-Attention-Based Fake Span Discovery. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9236–9240. https://doi.org/10.1109/ICASSP43922.2022.9746162

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, *66*, 130–153. https://doi.org/10.1016/j.specom.2014.10.005

Wu, Z., Yamagishi, J., Kinnunen, T., Hanilçi, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M., & Delgado, H. (2017). ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge. *IEEE Journal of Selected Topics in Signal Processing*, *11*(4), 588–604. IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2017.2671435