Preprints.org

Article

Not peer-reviewed version

Fake Voice Detection: A Comparative Analysis of Complex-Valued Deep Learning and Transformer Models across Multiple Languages

Mario Jojoa *, Alfonso Bahillo, Dávid Sztahó, Giovanni Hernandez, Géza Nemeth

Posted Date: 3 February 2026

doi: 10.20944/preprints202601.2360.v1

Keywords: fake voice detection; audio deepfakes; complex-valued deep learning; Wav2Vec 2.0; speech security

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Fake Voice Detection: A Comparative Analysis of Complex-Valued Deep Learning and Transformer Models across Multiple Languages

Mario Jojoa-Acosta [1,*], Alfonso Bahillo [1], Dávid Sztahó [2], Giovanni Hernandez [3] and Géza Nemeth [2]

[1] Department of Signal Theory and Communications, Universidad de Valladolid, Campus Miguel Delibes, Paseo de Belén, 1, 47011 Valladolid, Spain
[2] Budapest University of Technology and Economics (BME), Műegyetem rkp. 3, 1111 Budapest, Hungary
[3] Universidad Mariana de Pasto, Calle 18 No. 50-48, Pasto, Colombia
* Correspondence: mariofernando.jojoa@uva.es; Tel.: +(34)-602454625

Abstract

The rapid progress of modern text-to-speech (TTS) systems has led to synthetic voices that are increasingly indistinguishable from real human speech, raising serious concerns for security, audio forensics, and biometric authentication. As a result, automatic fake voice detection has become a relevant and challenging research problem. This work addresses the problem of distinguishing synthetically generated voices from real human speech using artificial intelligence techniques. Two state-of-the-art approaches are evaluated. The first approach is based on complex-valued deep learning and is motivated by the hypothesis that discriminative information between real and synthetic speech is partially embedded in the phase structure of the signal. By representing audio features in the complex domain, this model explicitly captures both magnitude and phase components, enabling the detection of subtle artifacts introduced during synthetic speech generation. The second approach relies on the pretrained Wav2Vec 2.0 transformer model, which learns robust speech representations through large-scale self-supervised training. Training and evaluation are conducted using a multilingual dataset collected from different countries and linguistic contexts. The dataset includes English speech from Ugandan speakers, Spanish speech from Colombian speakers, and Hungarian speech from native Hungarian speakers. Experimental results show that the Wav2Vec 2.0 model achieves F1-scores of 0.90 for English and 0.98 for Spanish, while the Complex-valued Convolutional Neural Network obtains an F1-score of 0.83 for Hungarian. These findings highlight the potential of both complex-valued models and foundation speech models to improve the security of synthetic voice generation systems in multilingual and cross-domain scenarios.

Keywords: fake voice detection; audio deepfakes; complex-valued deep learning; Wav2Vec 2.0; speech security

1. Introduction

The ability to generate human-like speech using artificial systems has undergone a remarkable transformation in recent years. Advances in deep learning, particularly neural vocoders, sequence-to-sequence architectures, and large-scale self-supervised learning, have enabled text-to-speech (TTS) systems to produce audio signals that are increasingly indistinguishable from real human voices (Shen et al. 2018; van den Oord et al. 2016; Kim et al. 2020; Ren et al. 2021). Modern neural vocoders such as WaveNet, WaveGlow, HiFi-GAN, and diffusion-based models have significantly reduced perceptual artifacts while improving naturalness and prosodic control (Kong et al. 2020; Kim et al. 2021; Chen et al. 2021). As a result, synthetic speech is now widely deployed in applications such as assistive technologies, conversational agents, and real-time translation systems (Tan et al. 2021).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

At the same time, these advances have introduced new vectors for malicious exploitation. Synthetic speech can be used to impersonate individuals, bypass voice-based biometric authentication systems, or fabricate convincing audio evidence for social engineering and disinformation campaigns (Yu et al. 2018; Wu et al. 2021). Unlike earlier generations of synthetic speech, which were often easily identifiable due to robotic prosody or spectral artifacts, modern neural TTS systems can generate speech that is perceptually convincing even to trained listeners (Cooper et al. 2020). This shift has made manual detection impractical and has placed automatic fake voice detection at the center of current research in audio security and digital forensics (Salih et al. 2025; Cao et al. 2022).

Detecting fake voices is therefore a problem of both scientific and societal relevance. In recent years, synthetic voice generation techniques have increasingly been used for non-ethical or directly fraudulent purposes, including call-center scams, social engineering attacks, and identity spoofing (Khanjani et al. 2023). To mitigate these risks, there is a growing need for automatic detection systems capable of distinguishing real human speech from synthetically generated audio under realistic conditions (Dinkel et al. 2017). Such systems must exhibit not only high classification accuracy but also robustness across languages, accents, speakers, recording conditions, and synthesis technologies. Models that perform well only under controlled or homogeneous conditions are of limited value in real-world deployments (Wu et al. 2015).

In this work, we address the binary classification problem of distinguishing real human speech from synthetically generated speech. Given an audio signal $x(t)$, the goal is to determine whether it originates from a natural speech production process or from an artificial generation pipeline. This task is inherently challenging for several reasons. First, speech synthesis systems are evolving rapidly, continuously reducing the perceptual gap between real and artificial signals (Ren et al. 2021; Donahue et al. 2021). Second, speech signals are highly variable and depend on factors such as language, accent, speaking style, emotional state, and recording environment (Schuller 2018). Third, machine learning models are prone to learning spurious correlations, such as language- or channel-specific cues, instead of intrinsic properties of synthetic speech (Chen et al. 2020). A major concern is domain bias: when real and synthetic samples come from different linguistic or demographic distributions, classifiers may exploit superficial cues rather than genuine synthesis artifacts (Nguyen and Do 2025).

Early approaches to synthetic speech detection relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch statistics, jitter, shimmer, and spectral flatness measures, combined with classical classifiers (Evans et al. 2014). Although these methods provided initial insights, they have proven insufficient against modern neural TTS systems, which generate much more natural and artifact-free speech (Dinkel et al. 2017). More recent approaches leverage deep learning models that learn discriminative representations directly from raw or minimally processed audio, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid architectures (Salih et al. 2025; Cao et al. 2022). In parallel, self-supervised learning has emerged as a powerful paradigm for speech representation learning, with models such as Wav2Vec 2.0 and HuBERT achieving state-of-the-art results across multiple speech tasks (Baevski et al. 2020; Hsu et al. 2021).

Despite their success, most existing detection approaches operate in the real-valued domain and focus primarily on magnitude-based representations of the signal. Phase information, which plays a fundamental role in signal reconstruction and fine temporal structure, is often discarded or only indirectly modeled (Shi et al. 2006). From a signal processing perspective, human speech production is a continuous physical process, whereas synthetic speech is generated through discrete computational pipelines that may introduce subtle phase-related artifacts or inconsistencies (Hemavathi and Kumaraswamy 2021). These artifacts may be difficult to perceive in the waveform domain but can become more evident in complex time–frequency representations (Trabelsi et al. 2018).

Motivated by these considerations, the first approach explored in this work is based on complex-valued deep learning. Complex-valued neural networks extend conventional real-valued architectures by allowing weights, activations, and intermediate representations to operate in the complex domain,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

enabling the explicit modeling of both magnitude and phase information (Trabelsi et al. 2018; Cole et al. 2021). Recent studies have demonstrated the potential of complex-valued models in audio and signal processing tasks, including speech enhancement and classification (Agrawal 2025). By preserving phase information throughout the processing pipeline, complex-valued models may better capture fine-grained temporal and spectral structures that differentiate real from synthetic speech, particularly in the presence of neural vocoders that prioritize perceptual quality over strict physical consistency.

We also evaluate approaches based on pretrained self-supervised speech models, specifically Wav2Vec 2.0 transformer. These models are trained on large-scale unlabeled speech corpora and learn hierarchical representations encoding phonetic, prosodic, and higher-level information (Baevski et al. 2020; Hsu et al. 2021). Wav2Vec 2.0 model have demonstrated strong performance in tasks such as speaker recognition, speech recognition, and emotion analysis (Pepino et al. 2021), and their robustness across languages and recording conditions makes them attractive candidates for synthetic speech detection (Jiang et al. 2020). However, their reliance on high-level abstractions raises the question of whether low-level synthesis artifacts, particularly those related to phase structure, are captured as effectively as with explicitly phase-aware models.

To assess the robustness of the proposed approaches, we construct a multilingual dataset encompassing English, Spanish, and Hungarian speech from different geographical regions. Real speech data are collected from speakers in Uganda (English), Colombia (Spanish), and Hungary (Hungarian), while synthetic speech is generated using state-of-the-art TTS systems appropriate for each language, including commercial and open-source solutions. The use of controlled textual content for both real and synthetic speech reduces lexical bias, while the diversity of speakers and recording conditions introduces realistic variability. A Mean Opinion Score (MOS) evaluation confirms that the synthetic signals are perceived as highly natural, reinforcing the relevance and difficulty of the automatic detection task (Cooper et al. 2024).

2. Materials and Methods

The dataset used in this work was specifically designed to evaluate fake voice detection in a multilingual and cross-domain setting. This section describes the data collection platform, the text material and synthesis process, the dataset composition, and the preprocessing pipeline applied before model training.

The dataset was built using a web-based application developed for audio recording [1]. The main goal of this design was to enable data collection with generic laptops and built-in microphones, allowing speakers to participate remotely from different countries. This strategy increases variability in recording conditions and makes the dataset more representative of realistic usage scenarios. Each participant accessed the web application, where they could start and stop the recording process through a simple graphical interface. Participants were instructed to read a short fairy tale displayed on the screen. After completing the recording, they filled out a form with basic metadata and explicitly provided consent for research use of their recordings by selecting a consent checkbox.

To maintain a generic and reproducible setting, synthetic fairy tales were generated and used as reading material. The same texts were later synthesized using TTS systems, ensuring that real and synthetic speech shared the same linguistic content. This design reduces potential bias caused by lexical content, dialectal expressions, or topic differences, and facilitates a fair comparison between real and synthetic samples. For English and Spanish, the same set of fairy tales was used, presented in the corresponding language. Speakers were recruited in university environments in native-language countries, (Uganda for English and Colombia for Spanish). For the Hungarian dataset, a similar approach was followed, but the base texts were obtained from a German fairy-tale repository and then translated into Hungarian before being presented to the speakers and synthesized.

[1] https://github.com/mario42004

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

Synthetic speech was generated using different systems per language in order to better reflect realistic, heterogeneous conditions:

- English: Microsoft Azure Neural TTS, featuring both Senegalese and Kenyan voices (Not Uganda was available).
- Spanish: Piper TTS with approximately three different Colombian voices.
- Hungarian: an ad-hoc neural TTS pipeline adapted to the language.

The amount of real speech collected for Uganda English was 78 recordings of approximately 120 s duration each; for Colombian Spanish, 95 recordings of approximately 120 s; and for Hungarian, around 38 recordings of approximately 5 s. For each language, a comparable number of synthetic samples was generated using the TTS systems mentioned above. Each speaker gave explicit consent for the use of their recordings for research purposes.

To increase robustness and mitigate overfitting, additive noise was applied to the recordings. Specifically, Gaussian noise with a random signal-to-noise ratio (SNR) uniformly sampled between 0 and 20 dB was added to generate augmented versions of the original signals.

After augmentation, the data were split into training, validation, and test sets with proportions of 60%, 20%, and 20%, respectively. The split was performed at the level of raw recordings before augmentation and segmentation. This strategy ensures that segments derived from the same original recording do not appear in both training and evaluation sets, thereby avoiding data leakage and overoptimistic performance estimates.

Finally, each recording was segmented into non-overlapping 1-second clips, and silence-only segments were discarded. This segmentation strategy reduces issues related to long silence intervals and standardizes the input length for model training. Table 1 summarizes the number of raw recordings, augmented recordings, and 1-second segments per language and class.

Table 1. Dataset statistics per language and class.

Language / Class	Raw recordings	Augmented recordings	1s segments
Uganda English Real	95	95	6011
Colombia Spanish Real	78	78	6262
Hungarian Real	38	38	228
Uganda English Fake	101	101	13141
Colombia Spanish Fake	121	121	14586
Hungarian Fake	40	40	240

3. Methods

This section describes the methods used to train, validate, and test the algorithms for the proposed task. Two approaches are considered: a complex-valued deep learning model operating on time–frequency representations, and a transformer-based model leveraging pretrained Wav2Vec 2.0 representations.

3.1. Complex-Valued Deep Learning Model

The proposed complex-valued deep learning model operates directly on complex time–frequency representations of speech. Each 1-second audio segment is first transformed into the time–frequency domain. Let $x(t) \in \mathbb{R}$ denote a speech segment. We compute its complex-valued representation using the Continuous Wavelet Transform (CWT):

$$W_x(a,b) = \int_{-\infty}^{\infty} x(t)\, \psi^*_{a,b}(t)\, dt, \tag{1}$$

where $\psi_{a,b}(t)$ is a scaled and translated version of a complex mother wavelet, a is the scale parameter, b is the temporal shift, and $(\cdot)^*$ denotes complex conjugation. In this work, a complex Morlet wavelet is used due to its favorable time–frequency localization properties for speech.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

The resulting complex scalogram is denoted as

$$S \in \mathbb{C}^{T \times F}, \tag{2}$$

where T and F are the number of time steps and frequency bins, respectively. In the experiments, T and F are chosen such that the input is reshaped to a fixed size of $T \times F = 128 \times 128$, which provides a good compromise between temporal and frequency resolution and computational cost. This representation preserves both magnitude and phase information, allowing the model to capture fine-grained temporal and spectral patterns, including potential phase inconsistencies introduced by synthetic speech generation pipelines.

Let $Z = Z_r + iZ_i$ denote a complex feature map and $K = K_r + iK_i$ a complex convolutional kernel. The complex convolution is defined as:

$$Z * K = (Z_r * K_r - Z_i * K_i) + i(Z_r * K_i + Z_i * K_r), \tag{3}$$

where $*$ denotes the standard real-valued convolution. This operation preserves the algebraic structure of complex numbers and allows the network to jointly model magnitude and phase interactions across time–frequency regions.

Non-linearity is introduced using the modReLU activation function:

$$\text{modReLU}(z) = \text{ReLU}(|z| + b)\, \frac{z}{|z|}, \tag{4}$$

where b is a learnable bias parameter and $|z|$ denotes the complex modulus. This activation preserves the phase of z and applies a rectified shift to its magnitude, which has been shown to be effective in complex-valued networks. To stabilize training, complex batch normalization is applied independently to the real and imaginary parts of each feature map. Complex max-pooling is implemented by applying the pooling operator to the magnitude and propagating the corresponding complex values.

The proposed complex-valued convolutional neural network (CV-CNN) follows a relatively compact but expressive architecture with four convolutional blocks followed by two fully connected layers. The design targets a balance between model capacity and the amount of available training data, resulting in a total number of parameters on the order of million, which is suitable for the size of the dataset considered.

The architecture can be summarized as follows (dimensions are given for an input scalogram of size 128×128):

Figure 1 illustrates the overall architecture of the proposed complex-valued convolutional neural network. The input to the model is a complex time–frequency representation (scalogram) extracted from a 1-second speech segment. The network comprises four hierarchical complex convolutional blocks that progressively increase the number of feature channels while reducing spatial resolution, enabling the extraction of increasingly abstract phase-aware representations. After global complex pooling and a complex fully connected projection, the learned complex embedding is mapped to the real domain via magnitude projection. A lightweight real-valued classifier then produces a binary decision indicating whether the input speech signal is real or synthetically generated.

The total number of trainable parameters is on the order of million, which is compatible with the size of the training set after segmentation and augmentation. The network is optimized using the binary cross-entropy loss and the Adam optimizer, with early stopping based on the validation F1-score, as described in Section 4. This architecture is fully implementable using existing complex-valued deep learning extensions for PyTorch and can be trained end-to-end for the task of multilingual fake voice detection.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

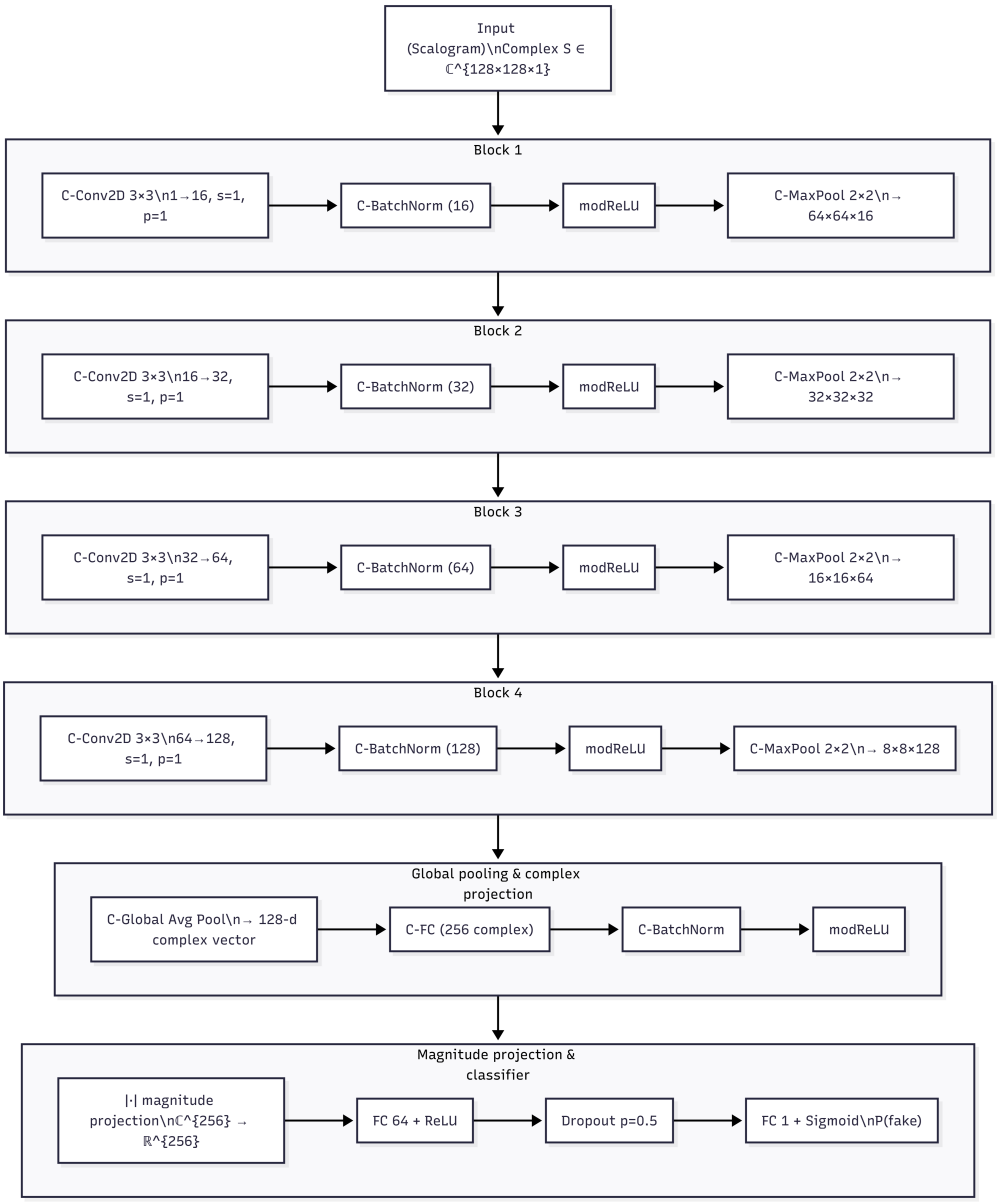

Figure 1. Architecture of the proposed complex-valued convolutional neural network (CV-CNN) for fake voice detection.

Table 2. Proposed complex-valued CNN architecture for fake voice detection (input size 128×128).

Layer	Type	Output shape
Input	Complex scalogram	$128 \times 128 \times 1$
Block 1	C-Conv2D + BN + modReLU + C-MaxPool	$64 \times 64 \times 16$
Block 2	C-Conv2D + BN + modReLU + C-MaxPool	$32 \times 32 \times 32$
Block 3	C-Conv2D + BN + modReLU + C-MaxPool	$16 \times 16 \times 64$
Block 4	C-Conv2D + BN + modReLU + C-MaxPool	$8 \times 8 \times 128$
Global pool	Complex global average pooling	128 (complex)
Dense (cplx)	Complex FC + BN + modReLU	256 (complex)
Projection	Magnitude	256 (real)
Dense (real)	FC + ReLU + Dropout(0.5)	64 (real)
Output	FC + Sigmoid	1 (real)

3.2. Transformer-Based Wav2Vec 2.0 Model

To compare the proposed complex-valued approach with a real-valued baseline, a pretrained transformer-based model Wav2Vec 2.0 was employed. The model operates directly on raw audio

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

waveforms, using 1-second segments extracted from the original recordings. No explicit normalization or handcrafted feature extraction is applied prior to the model input.

A pretrained Wav2Vec 2.0 encoder is used to extract contextualized latent representations from each input segment. Internally, the model maps the raw waveform into a sequence of frame-level hidden representations produced by the transformer encoder. Let h_t denote the hidden representation at time step t, where $t = 1, \ldots, T$ and T depends on the internal framing of the Wav2Vec 2.0 encoder.

To obtain a fixed-dimensional representation suitable for classification, a global temporal mean pooling is applied over the sequence of hidden states:

$$h = \frac{1}{T} \sum_{t=1}^{T} h_t, \tag{5}$$

where h represents a single embedding summarizing the information contained in the 1-second audio segment. This operation aggregates information across time while preserving the statistics learned by the pretrained encoder.

The resulting pooled representation is passed to a task-specific classification head consisting of a fully connected layer followed by a sigmoid activation function. The output corresponds to the estimated probability that the input audio segment is synthetically generated.

During fine-tuning, only the upper transformer layers and the classification head are updated, while the lower layers of the pretrained Wav2Vec 2.0 encoder remain frozen. This training strategy limits the number of trainable parameters and reduces the risk of overfitting, while still allowing the model to adapt the high-level representations to the fake voice detection task.

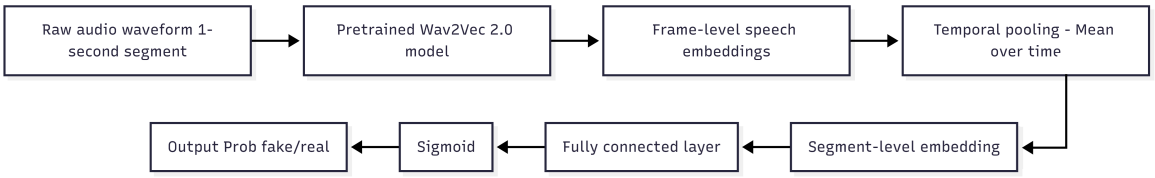

Figure 2. Schematic representation of the Transformer-based Wav2Vec 2.0 model (W2V) used for fake voice detection.

3.3. Training and Validation

Both models are trained using the same train/validation/test partitions described in Section 2. The binary cross-entropy loss is minimized using the Adam optimizer. Early stopping is employed to prevent overfitting, with the validation F1-score used as the monitoring metric and a patience of 10 epochs. The model parameters corresponding to the best validation F1-score are retained for final evaluation on the test set.

Hyperparameters such as learning rate, batch size, and maximum number of epochs are selected based on validation performance using Grid Search method. Table 3 summarizes the configuration used for both models.

Table 3. Hyperparameters used for the complex-valued CNN and Wav2Vec 2.0 model.

Parameter	CV-CNN	Wav2Vec 2.0
Optimizer	Adam	Adam
Learning rate	1×10^{-4}	1×10^{-5}
Batch size	32	16
Maximum epochs	100	50
Early stopping patience	10	10
Loss function	Binary cross-entropy	Binary cross-entropy

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

4. Experimental Setup and Evaluation Metrics

Experiments are conducted separately for each language (English, Spanish, and Hungarian) in order to analyze language-dependent performance. For each language, identical train, validation, and test splits are used across all models to ensure a fair and consistent comparison.

All audio recordings for the English and Spanish datasets are acquired at a sampling frequency of 24 kHz with a resolution of 16 bits per sample. These acquisition settings are preserved throughout the entire processing pipeline, including segmentation, data augmentation, and model training. Audio recordings for the Hungarian dataset are acquired using language-specific recording conditions consistent with the corresponding data collection protocol.

Both the complex-valued convolutional neural network and the transformer-based Wav2Vec 2.0 model are implemented in Python using standard deep learning frameworks. All experiments are conducted on a workstation equipped with an NVIDIA RTX 3090 Ti GPU featuring 24 GB of GDDR6X memory. The GPU provides a peak theoretical performance of approximately 40 TFLOPS in single-precision (FP32), which is sufficient to support the training and fine-tuning of the proposed models.

For each language, models are trained independently using the corresponding dataset splits. Performance is evaluated using Accuracy, F1-score macro, Sensitivity (recall for the synthetic class), Specificity (recall for the real class), and the Area Under the Receiver Operating Characteristic Curve (AUC). The F1-score is considered the primary evaluation metric, as it provides a balanced measure of precision and recall and is particularly informative in the presence of potential class imbalance. The AUC metric is additionally reported to assess the models' discriminative capability across different decision thresholds. For completeness, all metrics are reported for each language and model configuration.

5. Results

The evaluation of synthetic speech was conducted using the Mean Opinion Score (MOS) method, a common subjective metric for assessing the quality of speech synthesis. The MOS scores were assigned by a group of speakers, who evaluated the synthetic speech system in three different languages: Hungarian, Spanish, and English. The evaluation was based on the clarity, naturalness, and overall quality of the generated speech.

As shown in Table 4, the synthetic speech in Hungarian received the highest MOS score of 5, indicating a high level of satisfaction from the evaluators. The Spanish synthetic speech, however, scored lower with an average of 3.7, reflecting some dissatisfaction with the quality. The English synthetic speech scored 4.5, demonstrating a relatively good quality, but still below the optimal score of 5.

Table 4. Mean Opinion Score (MOS) Evaluation of Synthetic Speech for Different Languages.

Language	MOS Score
Hungarian	5.0
Spanish	3.7
English	4.5

The Table 5 summarizes the performance of the complex-valued deep learning model (CV-CNN) and the Wav2Vec 2.0 transformer model on the test sets for each language. The reported metrics are Accuracy, F1-score, Sensitivity, Specificity, and AUC.

For the Wav2Vec 2.0 model, in English, the model achieved an accuracy of 0.94, with a sensitivity of 0.96 and specificity of 0.93. Although the accuracy is quite high, the F1-score (0.90) could be improved, suggesting a slight mismatch between true positive and false negative rates. In Spanish, the model performed outstandingly with an accuracy of 0.97, an F1-score of 0.98, and specificity of 0.98. This performance is significantly higher than that of the English model, indicating better adaptation to

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

Spanish data. In Hungarian, the performance was lower, with an accuracy of 0.79 and sensitivity of 0.81, indicating that the model has more difficulty detecting fake voices in Hungarian compared to English or Spanish.

Table 5. Performance of the evaluated models on the test sets for each language.

Model	Lang	TN	FP	FN	TP	Acc	F1	Sens	Spec	AUC
W2V	EN	2592	195	37	1008	0.94	0.90	0.96	0.93	0.98
W2V	ES	1164	22	89	2896	0.97	0.98	0.97	0.98	0.99
W2V	HU	39	12	8	35	0.79	0.78	0.81	0.76	0.88
CV-CNN	EN	2562	186	67	1017	0.93	0.89	0.94	0.93	0.97
CV-CNN	ES	1181	126	72	2792	0.95	0.97	0.97	0.90	0.98
CV-CNN	HU	42	10	5	37	0.84	0.83	0.88	0.81	0.88

For the Complex-valued CNN model, in Spanish, the model achieved an accuracy of 0.95, with a sensitivity of 0.97 and specificity of 0.90, showing strong capability in detecting both real and fake voices. In English, the model obtained similar performance to the Wav2Vec 2.0 model with an accuracy of 0.93, sensitivity of 0.94, and specificity of 0.93. Although sensitivity is slightly higher than in the Wav2Vec 2.0 model, the F1-score (0.89) is somewhat lower. For Hungarian, the Complex-valued CNN model had a performance of 0.84 accuracy, with a sensitivity of 0.88 and specificity of 0.81. While this model performs better than the Wav2Vec 2.0 model in this language, the precision remains relatively low compared to the other languages.

In the comparative summary, the Wav2Vec 2.0 model outperformed the Complex-valued CNN model in Spanish, showing significantly better results in terms of accuracy, F1-score, and specificity. In English, both models performed similarly, although the Wav2Vec 2.0 model achieved a slight advantage in accuracy. For Hungarian, both models showed reduced performance, though the Complex-valued CNN model appeared to be somewhat more robust than the Wav2Vec 2.0 model.

Finally, the Wav2Vec 2.0 model is clearly more effective in Spanish, while the Complex-valued CNN model shows better handling of fake voices in Hungarian. The Wav2Vec 2.0 model's ability to handle language variations is an important factor, as it performs consistently between English and Spanish. Both models show areas for improvement in terms of F1-score and sensitivity in certain languages, suggesting that further customization or training with more language-specific data is needed.

6. Conclusions and Future Work

The results obtained in this work show that while phase-aware complex-valued modeling is theoretically motivated for synthetic speech detection, its practical benefits remain limited under certain multilingual conditions. Contrary to our expectations, the complex-valued model did not outperform the Wav2Vec 2.0-based approach in Spanish and English, suggesting that pretrained transformer representations may already encode discriminative cues related to synthetic artifacts, even without explicitly modeling phase information.

The superior performance of Wav2Vec 2.0 in these two languages indicates that large-scale self-supervised representations are highly effective for detecting manipulated audio, especially when the target languages are well represented in the pretraining data. Nonetheless, the complex-valued model demonstrated a relative advantage in Hungarian, a more phonetically distinct and low-resource scenario. This suggests that explicit phase modeling may provide additional robustness when pretrained features are less linguistically aligned with the evaluation domain.

Overall, the findings highlight an important limitation, complex-valued architectures alone may not guarantee improved generalization across diverse languages and recording conditions. Instead,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

our results point toward the complementary nature of both approaches, where phase modeling and pretrained representations capture different forms of discriminative evidence. Ensuring robust multilingual fake voice detection will likely require leveraging the strengths of both.

Future work will explore hybrid and fusion-based architectures that jointly exploit explicit phase cues and high-level self-supervised speech representations. Additionally, cross-lingual adaptation and transfer learning techniques will be investigated to better support underrepresented languages. Expanding the evaluation to biometric spoofing scenarios and tampering localization tasks is also envisioned, enabling a more comprehensive deployment of these technologies in security-critical applications.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank all volunteers who contributed speech recordings for this study, as well as the institutions that facilitated data collection. This work was supported in part by the **ENFIELD Project** (Grant No. oc2-2024-TES-02_15) and by the **Spanish Ministry** under the **Aginplace Project** (Grant Nos. PID2023-146254OB-C41 and PID2023-146254OA-C44).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In Proceedings of the Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. https://doi.org/10.1109/ICASSP.2018.8461368.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499* **2016**.

Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.

Kim, J.; Kong, J.; Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, Vol. 139, *Proceedings of Machine Learning Research*, pp. 5530–5540.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R.; Norouzi, M.; Chan, W. WaveGrad: Estimating Gradients for Waveform Generation. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

Tan, X.; Qin, T.; Soong, F.; Liu, T.Y. A Survey on Neural Speech Synthesis. *arXiv preprint arXiv:2106.15561* **2021**.

Yu, H.; Tan, Z.H.; Ma, Z.; Martin, R.; Guo, J. Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *29*, 4633–4644.

Wu, Z.; Yamagishi, J.; et al. ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge. In Proceedings of the Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

Cooper, E.; Lai, C.I.J.; Yasuda, Y.; Fang, F.; Wang, X.; Chen, N.; Yamagishi, J. Zero-Shot Multi-Speaker Text-To-Speech with State-of-the-Art Neural Speaker Embeddings. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

Salih, A.O.M.; Emam, A.H.M.; Suliman, A.; Babiker, N.B.M. Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review. *Engineering, Technology & Applied Science Research* **2025**, *15*, 29824–29832.

Cao, R.; Abdulatif, S.; Yang, B. CMGAN: Conformer-Based Metric GAN for Speech Enhancement. In Proceedings of the Proc. Interspeech, 2022.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2026 doi:10.20944/preprints202601.2360.v1

Khanjani, Z.; Watson, G.; Janeja, V.P. Audio deepfakes: A survey. *Frontiers in Big Data* **2023**, *5*, 1001063. https://doi.org/10.3389/fdata.2022.1001063.

Dinkel, H.; Chen, N.; Qian, Y.; Yu, K. End-to-End Spoofing Detection with Raw Waveform CLDNNS. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

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

Donahue, C.; Dieleman, S.; Zen, H. End-to-End Adversarial Text-to-Speech. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM* **2018**, *61*, 90–99.

Chen, T.; Kumar, A.; Nagarsheth, P.; Sivaraman, G.; Khoury, E. Generalization of Audio Deepfake Detection. In Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Tokyo, Japan, 2020; pp. 132–137. https://doi.org/10.21437/Odyssey.2020-19.

Nguyen, T.T.H.; Do, T.N.D. Cross-lingual XLSR-Wav2Vec2-based Speech Spoofing Detection for Vietnamese Speech. In Proceedings of the 2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2025, pp. 1–6.

Evans, N.; Kinnunen, T.; Yamagishi, J. Speaker Recognition Anti-Spoofing. In *Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks*; Marcel, S.; Nixon, M.S., Eds.; 2014; pp. 125–146.

Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.

Hsu, W.N.; et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**.

Shi, G.; Shanechi, M.M.; Aarabi, P. On the importance of phase in human speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **2006**, *14*, 1867–1874. https://doi.org/10.1109/TSA.2006.879341.

Hemavathi, R.; Kumaraswamy, R. Voice conversion spoofing detection by exploring artifacts estimates. *Multimedia Tools and Applications* **2021**, *80*, 23561–23580.

Trabelsi, C.; et al. Deep Complex Networks. In Proceedings of the International Conference on Learning Representations (ICLR), 2018.

Cole, E.K.; Cheng, J.Y.; Pauly, J.M.; Vasanawala, S.S. Analysis of deep complex-valued convolutional neural networks for MRI reconstruction and phase-focused applications. *Magnetic Resonance in Medicine* **2021**, *86*, 1093–1109.

Agrawal, N. Phase-Aware Deep Learning with Complex-Valued CNNs for Audio Signal Applications. *arXiv* **2025**, *abs/2510.09926*, [arXiv:cs.LG/2510.09926]. https://doi.org/10.48550/arXiv.2510.09926.

Pepino, L.; Riera, P.; Ferrer, L. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In Proceedings of the Proc. Interspeech, 2021.

Jiang, Z.; Zhu, H.; Li, P.; Ding, W.; Ren, Y. Self-Supervised Spoofing Audio Detection Scheme. In Proceedings of the Interspeech 2020, 2020.

Cooper, E.; Yamagishi, J.; Henter, G.E. A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology* **2024**, *45*, 161–183. https://doi.org/10.1250/ast.45.12.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.