# Future-Proofing Multilingual Fake Speech Detection

Meriç Demirörs, Ahmet Murat Özbayoğlu, and Toygar Akgün

TOBB University of Economics and Technology, Ankara, Turkey

**Abstract.** Developments in the field of generative-AI have made it extremely difficult to distinguish artificially generated content from real content. As a result, their reliable detection has become more important. The topic of this research is detecting speeches that are generated by future generative-AI models in unknown languages. It focuses on answering ”With what information does a model distinguish fake from real audio, does it learn how spoken languages sound, or does it learn a specific trait of generated speech waves?” Multiple models are trained on various datasets to detect synthetic audio signals generated by generative-AI models. After multiple sessions of trains and tests, the best test accuracy scores for different test sessions are 94.92% for known language from unknown generative-AI model, 98. 44% for an unknown language from known generative-AI model, and 95. 18% for an unknown language from unknown generative-AI model.

**Keywords:** CNN, Bispectrum

## 1 Introduction

Recent advances have significantly transformed content creation. By 2025, text-to-image models based on Transformer [1] architectures (e.g., DALL-E [2]) and denoising diffusion probabilistic models (e.g., Stable Diffusion [3]) or models that use both (e.g., SORA [4], VEO2 [5]) are capable of producing highly realistic visuals. Building on these successes, audio generation has also been improved with models like Meta's Voicebox [6] and GPT-4o [7]. The realism of synthetic audio created by generative-AI has emphasized the need to differentiate between real and synthetic audio.

Proposed method's applicability is shown in earlier research [8], with this new step, main goal is to prove the effectiveness of the method by testing models on languages and state-of-the-art generative-AI models that are not in the training dataset. In addition to inspection of detection task results, it is also important to emphasizes differences between used audio features, effect of augmentation and different CNN (Convolutional Neural Network) models' performances and their other aspects. Contrary to previous research this work is built on, there are only CNN models, so architectural and training process differences such as flattening the inputs for FC (Fully Connected) models and SVMs (Support Vector Machines) or passing the whole dataset at once for SVM trainings are not comparable. Rather than that, only CNN models' performances, storage sizes, ease of trainings and some other aspects are considered at the comparison.

The remainder of this paper is structured as follows: Section 1 presents a summary of prior research in this area and emphasizes the unique contributions of this study; Section 2 details the characteristics of audio bispectrums, the process of generating the ”signature image” from audio, the applied data augmentations, the models tested, the datasets used along with their features, and lastly the training and testing sessions; Section 3 analyzes the results; Section 4 suggests possible directions for future researches; and finally, Section 5 concludes the paper.

### 1.1 Literature Review

Numerous studies have focused on detecting synthetic audio. Some [8] has demonstrated and discussed the feature effectiveness [8][9][10][11], detection techniques such as statistical methods [12], clustering algorithms [13], machine learning-based models [8][9][14][15][16][17] [18][19], FC and/or CNN architectures [8][17][18][20][21][22][23][24][25][26][27], and Transformer based architectures [28]. Both raw audio features [9][12][14][15] [16][17][18][20][22][23] and feature extractor models [8][13][21] have been utilized to extract relevant features. These researches, highlight the dominance of deep learning models, in tasks involving both short- and long-term signal features like bispectrum or MFCC (Mel Frequency Cepstral Coefficients). But none of these cited researches are focused to solve multilingual speech detection by testing the models on a dataset which it does not have any bias at any way.

## 2 Methods

### 2.1 Bispectrum

Bispectrum is used in this research as it was in the previous ones [8][12][14][15]. but unlike it, to obtain individual features, features from audio bispectrums are extracted, and various averaging and normalization methods were applied without any helper signal processing library. So the process of obtaining bispectrum features differ from the previous one but gives the same features.

The bispectrum of a signal offers insight into higher-order correlations within the Fourier domain. Specifically, it identifies third-order correlations, in equation (1), where Y($\omega$) represents the Fourier transform and Y*($\omega$) denotes complex conjugate. This highlights ”unnatural” higher-order correlations introduced by nonlinear interactions, particularly between frequency triplets such as [$\omega$1, $\omega$1, $\omega$1+$\omega$1], [$\omega$2, $\omega$2, $\omega$2+$\omega$2], [$\omega$1, $\omega$2, $\omega$1+$\omega$2], and [$\omega$1, -$\omega$2, $\omega$1-$\omega$2]. As the bispectrum is inherently a complex-valued function, it can be more practical and intuitive to analyze it in terms of its magnitude and phase, from the equations (2) and (3), respectively.

To simplify interpretation further, the normalized version of the bispectrum—referred to as the bicoherence—is often employed. This normalization typically involves segmenting the signal into K portions. The bicoherence is then determined by averaging the bicoherence values calculated for each individual segment, as shown in equation (4). All equations are shown below at the Fig. 1. For a deeper exploration of the bispectrum and its properties, consult references [12][14][15], as well as the cited sources within.

$$B(\omega_1,\omega_2)=Y(\omega_1)Y(\omega_2)Y^*(\omega_1+\omega_2) \tag{1}$$

$$|B(\omega_1,\omega_2)|=|Y(\omega_1)|\cdot|Y(\omega_2)|\cdot|Y(\omega_1+\omega_2)| \tag{2}$$

$$\angle B(\omega_1,\omega_2)=\angle Y(\omega_1)+\angle Y(\omega_2)-\angle Y(\omega_1+\omega_2) \tag{3}$$

$$\hat{B}_c(\omega_1,\omega_2)=\frac{\frac{1}{K}\sum_k Y_k(\omega_1)Y_k(\omega_2)Y_k^*(\omega_1+\omega_2)}{\sqrt{\frac{1}{K}\sum_k |Y_k(\omega_1)Y_k(\omega_2)|^2\frac{1}{K}\sum_k |Y_k^*(\omega_1+\omega_2)|^2}} \tag{4}$$

**Fig. 1.** Formulas for the bispectrum and bicoherence

### 2.2 Audio Processing

The process begins by dividing the audio into K segments, each consisting of 400 samples. This segment size is chosen based on the typical stationarity window of the human voice, with all audio processed at a sampling rate of 16 kHz in this study. An overlap of 200 samples is used between consecutive segments to ensure continuity. Initially, the 'cum3' feature is computed for each segment. Following this, a "signature image" is generated from the audio, from which the remaining four features —'absolute,' 'angle,' 'real,' and 'imag'—are derived. Lastly, min-max normalization is applied to the features.
Feature Extraction Algorithm:

```
– $\odot$ = Element wise Multiplication
– $\circ$ = Matrix Multiplication
– SS = 400 (segment size)
– OS = 200 (overlap size)
– audio_data = segment with SS sample from audio read at 16000 Hertz
– data = audio_data + random normal distribution noise
– cum3_sum = 0
–     for each segment in data:
–     cum3 = zero matrix with shape SSxSS
–     ind = indices from [0 ,OS)
–     zero_maxlag = zero vector with shape 1
–     signal = data reshaped to 1xSS and centralized
–     signal_t = transpose of signal
–     signal_r = reverse of signal
–     col = concat((sig[ind], zero_maxlag_t))
–     row = concat((rev_signal[0][ind_t], zero_maxlag[0]))
–     toep = toeplitz matrix created from col and row
–     rep_signal_r = repeated reverse signal SS times
–     cum3 = ((toep $\odot$ rev_signal) $\circ$ toep.transpose())
–     cum3 = cum3/SS
–     bispec = FFT_shift(2D_FFT(inverse_FFT_shift(cum3 * hamming_window)))
–     mag = absolute value of bispec
–     phase = angle of the bispec
–     cum3_sum = cum3_sum + cum3
–     R = mag * cos(phase)
–     C = mag * sin(phase)
– cum3 = cum3_sum/(number of segments)
– signature_image = empty matrix to hold complex values
– sum_RC = sum of RC_layers along the stacking axis
– for each index row i and column j at signature_image:
–     L = RC_layers[:,i,j,:] vector of K complex values along the stacking axis
–     top = sum_RC[i,j]
–     bottom = square_root((L.transpose $\circ$ L.conjugate).real)
–     signature_image[i,j] = top/(bottom + eps)
– absolute = absolute value of the signature image
– angle = angle value of the signature image
– cum3 = average 3rd order cumulant of all segments
– real = real part of the signature image
– imag = imaginary part of the signature image
```

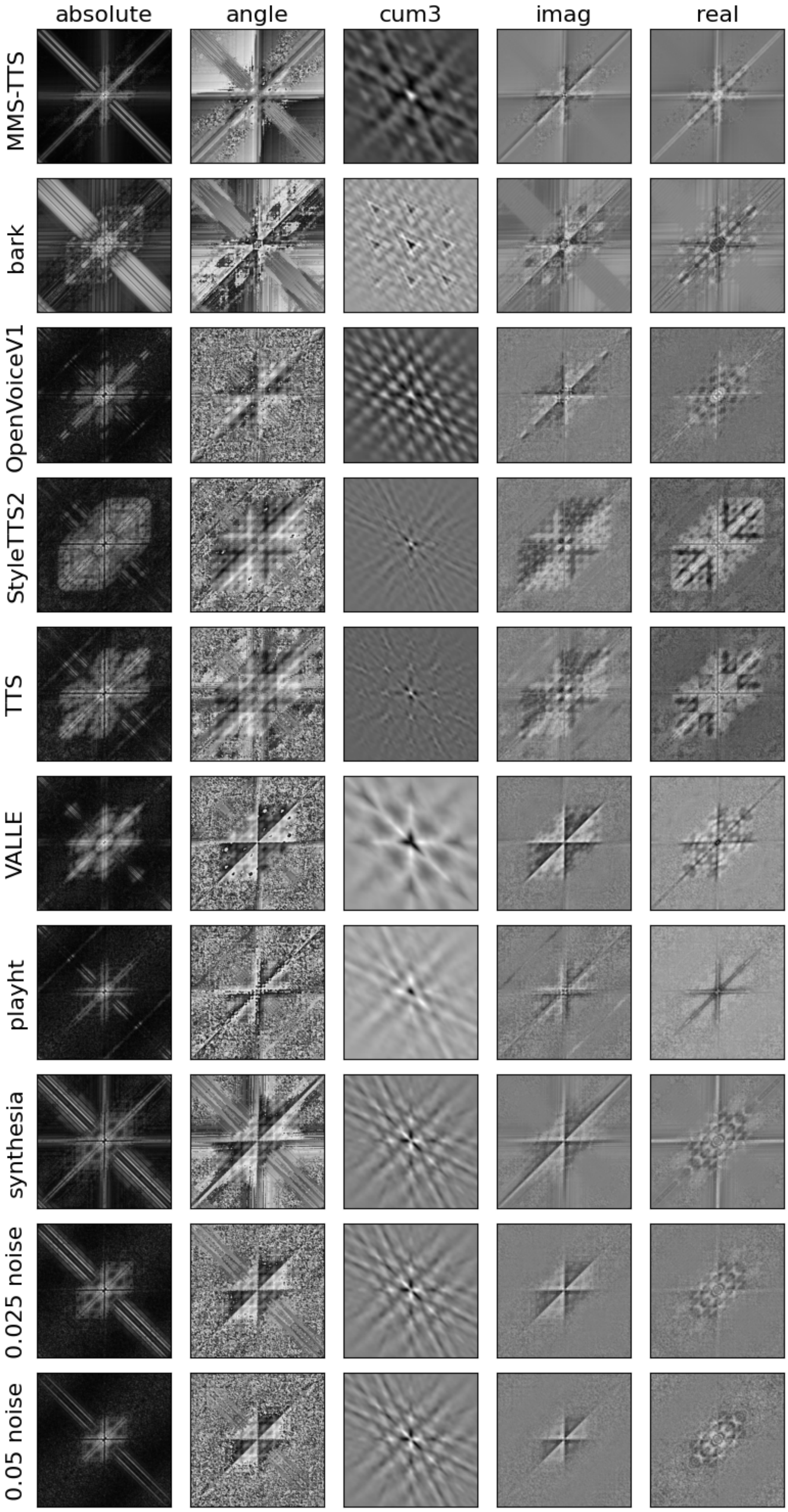

**Fig. 2.** Feature looks for each generative model, (noisy features are for demonstrating the noise's effect. Noise is added to synthesia model for easy comparison between rows)

And Fig. 2 above, is the different feature looks for the same sentence's speech from different methods, and two additional rows are added to show the effect of applied noise augmentation. Rows with noise are noisy versions of audio that is used in the last clean row. To create noise, random distribution with different sigma values are added to the original audio and then audio is clipped between -1 and 1. Used sigma values, which are 0.025 and 0.05, are selected after listening noise added audios with various sigmas and finding the values that make audios noticeably noisy but not inaudible. Audios types are named as 'no noise', '0.025 noise' and '0.05 noise'.

### 2.3 Models

Three different CNN models are trained, ResNet50, GoogLeNet and mid_CNN, which's architecture is illustrated below at the Fig. 3 (ReLU is used after each block, Sigmoid is used for the last layer). It is named mid_CNN, since it is an updated version of basic_CNN architecture from one of the previous studies [8]. Improvements such as additional hidden layers including more complex convolutional layers with dropout layers are added. But still it has potential to improve with other experimented and proved techniques such as skip connections (also called shortcut connections), inception modules or squeeze-and-excitation blocks.

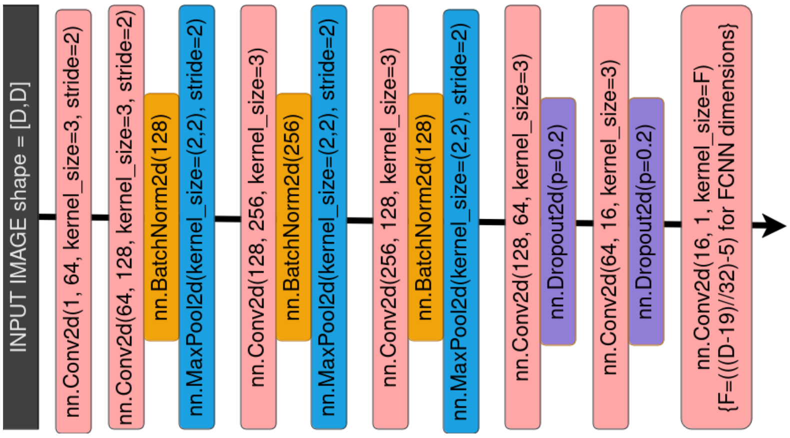
INPUT IMAGE shape = [D,D]
nn.Conv2d(1, 64, kernel_size=3, stride=2)
nn.Conv2d(64, 128, kernel_size=3, stride=2)
nn.BatchNorm2d(128)
nn.MaxPool2d(kernel_size=(2,2), stride=2)
nn.Conv2d(128, 256, kernel_size=3)
nn.BatchNorm2d(256)
nn.MaxPool2d(kernel_size=(2,2), stride=2)
nn.Conv2d(256, 128, kernel_size=3)
nn.BatchNorm2d(128)
nn.MaxPool2d(kernel_size=(2,2), stride=2)
nn.Conv2d(128, 64, kernel_size=3)
nn.Dropout2d(p=0.2)
nn.Conv2d(64, 16, kernel_size=3)
nn.Dropout2d(p=0.2)
nn.Conv2d(16, 1, kernel_size=F)
{F=(((D-19)//32)-5) for FCNN dimensions}

**Fig. 3.** Architecture of Mid_CNN model

Every CNN model has six training types: five trainings on individual features and the one with multi form data, which is created by stacking five individual features into one input (output layers of ResNet50 and GoogLeNet models are modified to perform binary classification).

### 2.4 English Datasets

- 'In-the-Wild' Dataset [29] (2022): contains real and synthetic voice recordings of famous names. There are 19963 real and 11816 synthetic voice recordings, the longest of which is 24 seconds, all with a sample rate of 16 Kilohertz. The dataset's language is English, and it consists of 58 speakers.
- Fake or Real (FoR) [30] original Dataset (2019): contains real and synthetic audios. There are 32496 real and 34405 synthetic audios, the longest of which is 39 seconds, all with sample rate of 22050 Hertz. The dataset's language is English, and it contains multiple speakers.

- Fake or Real (FoR) [30] rerec Dataset (2019): is rerecorded version of the for-2second dataset, which contains two seconds of audios based on original dataset that passed through normalization and various processes, to simulate a scenario where an attacker sends an utterance through a voice channel. Contains real and synthetic audios. There are 6613 real and 6655 synthetic audios, the longest of which is two seconds, all with sample rate of 22050 Hertz. The dataset's language is English, and it contains multiple speakers.
- WaveFake [31] Dataset (2021): contains only synthetic audios. There are 60583 synthetic audios, the longest of which is 16 seconds, all with sample rate of 22050 Hertz. Dataset's contains audios in English and Japan, and it contains multiple AI models' outputs. Only English audios are used.
- VCTK [32] Dataset (2022): contains only real audios. There are 43873 real audios, the longest of which is 16 seconds, all with sample rate of 22050 Hertz. The dataset's language is English, and it contains 110 speakers.

### 2.5 Multilingual Datasets

- CommonLanguage [33] Dataset (2022): contains only real audios. There are 34047 real audios, the longest of which is 105 seconds, all with sample rate of 22050 Hertz. The dataset is multilingual with 45 languages, and it contains multiple speakers. This dataset is composed of speech recordings from of languages that were carefully selected from the CommonVoice [34] database.
- ELTOLSM (80 lines 31 languages 7 methods) [35] Dataset (2024): generated during the research process using open source text-to-speech AI models. 80 lines are generated in English with OpenAI's GPT-4 [36] around five topics: fake audios, famous books, famous movies, famous paintings and famous songs 16 lines for each topic. Then translated to 30 other language: Bengali, Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hindi, Hungarian, Irish, Italian, Japanese, Korean, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish. After experimenting and eliminating unrealistic sounding ones, seven method has selected to generate audios with: bark [37], StyleTTS2 [38], TTS [39], VALL-E-X [40], OpenVoice [41], synthesia [42] and PlayHT [43]. Fake audios are generated in every possible language these models offer (union of these models' languages create a language pool with 31 language) with realistic sounding range of various experimented parameters. The only possible language for PlayHT and synthesia was English, so different speakers are used to create variety instead of different languages or parameters. After listening the generated audios; noisy, unrealistic and faulty ones are eliminated. After elimination of more than 1000 audios there are 4980 left, the longest of which is 22 seconds, all with sample rate of 22050 Hertz.

### 2.6 Hypothesis Datasets

Two multilingual datasets are generated with same processes under this title, which are called HULSIL (100 lines 16 languages) and HULTL (100 lines 12 languages), with a purpose of proving our hypothesis. HULSIL is primarily for testing purposes and HULTL is for training. They are split under these names just to clarify namings in the paper. The whole dataset is called HULTEL (100 lines 28 languages) [44], which is explained below.

- HULTEL (100 lines 28 languages) [44] Dataset (2024): Also another multilingual dataset is generated by using 100 lines. From ten different topics: fashion, tech, economy, education, travel, architecture, sports, video games, history, geography. Ten lines from each are generated by Gemini AI [45]. 16 languages that are not in ELTOLSM is selected to create a dataset to simulate ”fake speech in unknown language by unknown AI model” which are: Albanian, Arabic, Bali, Catalan, Hebrew, Indonesian, Kazakh, Malay, Mongolian, Persian, Somali, Swahili, Thai, Ukrainian, Vietnamese, Welsh. Also, another twelve popular languages, that ELTOLSM contains, are selected to create a dataset. These languages are: Bengali, English, Finnish, French, German, Hungarian, Latvian, Polish, Portuguese, Romanian, Spanish and Swedish. This dataset is used to help with proving the hypothesis from Results and Discussion section. A total of 2800 audios are generated by 28 languages using META's Massively Multilingual Speech text-to-speech model (MMS-TTS) [46]. Longest of which is 15 seconds, all with sample rate of 22050 Hertz. For the purpose of clearity, They will be named as two different datasets: HULSIL (100 lines 16 languages) which contains languages that are not in the ELTOLSM, and HULTL (100 lines 12 languages) which contains languages that are in the ELTOLSM.

To have a balanced dataset distribution in training dataset, at most 5000 samples are selected from every dataset and three sets of datasets are generated from each: no noise dataset, 0.025 noise dataset and 0.05 noise dataset. With all these datasets there are more than 250000 audios, and after feature extractions whole research dataset contains close to 800000 extracted feature images. Where, over 600000 are from English datasets, around 200000 are from multilingual datasets. More details are shown at the Fig. 4 below.

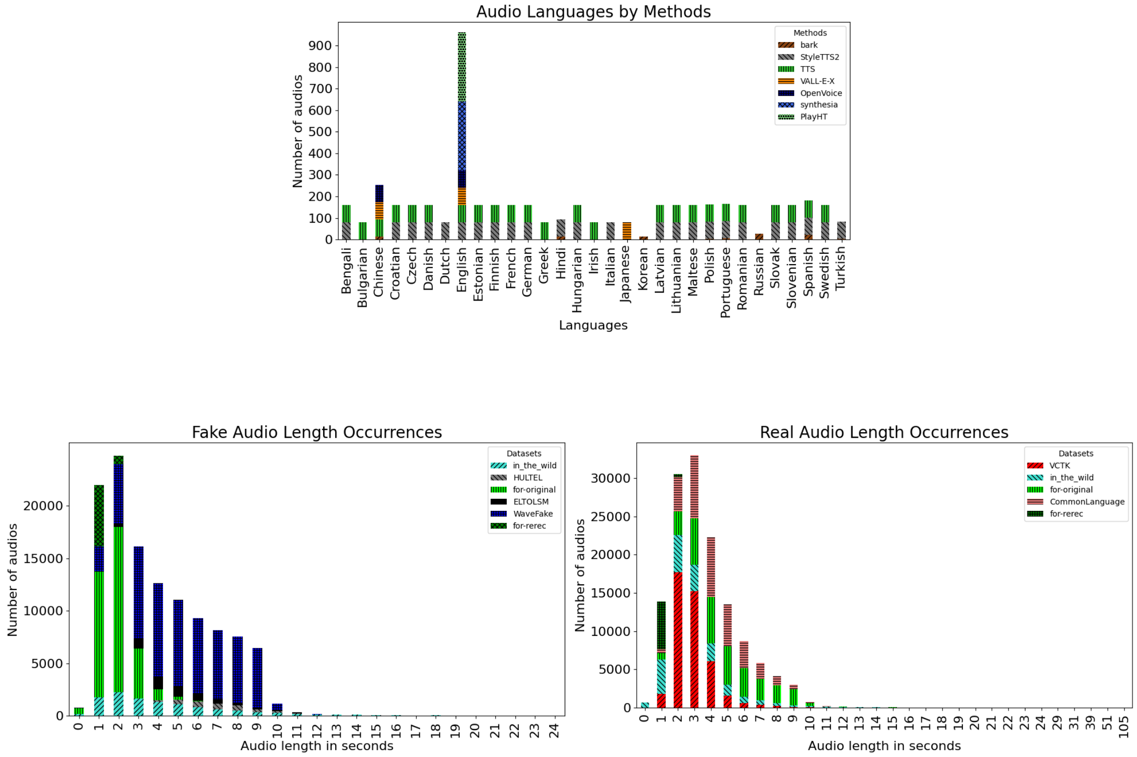

**Fig. 4.** Language distribution graph(upper graph) with Fake (lower left )and Real(lower right) audios' length distribution graphs

### 2.7 Training and Test Sessions

To explain how this study's training and test results are acquired, first there are two type of tests needed to be explained:

- Biased Test Score: Performance of a model on the test portion of dataset it is trained with. This type of score is acquired at the end of every training session by performing detection over 10% (test portion) of the whole dataset which 80% (training portion) is trained on and 10% (validation portion) is used to evaluate for early stopping. So this type of test score is called Biased Test Score because the model has a bias about the test portion of the dataset since it is trained on somewhat similar data.
- Unbiased Test Score: Performance of a model on a dataset which the model is seeing for the first time. This type of score is acquired by performing detection on all of the data of another dataset which the model encounters first time.

Here are the trained model sets that are used in this research:

- English Models: Trained only using English datasets
- Clear English Models: Trained only using no noise part of English datasets
- Complete Models: Trained using all English and multilingual datasets merged
- Clear Complete Models: Trained using all English and multilingual datasets' no noise part merged
- Multilingual Models: Trained only using multilingual datasets
- Clear Multilingual Models: Trained only using no noise part of multilingual datasets

And here are the information about train test sessions:

- All the models that are listed in above six items have Biased Test Scores acquired at the end of their training session. Which indicates how well can a model learn.
- English Models have three Unbiased Test Scores from multilingual datasets' no noise, 0.025 and 0.05 noise types. Which indicates how does an only English learned model perform on other languages. And, same are done also with the Clear English Models to check the augmentations effect on the Unbiased Test Scores.
- English Models have three Unbiased Test Scores from HULSIL's no noise, 0.025 and 0.05 noise types. Which indicates how is the performance affected when audios are generated by a newer AI compared to Unbiased Test Scores from multilingual datasets. And, same are done also with the Clear English Models to check the augmentations effect on the Unbiased Test Scores.
- Multilingual Models have three Unbiased Test Scores from HULSIL's no noise, 0.025 and 0.05 noise types. Which indicates how does multilingual trained models perform on languages that they are not trained on and generated by a newer generative-AI. And, same are done also with the Clear Multilingual Models to check the augmentations effect on the Unbiased Test Scores.
- Complete Models have three Unbiased Test Scores from HULSIL's no noise, 0.025 and 0.05 noise types. Which indicates how does a model, that is trained on a big dataset of various languages, will perform on future attacks. And, same are done also with the Clear Complete Models to check the augmentations effect on the Unbiased Test Scores.

# 3 Results Discussions

## 3.1 Different Detection Cases

Performance of methods are examined focusing only on the languages and the sources of the audios. So there are four different concepts of fake speeches:

- known languages, known generative-AI: which the models performed well on as shown in the 'Biased Test Scores' subtitle.
- unknown languages, unknown generative-AI: which is generally performed poor as shown in the 'Unbiased Test Scores' subtitle. But also there are some models that perform well (above 90 percent accuracy). So there is a lack of information about how to detect these type of audios, which will be cleared in below cases:
- known languages, unknown generative-AI: which is shown under the 'Unbiased Test Scores on HULTL' subtitle. Similarities between these results and other results under the 'Unbiased Test Scores' subtitle indicate that the key ingredient to do detection is not the language. Otherwise, scores under the 'Unbiased Test Scores on HULTL' subtitle would be better than the ones that are under the 'Unbiased Test Scores' subtitle which takes us to item four.
- unknown languages, known generative-AI: which can be tested on HULSIL with models that are trained on HULTL or vice versa, but models are trained on HULTL and tested them on HULSIL since it is also wanted to check models Unbiased Test Scores on the ELTOLSM. Test scores under the 'Hypothesis Test Scores' subtitle show that there is a big difference of performance at testing models on HULSIL and ELTOLSM. These results support the hypothesis of models learning and generalizing over known generative-AI models' traits rather than the languages.

## 3.2 Biased Test Scores (known languages, known generative-AI)

Scores of trainings with augmentation are higher as expected, which indicates that applied noise augmentation enhances the performance in the same domain as the training dataset is in. Also, gradually increasing performance scores of Complete, English and Multilingual training sessions on both types of datasets show that it is harder to generalize in a larger domain even if the dataset is larger. More detailed scores are shown at the Fig. 5 below.

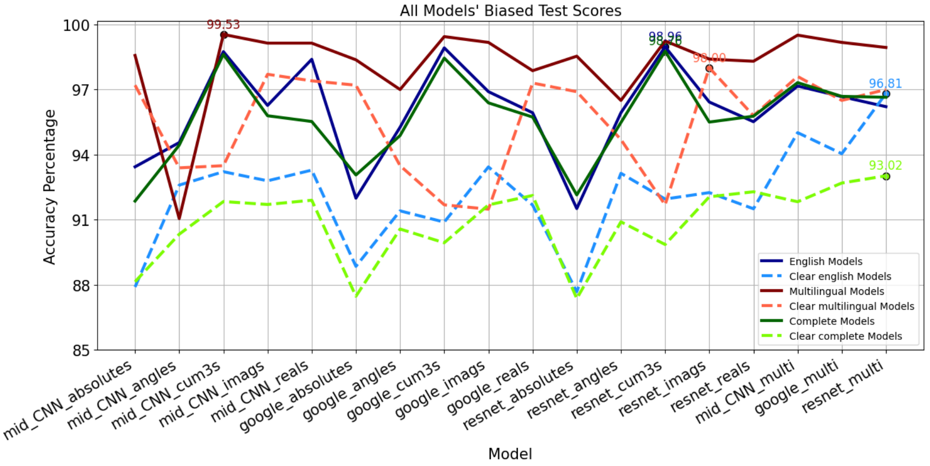

**Fig. 5.** Biased test results of all models

### 3.3 Unbiased Test Scores (unknown languages, unknown generative-AI)

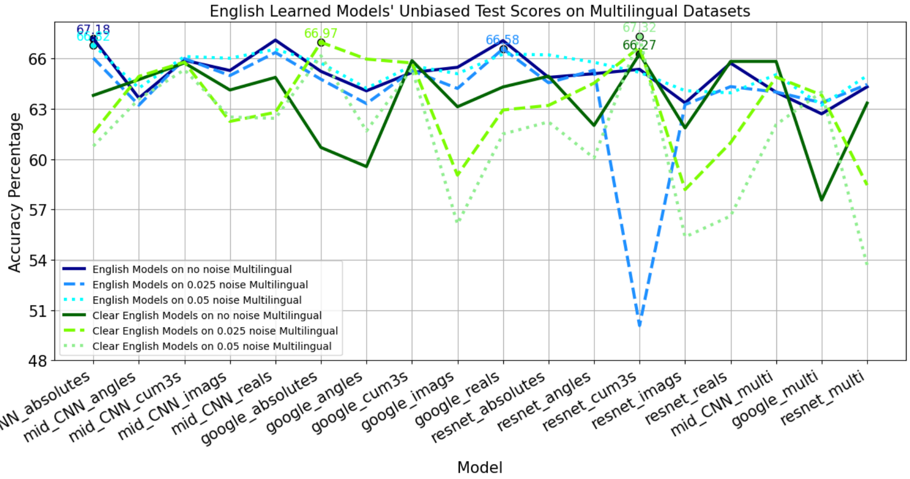

**Fig. 6.** Unbiased test results of english learned models on multilingual datasets

Scores of English Models with augmentation are generally higher than the Clear English Models scores, which indicates that augmentation helps, but not distinctly in a foreign domain. Also, tests on no noise datasets mostly resulted in better scores than noisy datasets. Which supports the claim of noise augmentation making detection harder. More detailed scores are shown at the Fig. 6 above.

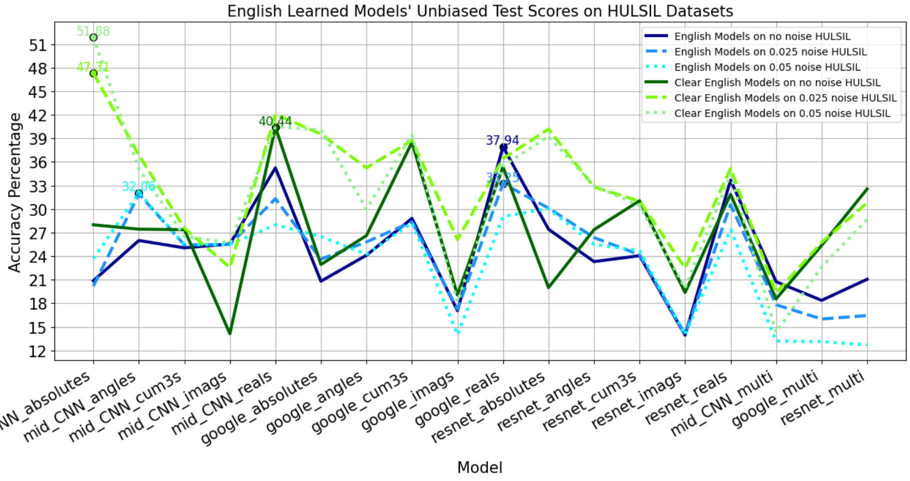

**Fig. 7.** Unbiased test results of english learned models on HULSIL dataset

Best score is just above the fully random decision-making. When other low scores are also examined, it is right to say that HULSIL's audios are considered as real audios to English Models and Clear English Models. Which shows the main purpose of this research: it is not possible to distinguish real and fake audios when they fake ones are generated by an unknown generative-AI in unknown language using bispectrum. It is demonstrated that learning English from older generative-AI models does not give any insight on future tasks. More detailed scores are shown at the Fig. 7 above.

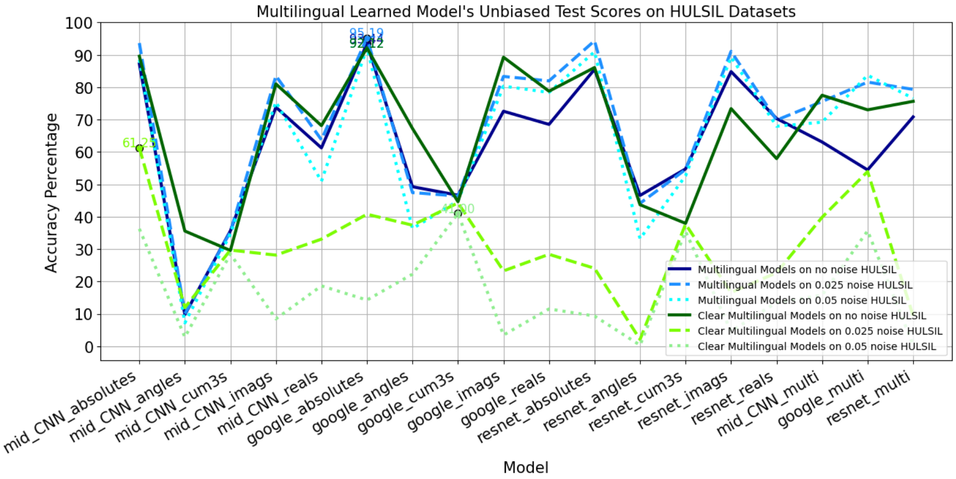

**Fig. 8.** Unbiased test results of multilingual learned models on HULSIL dataset

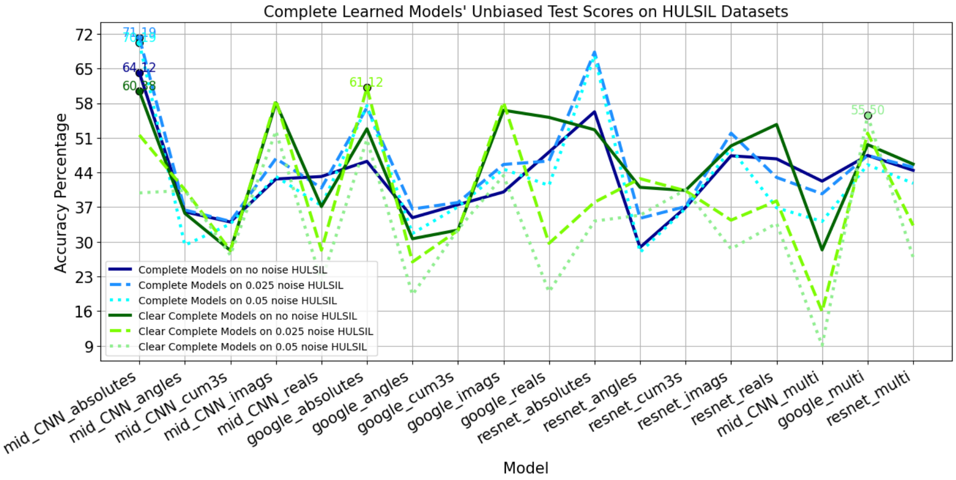

**Fig. 9.** Unbiased test results of complete learned models on HULSIL dataset

English Learned Models' scores are between 12-51 range, Multilingual Learned Models' scores are between 0-93 range and Complete Learned Models' scores are between 9-72 range. Which emphasizes a few points: adding more languages to the training dataset helps with the detection but is not reliable only by itself, since Complete Learned Models' scores are lower than Multilingual Learned Models. Also, since Complete Learned Models are trained on a training dataset that consist mostly English audios and their scores are lower than the Multilingual Learned ones, it is safe to say that balance of the dataset languages also plays an important role when teaching models to generalize. More detailed scores are shown at the Fig. 8 and 9 above.

Also when scores are examined paying attention to the features, downwards spikes are happened at 'angle' and 'cum3' features, while upwards spikes are happened at 'imag' and 'absolute'. And models with multi form input do not have the best scores as expected from the earlier study's [8] findings. Which can indicate few important things that may be overlooked at it. Firstly, features such as 'imag' and 'absolute' may be better to generalize for tasks at foreign domains; secondly, features such as 'angle' and 'cum3' might be redundant, if not harmful, to models with multi form input. This can be also supported when scores on 'cum3' feature are paid attention on. It seems like 'cum3' feature is the least polluted one to human eye when noise augmentation is applied as shown in the feature looks image. Also, test scores of the cum3 models are scattered around smaller areas than the other features. If smaller scattering is related to noise immunity of 'cum3' feature, it is also possible to say that 'cum3' does not carry enough information regarding the speech relations. So noise addition does not have an effect as much as it have on other features and 'cum3' does not provide enough information to generalize on foreign domain.

### 3.4 Unbiased Test Scores on HULTL (known languages, unknown generative-AI)

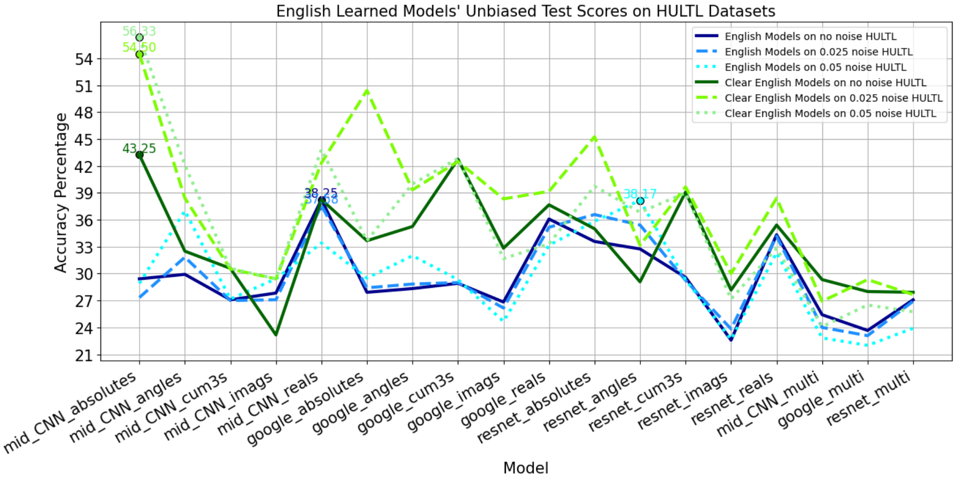

**Fig. 10.** Unbiased test results of english learned models on HULTL dataset

As seen from the tests with HULTL, Unbiased Test Scores are very similar to HULSIL test scores, which indicates that models are unable to detect fake speeches by newer generative-AI models even if language is known. More detailed scores are shown at the Fig. 10 above, Fig. 11 and 12 below. So this takes us to the fourth detection case.

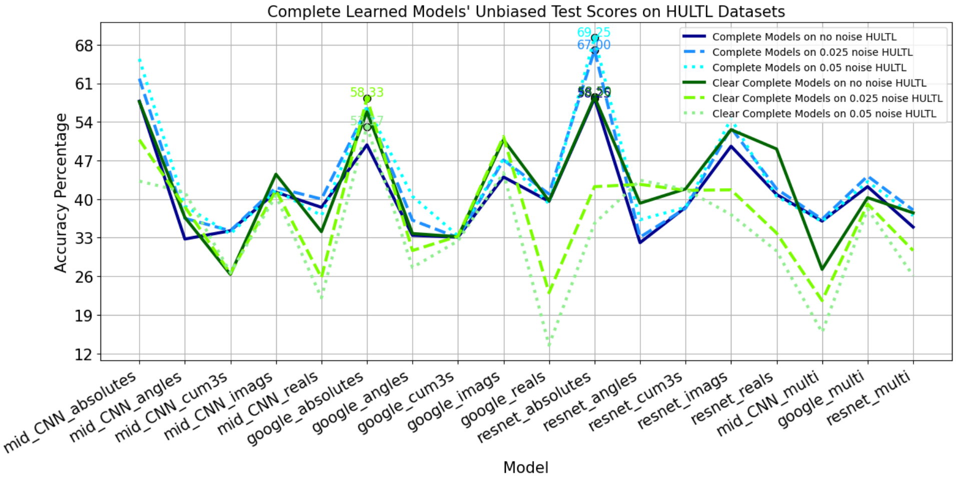

**Fig. 11.** Unbiased test results of complete learned models on HULTL dataset

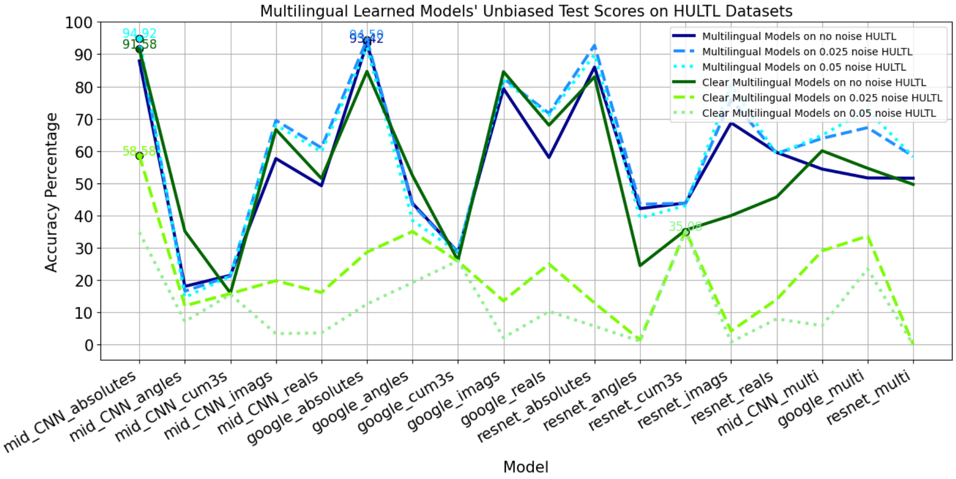

**Fig. 12.** Unbiased test results of multilingual learned models on HULTL dataset

### 3.5 Hypothesis Test Scores (unknown languages, known generative-AI)

So after considering these results here is the hypothesis: A model must be trained on a multilingual dataset, augmentations must be applied have better generalization and the dataset must be balanced across the languages. This way the model can detect any fake speech in the future, ideally.

To prove this hypothesis, new models will be trained on a training dataset that contains 1200 fake audios from HULTL and 1200 randomly selected real audios from CommonLanguage. This model set will be called Hypothesis Models, and below are the scores for Hypothesis Models (Unbiased Test Scores of Hypothesis Models on HULSIL is not completely unbiased since both HULTL and HULSIL are a portion of the same dataset, but it will be named as 'Unbiased' to not introduce another type of test score and to keep terms as simple as can be).

It is expected from Hypothesis Models to perform well on Biased Test Scores, even better than the others since the dataset is smaller. And if they perform well also on Unbiased Test with HULSIL, that means models are able to learn generative-AI models and detect fake speech over detecting known generative-AI models. And to further support the hypothesis of models learning generative-AIs not the languages, Hypothesis Models can be tested on ELTOLSM. If Hypothesis Models perform poorly on ELTOLSM test, that supports the claim, but If Hypothesis Models perform well also on the ELTOLSM test, that means generative-AI models are getting closer to real speech and threshold line for speech detection should be drawn closer to real audios on the speech space from now on. And below, at the Fig. 13, are the Biased Test Scores of the Hypothesis Models and Unbiased Test Scores of Hypothesis Models on HULSIL and ELTOLSM:

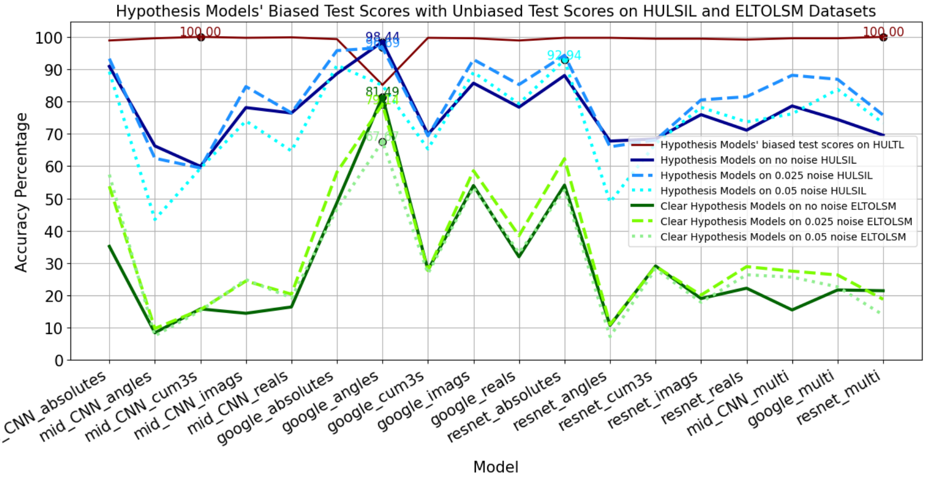

**Fig. 13.** Unbiased test results of hypothesis models on HULSIL and ELTOLSM dataset

As expected, almost all Biased Test Scores are close to perfect 100% accuracy. And Unbiased Test Scores are higher at the HULSIL than the ELTOLSM, as predicted, with a margin of at least 17% and at most 64%. These results support the claim of models learning generative-AI models rather than generalizing over languages. Although there are scores as high as 81% in the ELTOLSM tests, most of them being under the 50% line indicates that, most models are unable to generalize to unknown languages.

When it comes to model sizes, all mid_CNN models are around 3MB, GoogLeNet models are 40MB and ResNet50 models are 90MB. About their performances, there is not any significant difference between, but when other aspects are considered such as training speed, inference speed and memory usage during training, mid_CNN models have the best numbers.

### 3.6 Further Test Scores with Background Noise Augmentation

Also another important test to hold is about the noise augmentation, since only gaussian noise is applied in this research this can only guess/demonstrate the effect of additive noises, and cannot have any insight about the augmentation that is done by putting background noise to audios. Models may be forced to predict the fake speeches as real if background noise from real environments added to them. And even one step further, any generated environmental noise may deceive the models to predict the real audio as fake ones. So another dataset is created by altering the multilingual audios from CommonLanguage and ELTOLSM. A total of 2169 real and fake environmental background noises from different sources [47][48][49][50][51][52] are gathered, and created 'BG Datasets'.

There are two BG datasets, one with adding background noise after scaling it with 0.5 and other with scaling 0.25. Scaling values are selected according to 0.5 being as loud as the speaker and 0.25 being heard and understandable while not suppressing the speaker. Models trained with BG Datasets are called BG Models, and another test session is done with BG, Multilingual and Clear Multilingual Models on BG Datasets to measure the performance of models that are trained with Gaussian Noise augmentation. Below, the Fig. 14, is the table containing Biased Test Scores of BG Models and Unbiased Test Scores of Multilingual and Clear Multilingual Models on BG Datasets:

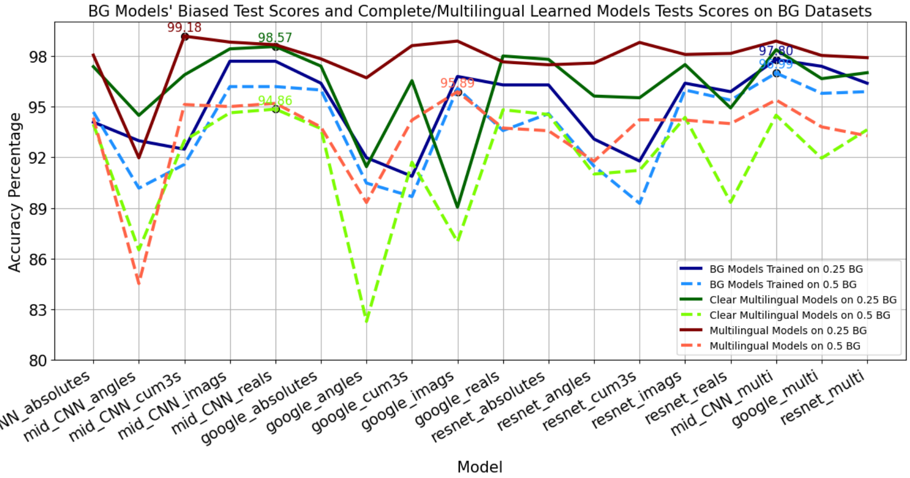

**Fig. 14.** Biased test results of BG models and unbiased test results of complete and multilingual learned models on BG dataset

Successful test scores indicates that it is possible to distinguish audios with added decisive background noises. One important conclusion is that the Multilingual Models' performances suppressed the BG Models' performances on the dataset. Which is quite interesting that BG Models are unable to generalize to a dataset that they are trained on, as good as Multilingual Models, which trained on the same audios, but with only difference of augmentation methods. It is possible to say that Gaussian Noise is more regularizing than a synthetic augmentation that is achieved by altering the data with other real world audios. Also as shown from the figure, models' performances on the BG dataset with 0.25 scale is always better than the same model's performance on dataset with 0.5 scale, which supports the previous claims about noise augmentation.

## 4 Future Work

More diverse features could be experimented with such as MFCC, power spectrums or even other higher-order spectrums to detect correlations beyond second- and third-order. For example, another feature to check may be trispectrum which is sensitive to fourth-order correlations. And below, at the Fig. 15, is the formula for trispectrum:

$$T(\omega_1, \omega_2, \omega_3) = Y(\omega_1)Y(\omega_2)Y(\omega_3)Y^*(\omega_1 + \omega_2 + \omega_3)$$

**Fig. 15.** Formula for fourth order correlations

And in general the Nth-order polyspectrum is sensitive to (N+1)st-order correlations. These polyspectra can also be computer for higher-dimensional signals. For example, a 2D signal with Fourier transform Y(ωx, ωy) has a 4D bispectrum as shown in the Fig. 16 below:

$$\begin{aligned} B(\omega_{x1}, \omega_{y1}, \omega_{x2}, \omega_{y2}) &= \\ & Y(\omega_{x1}, \omega_{y1})Y(\omega_{x2}, \omega_{y2})Y^*(\omega_{x1} + \omega_{x2}, \omega_{y1} + \omega_{y2}) \end{aligned}$$

**Fig. 16.** 4D bispectrum of the 2D signal with Fourier transform

More advanced models can be trained. Instead of CNNs, models with better understanding of sequences such as MAMBA [53] models can be used. And with these sequenceable models, detection task can change from "deciding over the whole signal" to "sectional decisions". Better way to understand these use-cases is demonstrated and experimented with at one of the previous researches [8] with multiple experiments.

Also, more dynamic models could be trained, models trained for this research are trained only on the features that are extracted from 16 Kilohertz audios' segments with 400 sample. Used segments size for other audios with different sample rates would be different. It would be 552 for 22050, 600 for 24000 and 1200 for 48000 Hertz sample rates. Training a model on features from dynamic sample rates could also improve the performance, since downsampling audios to 16 Kilohertz may result in information loss.

Data augmentation can be made more effective with various additive noises and the methods given in the "Background Noise Augmentation" section.

## 5 Conclusion

In this work, multiple models and datasets are presented to find a possible solution to the problem of detecting generated speech from real ones, especially the ones that are from unknown generative models in unknown languages. Best obtained test scores for different test scenarios are 94.92 for unknown model-known language, 98.44 for known model-unknown language and 95.18 for unknown model-unknown language .

Even though they are clearly successful scores, there are more scores that can be considered ”unsuccessful”. For the ground of this ”unsuccessful” results, few reasons are proposed and tested to find out what does a ”successful” result come out of. To test the presented hypothesis, different training and testing sessions on newly created datasets are experimented with. And as a result, it is found that models are not learning to discriminate speeches by generalizing over how languages sound or are spoken, but how generative-AI models sound uniquely.

## References

1. Ashish Vaswani et al., ”Attention Is All You Need”, June 2017
2. Aditya Ramesh et al., ”Zero-Shot Text-to-Image Generation”, February 2021
3. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer, ”High-Resolution Image Synthesis with Latent Diffusion Models”, April 2022
4. Yixin Liu et al., ”Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models”, February 2024
5. Veo-Team et al., ”Veo 2”, 2024, https://deepmind.google/technologies/veo/veo-2/
6. Matthew Le et al., ”Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale”, October 2023
7. OpenAI, ”GPT-4o System Card”, August 2024
8. [Redacted for Blind Review], ”AI Generated Speech Detection Using CNN”, January 2025
9. Md Sahidullah, Tomi Kinnunen and Cemal Hanilçi, ”A Comparison of features for synthetic speech detection”, September 2015
10. Zaynab Almutairi and Hebah Elgibreen, ”A review of modern audio deepfake detection methods: challenges and future directions”, May 2022
11. Jiangyan Yi et al., ”Audio deepfake detection: a survey”, Aug 2023
12. Hany Farid, ”detecting digital forgeries using bispectral analysis”, August 2001
13. Alessandro Pianese, Davide Cozzolino, Giovanni Poggi and Luisa Verdoliva, ”Deepfake audio detection by speaker verification”, September 2022
14. Ehab A. AlBadawy, Siwei Lyu and Hany Farid, ”Detecting AI-synthesized speech using bispectral analysis”, June 2019
15. Arun Kumar Singh and Priyanka Singh, ”Detection of AI-synthesized speech using cepstral & bispectral statistics”, April 2021
16. Ameer Hamza et al., ”Deepfake audio detection via MFCC features using machine learning”, January 2022
17. Tianyun Liu, Diqun Yan, Rangding Wang, Nan Yan and Gang Chen, ”Identification of fake stereo audio using SVM and CNN”, June 2021
18. Mohammed LataifehNassif, Ashraf Elnagar, Ismail Shahin and Ali Bou Nassif, ”Arabic audio clips: Identification and discrimination of authentic cantillations from imitations”, December 2020
19. Devesh Kumar, Pavan Kumar V. Patil, Ayush Agarwal and S. R. Mahadeva Prasanna, ”Fake Speech Detection Using OpenSMILE Features”, November 2022
20. Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar and Faruk Kazi, ”A Deep learning framework for audio deepfake detection”, November 2021
21. Chengzhe Sun, Shan Jia, Shuwei Hou and Siwei Lyu, ”AI-synthesized voice detection using neural vocoder artifacts”, April 2023
22. Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman and Elie Khoury, ”Generalization of audio deepfake detection”, April 2023
23. Yipin Zhou and Ser-Nam Lim, ”Joint audio-visual deepfake detection”, October 2021
24. Raghav Magazine, Ayush Agarwal , Anand Hedge and S. R. Mahadeva Prasanna, ”Fake Speech Detection Using Modulation Spectrogram”, November 2022

25. Steven Camacho, Dora Maria Ballesteros, and Diego Renza, ”Fake Speech Recognition Using Deep Learning”, September 2021
26. Hafiz Malik and Raghavendar Changalvala, ”Fighting AI with AI Fake Speech Detection using Deep Learning”, June 2019
27. Nishant Subramani and Delip Rao, ”Learning Efficient Representations for Fake Speech Detection”, April 2020
28. Zhenyu Zhang, Xiaowei Yi and Xianfeng Zhao, ”Fake Speech Detection Using Residual Network with Transformer Encoder”, June 2021
29. Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar and Konstantin Böttinger, ”Does audio deepfake detection generalize?”, Interspeech, 2022
30. Ricardo Reimao and Vassilios Tzerpos, ”FoR: A Dataset for Synthetic Speech Detection”, October 2019
31. Frank, J., & Schönherr, L. (2021). WaveFake: A data set to facilitate audio DeepFake detection (1.2.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.5642694
32. Yamagishi, Junichi; Veaux, Christophe; MacDonald, Kirsten. (2019). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645.
33. Ganesh Sinisetty, Pavlo Ruban, Oleksandr Dymov, & Mirco Ravanelli. (2021). CommonLanguage (0.1) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.5036977
34. ”Common Voice dataset by Mozilla”, https://commonvoice.mozilla.org/en/datasets
35. ”ELTOLSM Dataset”, https://drive.google.com/drive/u/1/folders/1SVSou6rZkQYgmZhVCFCOj6bPEkrZrBvT, last accessed: 15/03/2025
36. OpenAI, https://chatgpt.com/, 2022
37. ”bark”, 2023, https://github.com/suno-ai/bark
38. Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler and Nima Mesgarani, ”StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models”, November 2023
39. The Coqui TTS Team, ”A deep learning toolkit for Text-to-Speech, battle-tested in research and production”, January 2021, https://github.com/coqui-ai/TTS
40. ”VALL-E X: Multilingual Text-to-Speech Synthesis and Voice Cloning”, 2023, https://github.com/Plachtaa/VALL-E-X
41. Qin, Zengyi and Zhao, Wenliang and Yu, Xumin and Sun, Xin, ”OpenVoice: Versatile Instant Voice Cloning”, 2023, arXiv preprint arXiv:2312.01479
42. https://www.synthesia.io/features/ai-voice-generator
43. https://play.ht/
44. ”HULTEL Dataset”, https://drive.google.com/drive/u/6/folders/1EXysipBgs3tpQOTU7hL_3RPjB3ZqDkIW, last accessed: 15/03/2025
45. Gemini Team Google, ”Gemini: A Family of Highly Capable Multimodal Models”, December 2023, https://gemini.google.com/app
46. Vineel Pratap et al., ”Scaling Speech Technology to 1,000+ Languages”, May 2023, https://ai.meta.com/blog/multilingual-model-speech-recognition/
47. K. J. Piczak. ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia, 2015.
48. https://freesound.org/people/FALL-E/packs/38822/
49. https://freesound.org/people/FALL-E/packs/38926/
50. https://freesound.org/people/DataJuggler/packs/41547/
51. https://www.gaudiolab.com/technology/fall-e
52. Zach Evans et al., ”Stable Audio Open”, July 2024, https://stability.ai/news/introducing-stable-audio-open
53. Albert Gu and Tri Dao, ”Mamba: Linear-Time Sequence Modeling with Selective State Spaces”, May 2024

## Authors

**Meriç Demirörs** received his BSc in AI Engineering. Studying at TU Darmstadt AI&ML Master's. His research interests include computer vision, deep learning, signal processing.

**Ahmet Murat Özbayoğlu** received his PhD in Electrical and Electronics Engineering. He is a professor at the Department of Computer Engineering at TOBB ETÜ. His research interests include deep learning, computer vision, and biomedical signal processing.

**Toygar Akgün** received his PhD in Computer Science. He is currently a faculty member at the Department of Computer Engineering at TOBB ETÜ. His research interests include image processing, pattern recognition, and artificial intelligence.