Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics

Arun Kumar Singh
Indian Institute of Technology, Jammu
Jammu J&K, India
2019PCS0017@iitjammu.ac.in

Priyanka Singh
Dhirubhai Ambani Institute of
Information and Communication Technology
Gandhinagar, Gujarat, India
priyanka_singh@daiict.ac.in

Abstract—**Digital technology has made possible unimaginable applications come true. It seems exciting to have a handful of tools for easy editing and manipulation, but it raises alarming concerns that can propagate as speech clones, duplicates, or maybe deep fakes. Validating the authenticity of a speech is one of the primary problems of digital audio forensics. We propose an approach to distinguish human speech from AI synthesized speech exploiting the Bi-spectral and Cepstral analysis. Higher-order statistics have less correlation for human speech in comparison to a synthesized speech. Also, Cepstral analysis revealed a durable power component in human speech that is missing for a synthesized speech. We integrate both these analyses and propose a model to detect AI synthesized speech.**

Index Terms—**AI-synthesized speech, Bi-spectral Analysis, Higher Order Correlations, Cepstral Analysis, MFCC, Multimedia Forensics.**

I. INTRODUCTION

In today's era, advances in Artificial Intelligence and Deep Neural Networks have led to very significant results in creating a more realistic type of synthesized audio and speech [2], [4]. Speech cloning and duplication via training the neural networks using powerful AI algorithms lead to synthesized speech. However, these advancements had also resulted in many misuses of this technology. Many dangerous fakes have been created, which can harm in many possible ways. Hence, the authenticity of digital data is the prime concern for all of us.

The field of speech forensics has progressed a lot over the past few decades, but very few schemes have proposed detecting AI synthesized content and speech. Some techniques address speech spoofing [6] and tampering, but they are not explicit for detecting AI synthesized speech. Hany discussed how the forgery of a signal affects the correlations of the higher-order statistics [1] but not discussed AI synthesized content.

A comparison between the different features for synthetic speech, spectral magnitudes, and phase of the statistical measures presented in [3] to distinguish the human speech from the AI synthesized speech. First-order Fourier coefficients or second-order power spectrum correlations can be easily tuned to match the human speech while synthesizing speech. On the other hand, third-order bispectrum correlations are hard to adjust and can discriminate between human and AI speech

[7]. One such work was proposed in [7] where they used bicoherence magnitude and phase to distinguish human speech from AI speech collected from various sources like Google AI, Siri, Baidu etc. Muda et al. [8] presented the distinction between male and female speeches using Mel Frequency Cepstral Coefficient (MFCC). MFCC are useful features to identify vocal tracts. Synthetic speech detection using the temporal modulation technique presented in [9] also used MFCC as one of their features. However, they did not include two primary features related to the MFCC: Δ-Cepstral and Δ^2-Cepstral.

The motivation of the proposed algorithm comes from the urgent need of more robust features for discriminating AI synthesized speech from human speech. In the proposed work, we have considered synthesizers that are not too famous but readily available to common users. We combine multiple primary features, specifically higher-order spectral correlations revealed by Bispectral analysis, and MFCC, Δ-Cepstral and Δ^2-Cepstral revealed by Mel Cepstral analysis, to account for the enhanced discrimination between human speech and AI synthesized speech. Bispectral analysis can identify components in AI synthesized speech that are absent in human speech [7]. The process of AI synthesis may induce some correlations in the AI synthesized speech as it passes through different layers of neural network. These correlations are hard to remove as they are likely to be generated due to the fundamental properties of the synthesis process [1]. These correlations are lacking in the recorded human speech. Mel Cepstral analysis of the speech reveals strong power components in human speech, which is not present in the AI synthesized speech. These power components may be present in human speech due to the vocal tract, which in contrast, is not the case with AI-Synthesised speech. We integrate Bispectral analysis and Mel Cepstral analysis and perform various experiments to test the efficacy for classifying the Human and AI speech. The audio data we accounted for, in our experiments, is uncleaned audio data which may contain noise and interference as we get in real life scenarios.

The rest of the paper is organized as follows. Section II gives a brief overview of the key concepts used in the proposed algorithm, section III provides the details of the dataset, classification model and parameters used. Section IV discusses the result findings of the proposed method.

II. PRELIMINARIES

In this section, a brief overview of the higher-order statistics is given that we have used in our proposed algorithm as distinguishing features. Analysis of Mel Frequency Cepstral Coefficient (MFCC) and visualization using the Mel spectrogram is described. Delta and Delta Square related to Mel Cepstrum is briefed.

A. Bispectral Analysis

The bispectrum of the signal represents a higher-order correlation in the Fourier domain. The simple Fourier coefficients represent the first-order correlation or first-order statistics.

An audio signal $y(k)$ is decomposed into different frequencies according to the Fourier transform :

$$Y(\omega) = \sum_{k=-\infty}^{\infty} y(k).e^{-ik\omega} \quad (1)$$

with $\omega \; \epsilon \; [-\pi, \pi]$. Power spectrum of the signal $P(\omega)$ is generally used to detect second order correlations, given by : $P(\omega) = Y(\omega).Y^*(\omega)$ where $*$ denotes the complex conjugate. Power spectrum is blind to higher order correlations that means we cannot detect higher order correlations using power spectrum. However, these higher order correlations can be detected using bispectral analysis. We find bispectrum of the signal to calculate third order correlations which is given by:

$$B(\omega_1, \omega_2) = Y(\omega_1).Y(\omega_2).Y^*(\omega_1 + \omega_2) \quad (2)$$

Unlike the power spectrum, the bispectrum in the Equation (3) is a complex valued quantity. So for the purpose of simplicity and interpretation for our problem, it is suitable to represent or use the complex bispectrum with respect to it's magnitude :

$$\mid B(\omega_1, \omega_2) \mid = \mid Y(\omega_1) \mid . \mid Y(\omega_2) \mid . \mid Y(\omega_1 + \omega_2) \mid \quad (3)$$

and Phase :

$$\angle B(\omega_1, \omega_2) = \angle Y(\omega_1) + \angle Y(\omega_2) - \angle Y(\omega_1 + \omega_2) \quad (4)$$

Also for the purpose of scaling and simplicity in calculations, it will be helpful to use the normalized bispectrum [5], the bicoherence :

$$B_c(\omega_1, \omega_2) = \frac{Y(\omega_1).Y(\omega_2).Y^*(\omega_1 + \omega_2)}{\sqrt{\mid Y(\omega_1).Y(\omega_2) \mid^2 . \mid Y(\omega_1 + \omega_2) \mid^2}} \quad (5)$$

This normalized bispectrum yields magnitude in the range $[0, 1]$. But we have used the other normalized process for bispectral magnitude and phase which also yields the range into $[0, 1]$.

For the purpose of efficient calculation, we divide each speech samples of length N into approximately $K \approx 100$ smaller samples of length N/K. The bispectral magnitude and phase of these K segments are summed over different ω values and average value is taken.

$$\mid \widehat{B}(\omega_1, \omega_2) \mid = \frac{1}{K} \sum_K (\mid Y_K(\omega_1) \mid\mid Y_K(\omega_2) \mid\mid Y_K(\omega_1 + \omega_2) \mid) \quad (6)$$

$$\angle \widehat{B}(\omega_1, \omega_2) = \frac{1}{K} \sum_K (\angle Y_K(\omega_1) + \angle Y_K(\omega_2) - \angle Y_K(\omega_1 + \omega_2)) \quad (7)$$

B. Mel Frequency Cepstral Coefficient (MFCC) and Analysis

The speech generated by humans can be uniquely identified due to the vocal tract's shape that includes human oral organs, during the speaking. In general, these different vocal tract shapes helps in determining the type of sound that we speak. The Mel Frequency Cepstral represents the short-time power spectrum of the audio. It can be used to filter speech based on vocal tract shape. This is represented using MFCCs. On a similar hypothesis, there will be differences in MFCCs values of Human speech and AI-synthesized speech as the speech generated by AI are not generated from vocal tracts.

For our study, we have considered four types of speech. The first one is human speech, and the other three are AI synthesized speech from three different sources, namely Spik.AI, Natural Reader, and Replica AI. These three kinds of AI speech are generated from different text to speech-generating AI engines. Mel spectrogram for the four types of speeches is represented in Figure 2.

The MFCCs are calculated from the magnitude spectrum of short term Fourier transform of the audio signal. The short term Fourier transform of the audio signal $y(k)$ is given by :

$$Y(\omega) = \mid Y(\omega) \mid e^{\; j\phi(\omega)} \quad (8)$$

where $\mid Y(\omega) \mid$ is magnitude spectrum and $\phi(\omega)$ is the phase spectrum. For calculating MFCC, the entire speech is split into overlapping segments called windows. After that, a Fourier transform is performed for each segment, which is used to derive the power spectrum. Mel Frequency Filter is applied to the power spectrum obtained, and then discrete cosine transform (DCT) of the Mel log power is taken. The MFCCs represents the amplitude of the obtained spectrum after DCT is performed.

Other parameters associated with MFCC useful as a feature for the distinction of speech are Δ-Cepstrum and Δ^2-Cepstrum. Change in MFCC coefficients is given by Δ-Cepstrum, i.e., Δ-Cepstrum is the difference between the current MFCC coefficient and the previous MFCC coefficient. Similarly, Change in Δ-Cepstrum values is given by Δ^2-Cepstrum, i.e., Δ^2-Cepstrum is the difference between current Δ-Cepstrum value and the previous Δ-Cepstrum value. All these three parameters act as strong traits for representing Cepstral Analysis.

III. EXPERIMENTS AND ANALYSIS

In this section, we have described the data set created by us by collecting speech samples from different sources. Some relics for the observations, the classification model, and the parameters used in the study are also described.

A. Data Set

We collected speech samples from different sources and created a data set for our experiment. Our data set comprises of
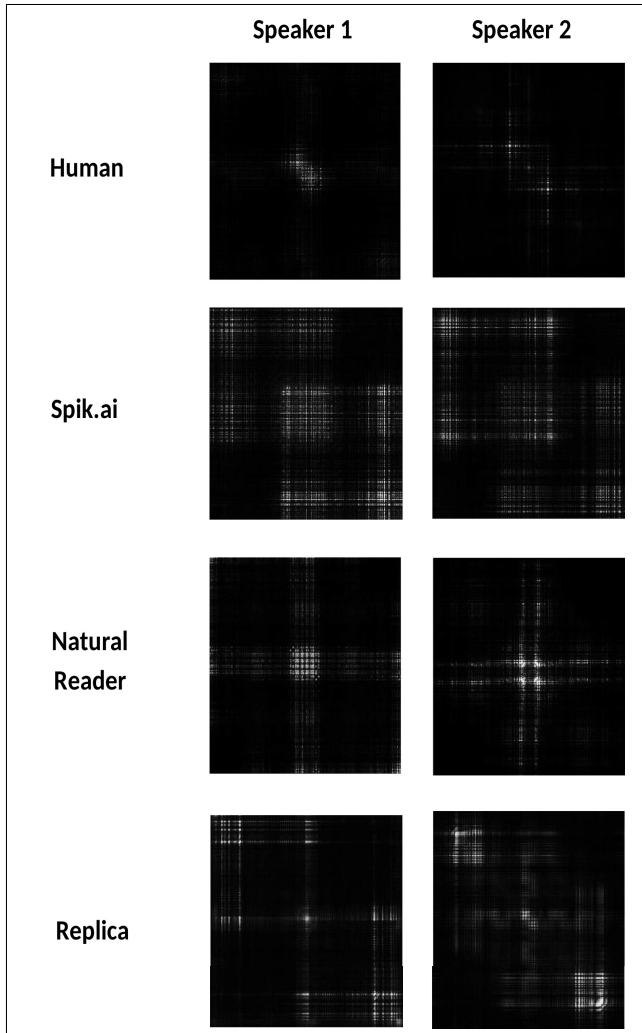
413

Fig. 1. Bicoherent magnitude of two speakers for human and three synthesized speech. The magnitude plot is shown on an intensity of the scale $[0, 1]$

total 704 different samples. For human speech, we have a total of 250 samples, where 110 speech samples are taken from the Kaggle data repository, and 140 speech samples are recorded with the microphone. For diversity in our data set, we took human speech samples from both male and female speeches. For AI synthesized speech, we have collected samples from three different sources: Natural Reader, Spik.AI, and Replica. We have taken 79 samples from Natural Reader, 230 samples from Spik.AI, and 126 samples from Replica AI. These AI synthesized speech engines are based on text-to-speech synthesis. We collected the sample from these not too famous synthesizers because, the AI advancement has increased to so much extent that these types of unnoticed sources has increased a lot and readily available for use. And the old techniques were limited to the testing over some famous synthesizers only. Before processing each speech sample (both AI synthesized and human), it is trimmed to average of 5 secs each. Since audios exist in two channels, i.e., mono and stereo,

we converted all speech samples into monotype for the equal basis of comparison. These samples were then distributed randomly into 622 training and 82 testing data. Also we didn't pre-process our data so the noise and interference are readily available. The reason behind to use unprocessed data is that, in real life applications we don't get processed data and we want to test how much our features (independently and combined) can be effective over unprocessed audio data.

B. Relics for the Observations

We demonstrated some of the preliminary results in the form of images in Figure 1, showing the different patterns achieved after processing different speeches. The bicoherence magnitude for two speakers of each category is represented in Figure 1. The first row represents the relic of human speech. Similarly, second, third, and fourth row represent relics from AI synthesized speech, i.e., Spik.AI, Natural Reader, and Replica AI. Similarly, the columns represent the normalized bispectral magnitude for speaker 1 and speaker 2 for the corresponding speeches. All images are shown on the same intensity scale.

On observing the relics, we can see a glaring difference between the magnitude of human speech and all other AI synthesized speech. These variations can be due to significant spectral correlations present in AI synthesized speech but absent in human speech. These spectral correlations in AI synthesized speech are induced due to neural network architecture. In particular, the long-range temporal connections between the layers of neural networks may be the cause.

Four Mel spectrograms of different four speeches, i.e., Spik.AI, Replica, Natural Reader, and Human speech is shown in Figure 2. Mel power scale is given on the right of each spectrogram. From reference from scale we can see that dark blue colour indicate weak power component and yellow colour represents strong power component in the spectrogram 2. It is observed that a strong power component is missing in all types of AI-synthesized speech, which is not in the case of human speech. It may be due to the absence of vocal tract during the generation of AI-synthesized speech. These differences shown in the Mel spectrogram encourage us to add MFCC as a feature for discriminating speeches.

C. Classification Models, Parameters, Scenarios and Test Cases

The bicoherence magnitude and phase are calculated, as mentioned in Equation 6 and Equation 7, respectively, for all human speech samples and three types of AI synthesized speech. These quantities are then normalized to the range [0,1] and then used to calculate the machine learning model's higher-order statistical parameters.

We calculate the first four statistical moments for both magnitude and phase. Let M and P be the random variables denoting the underlying distribution of bicoherence magnitude and phase. The first four statistical moments are given by:
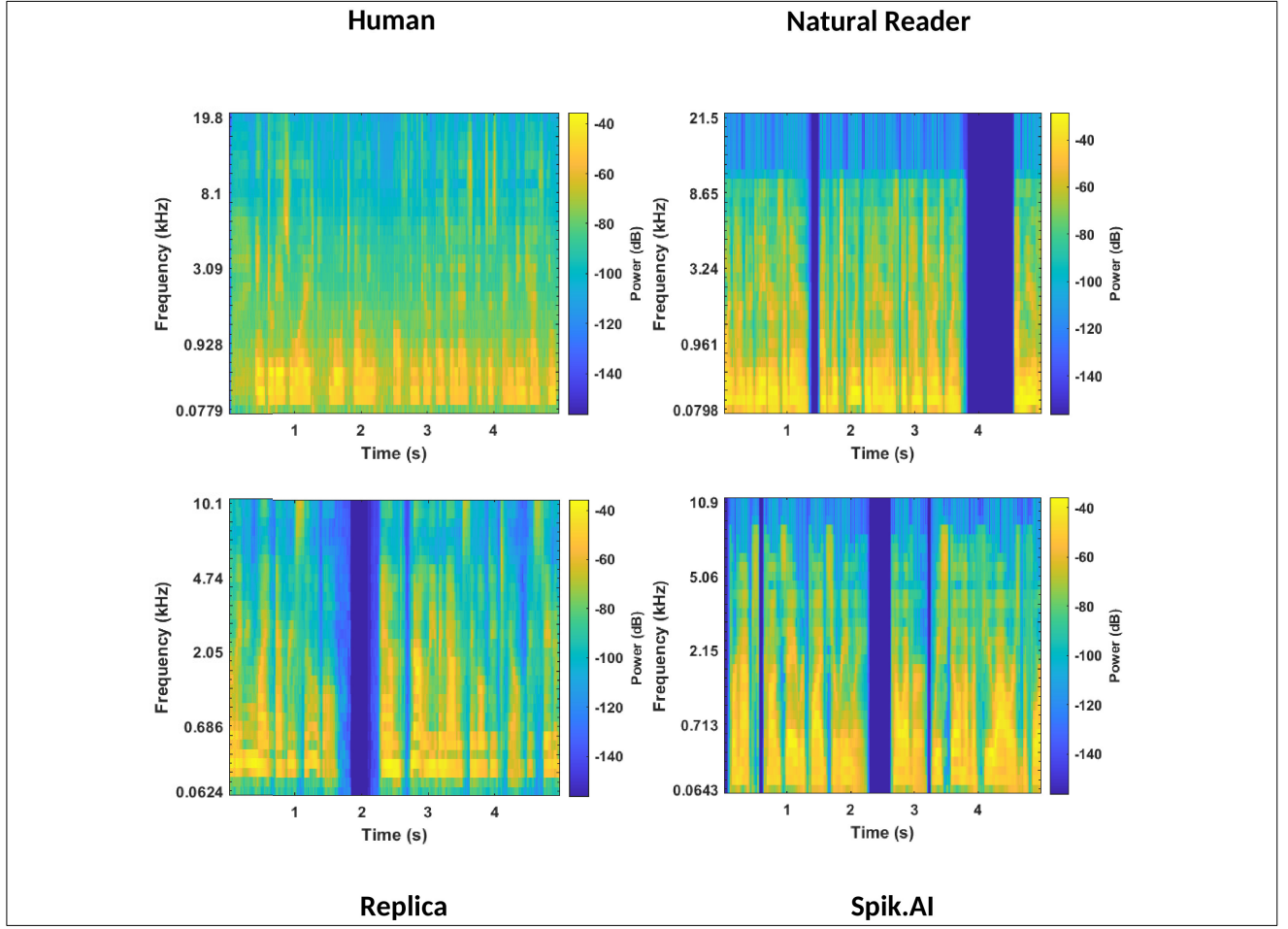
Fig. 2. Melspectrogram for different four types of speeches

- Mean , $\mu_X = E_X[X]$

- Variance , $\sigma_X = E_X[(X - \mu_X)^2]$

- Skewness , $\gamma_X = E_X[(\frac{X - \mu_X}{\sigma_X})^3]$

- Kurtosis , $\kappa_X = E_X[(\frac{X - \mu_X}{\sigma_X})^4]$

where $E_X[.]$ is the expected value operator with regards to random variable X. For the magnitude, we represent X = M, and for the phase X = P, these four moments are calculated by replacing this expected value operator with average. Also, for each speech sample, the mean and variance of MFFC, Δ-Cepstrum, and Δ^2-Cepstrum are calculated. It contributes to a 15-D feature vector for each speech sample. The first 8 entries represent the four moments for magnitude and phase. The next 6 entries represent the mean and variance of MFCC, Δ-Cepstrum, and Δ^2-Cepstrum and last entry represent the class of speech, i.e., Human, Natural Reader, Spik.AI or Replica.

We perform experiments considering following scenarios:

- Scenario 1: In this setup, we classify speech samples into two classes i.e., Human vs. AI-synthesized (Natural Reader, Spik.AI or Replica). It is a binary classification and main focus of this paper.
- Scenario 2: In this experiment, we classify speech samples into multiple classes i.e., Human, Natural Reader, Spik.AI and Replica.

The 15-D feature representation for these experiment scenarios differs only in the last entry. For scenario 1, the last entry can represent either of the two values, i.e., Human or AI synthesized. However, for scenario 2, it can represent any of the four classes, i.e., Human, Natural Reader, Spik.AI, or Replica. We perform machine learning-based classification for both the experiment scenarios with the feature mentioned earlier, with the intuition of different expected outcomes.

We calculated Area Under the Curve (AUC) for ROC of different classifiers for both the scenarios mentioned in Table I. We did the experiments by using many different classifiers for the below two cases to justify how adding MFCC, Δ-Cepstrum, and Δ^2-Cepstrum into third-order bispectrum correlations can act as a set of good features:

- Case 1: Considering only Bicoherence Magnitude and

Various Models	Multi-Class		Binary Class	
	Bicoherence Magnitude & Phase	Bicoherence Magnitude & Phase + MFCC & Delta Cepstral + Delta Square Cepstral	Bicoherence Magnitude & Phase	Bicoherence Magnitude & Phase + MFCC & Delta Cepstral + Delta Square Cepstral
Linear Discriminiant	0.81	0.99	0.79	0.99
Quadratic Discriminant	0.63	0.98	0.71	0.98
Linear SVM	0.80	0.99	0.80	0.98
Quadratic SVM	**0.83**	**0.99**	**0.84**	**0.99**
Weighted KNN	0.82	0.99	0.82	0.99
Boosted Trees Ensemble	0.84	0.98	0.84	0.96
Logistic Regression	–	–	0.79	0.99

Phase as a feature for classification, similar to the experiment of [7].

- Case 2: Adding MFCC, Δ-Cepstrum, and Δ^2-Cepstrum along with Bicoherence Magnitude and Phase.

By visualizing the data for both types of classification and based on our intuition, we tried a few of the learning algorithms to train the Machine Learning model for classifying samples. For binary classification, we did experiment with seven ML algorithms for training, i.e, Linear and Quadratic Discriminant, Linear and Quadratic SVM, Weighted KNN, Boosted Trees and Logistic Regression with 5-fold cross-validation. For multi-class classification, we did experiment with six ML algorithms for training, i.e., Linear and Quadratic Discriminant, Linear and Quadratic SVM, Weighted KNN and Boosted Trees with same 5-fold cross-validation.

IV. RESULTS AND DISCUSSIONS

After testing each classifier's performance by 5-fold cross-validation, we found that the binary classification Quadratic Support Vector Machine(Q-SVM) algorithm-based machine learning model has the highest accuracy with 96.3%. For multi-class classification, also the Quadratic Support Vector Machine(Q-SVM) algorithm-based machine learning model has the highest accuracy with 96.1%. We choose Quadratic SVM over all other classifiers though there was a marginal difference because the Kernel trick in SVM gives us an edge over all other classifiers. The Q-SVM model we used in our experiment had kernel function set to quadratic and kernel scale set to 2. During our experiments, we also observed that in a binary classification setting up kernel scale to 3 increases the training accuracy to 97.6% and testing accuracy to 98.4%. But that doesn't show a good result with multi-class classification. Hence, for both binary class classification and multi-class classification, we chose Quadratic SVM as our classifier for constructing the trained ML model. The classifier's exact accuracy can be visualized with a confusion matrix, which indicates how many true values are falsely recognized. The confusion matrix for both the classifier Quadratic SVM binary
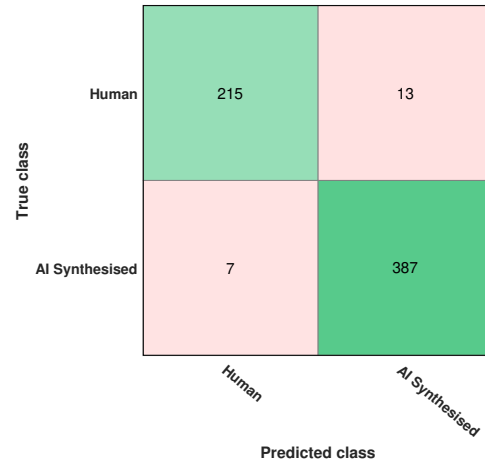

Fig. 3. Confusion matrix for binary classification on training data set

class and Quadratic SVM multi-class is presented in Figure 3 and Figure 4, respectively.

By observing both the confusion matrix, we can see that binary classification results are more potent than the multi-class classification. In multi-class classification, we have false positives between Natural Reader and Spik.AI, Spik.AI and Replica, etc., which will be treated inside one class, i.e., AI synthesized in binary class classification. In the confusion matrix, we can observe that many of the Natural Reader speech samples are misclassified as Spik.AI samples. This may be due to their similarity in neural network engines. However, this thing does not matter much in the binary classification because all AI synthesized speech is classified as AI synthesized irrespective of what neural network architecture they follow. That is why binary classification between Human speech vs. AI synthesized speech has much better accuracy on cross-validation than multi-class classification.

After choosing our classifier and training the model, we tested our test data set over the trained model for the pre-
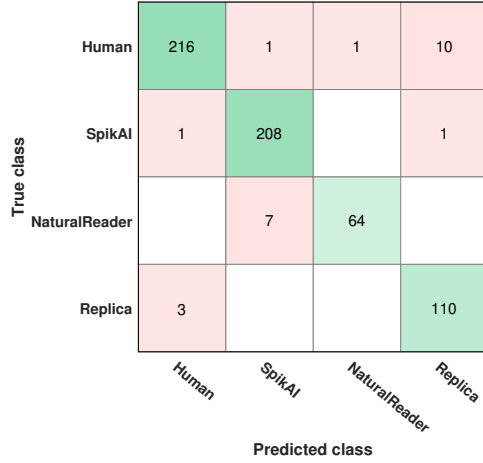
416

Fig. 4. Confusion matrix for multi-class classification on training data set

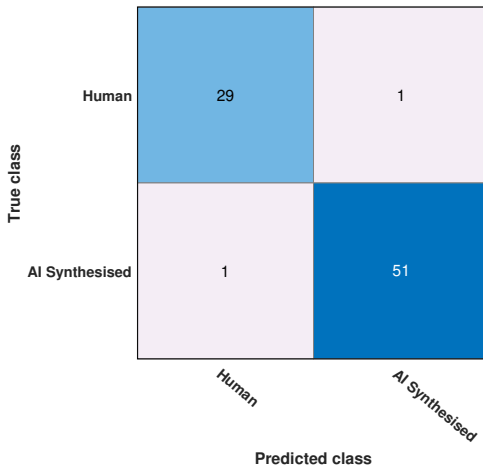Fig. 6. Confusion matrix for multi-class classification on test data set

Fig. 5. Confusion matrix for binary classification on test data set

diction. The test data consist of 82 random samples from all four types of speeches. Prediction is performed for both binary class classification and multi-class classification. Figure 5 and 6 represents the confusion matrix for our final result of predicted data on binary classification and multiclass classification respectively.

The accuracy achieved on the prediction of test data on binary classification, which was our main motto, is 97.56% with a miss classification rate of 2.43%. For multi-class classification,we achieve accuracy of 93.9%, with a miss classification rate of 6.09%.

We plan to study and integrate other discriminatory features to improve upon the accuracy and decrease the miss classification rate for our future work. Also, the scalability of the proposed model can be validated by testing with more massive datasets. More variants of experiment scenarios like classification based on gender, age, and accent can be done.
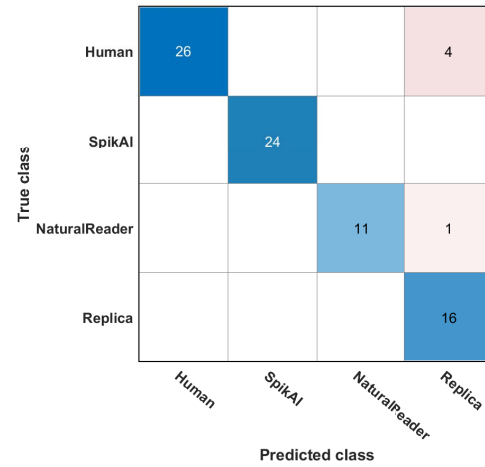
REFERENCES

[1] Hany Farid. Detecting digital forgeries using bispectral analysis. Technical Report AI Memo 1657, MIT, June 1999. 2
[2] Yu Gu and Yongguo Kang. Multi-task WaveNet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions. In Interspeech, Hyderabad, India, 2018. 1
[3] Md Sahidullah, Tomo Kinnunen, and Cemal Hanilci. A comparison of features for synthetic speech detection. In Interspeech, Dresden, Germany, 2015. 1
[4] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep speech 3: 2000-speaker neural text-to-speech. arXiv preprint arXiv:1710.07654, 2017. 1
[5] J.W.A. Fackrell and Stephen McLaughlin. Detecting nonlinearities in speech sounds using the bicoherence. Proceedings of the Institute of Acoustics, 18(9):123– 130, 1996. 1
[6] Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. Digital multimedia audio forensics: past, present and future. Multimedia Tools and Applications, 77(1):1009–1040, 2018. 1
[7] AlBadawy, E.A., Lyu, S., & Farid, H. (2019). Detecting AI-Synthesized Speech Using Bispectral Analysis. CVPR Workshops.
[8] Muda, Lindasalwa & Begam, Mumtaj & Elamvazuthi, Irraivan. (2010). speech Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. J Comput. 2.
[9] Z. Wu, X. Xiao, E. S. Chng and H. Li, "Synthetic speech detection using temporal modulation feature," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 7234-7238, doi: 10.1109/ICASSP.2013.6639067.
[10] Kinnunen, Tomi & Lee, Kong Aik & Li, Haizhou. (2008). Dimension reduction of the modulation spectrogram for speaker verification. Proceedings of Speaker Odyssey.
[11] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "Phonetic speaker recognition with support vector machines," in Proc. Neural Information Processing Systems (NIPS), Dec. 2003, pp. 1377–1384.
[12] S. Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998), Sydney, Australia, November 1998, pp. 3205–3208.