Literature Review Notes

Anthony White

A Comprehensive Study on Multimedia DeepFakes

1. Boujadjine et al 2023)

# Summary

Good outline of various DeepFake detection and generation techniques. Paper focuses mainly on DeepFake images and videos instead of speech specifically. Most (all?) methods presented use a deep learning approach. Outlines many methods developed by others and provides a good idea as to the current state of DeepFake detection.

Sections in this paper:

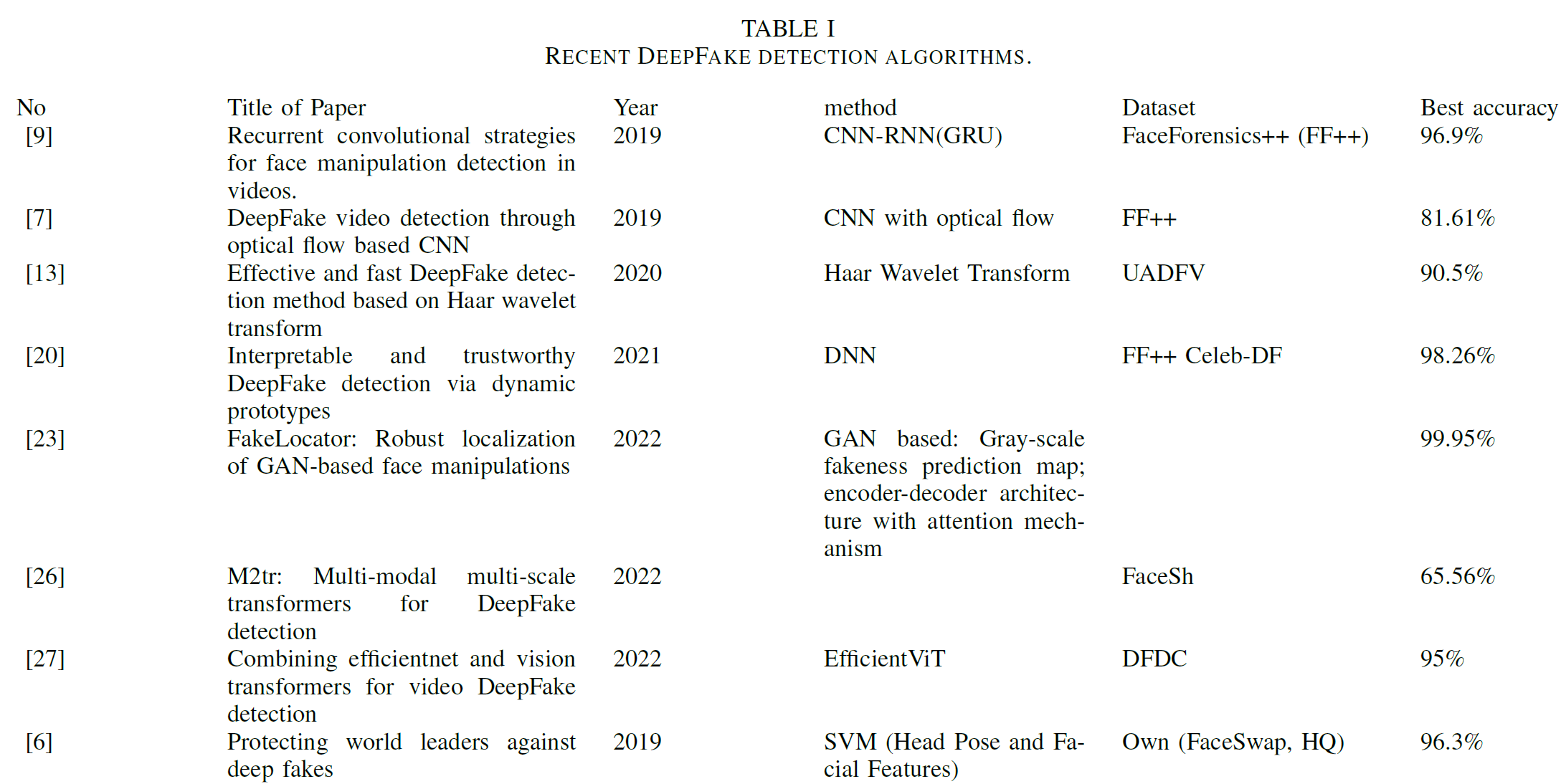
* Section 2: Commonly used techniques for DeepFake generation.
* Section 3: Methods of DeepFake detection
* Section 4: Good and bad aspects of DeepFakes
* Section 5: Threats and rising challenges of DeepFake recognition

# Snippets

Currently, generative adversarial networks (GANs), autoencoders, and neural radiation fields (NeRF) are the three most popular deep learning-based techniques for creating synthetic human faces. Traditional CGI has been around since the 1970s.

Ian Goodfellow and his colleagues [11] created the GAN, which consists of two neural networks that compete with one another: the discriminator and the generator. In this approach, as the discriminator becomes more adept at spotting AI-generated images, the generator creates ever-more realistic images in an effort to deceive them.

Numerous initiatives have been taken to enhance the effectiveness of face forgery detection, and several deep learning networks such as LSTMs, CNN [29], [30], RNN [9],



# Afchar et al. [2]:

Suggested two distinct CNN architectures with just a few layers: (a) a CNN made up of four convolutional layers followed by a fully-connected layer (Meso-4), and (b) an adaptation of Meso-4 constructed using a special Inception module called MesoInception-4.The dual CNN-introduced model demonstrated a 98% detection rate on DeepFake videos.

# Agarwal et al. [5]:

The first thing this approach does is determine the minimum distance between the original and the GAN-generated image distributions. This distance calculates the detection capability based on the findings of the hypothesis. For instance, DeepFakes are easily identified when this distance is increased. If the GAN delivers less accuracy, the distance will typically increase. Additionally, in order to produce modified images with high resolution that are more difficult to detect, a very exact GAN is required.

# Sabir et al. [9]:

suggested utilizing an end-to-end trained Recurrent Convolutional Network rather than a pre-trained model. FaceForensics++ dataset was used to evaluate their suggested detection method, and AUC results of over 96% were obtained.

Since The majority of DeepFake detection techniques are primarily focused on recognizing a specific artifact and rely on a precise approach, they are unable to generalize to manipulations in novel datasets.

Younus and Hasan [13]:

Proposed a method that compares the blurriness and sharpness of the edges in the facial region with the background boundary using the Haar Wavelet Transform.

Trinh et al [20]:

Dynamic prototype well-designed deep neural network (DPNet). adopting a more general approach than concentrating on a specific camera artifact and highlighting noise artifacts produced by the entire acquisition process, independent of their particular origin.

Ref [23]:

FakeLocator method: technique for the detection and localization of GAN-based face alterations is suggested. GANbased face alterations introduce artificial textures to the image when low-quality images are upsampled to high resolution. the application of this technology is limited to GAN-based face-altering techniques and cannot be utilized to identify fake faces or expressions.

Ref [26]:

developed the M2TR, a multi-modal and multi-scale transformer-designed feature representation based on different patch sizes to exploit both spatial and frequency domain artifacts. Multiple ViTs with various spatial embedding sizes.

Ref [27]

integrating different kinds of Vision Transformers with a convolutional EfficientNet B0 utilized for feature extraction. To determine if the video taken was edited or not, the faces from various frames of the video are individually examined, sorted, and voted on at inference time. The suggested method makes it feasible to handle scenarios like the existence of multiple people in a video where just one has been altered, for example, more effectively.

Ref [31]:

Observe the diversity of actual faces and false faces in the frequency domain.

Ref [6]:

The work aims to secure individuals by giving them soft traits that define them and are extremely difficult for a generator to imitate. Particularly, head motions and facial expressions were found to have a high correlation, and that changing the former without affecting the latter could reveal a manipulation. By capturing the peculiar behavioral traits of a specific individual, the time series of facial landmarks extracted from real films are used to identify DeepFakes.

Khochare et al. [29]:

feature-based and imagebased methods. This work made use of two brand-new Deep Leaning models, the Temporal CNN (TCN) and the Spatial Transformer (STN) via mel-spectrograms. The TCN works well with sequential data, but it cannot handle inputs with Short-Time Fourier Transform (STFT) or Mel Frequency Cepstral Coefficients (MFCC) features.

Real-Time Detection of AI-Generated Speech for DeepFake Voice Conversation

(J. Bird and A. Lotfi 2023)

# Summary

Deep-Voice dataset is generated, using eight well-known figures and their speech converted to one another using Retrieval-based Voice Conversion. Hyperparameter optimisation is implemented for

machine learning models to identify the source of speech. Following the training of 208 individual

machine learning models over 10-fold cross validation, it is found that the Extreme Gradient Boosting

model can achieve an average classification accuracy of 99.3% and can classify speech in real-time,

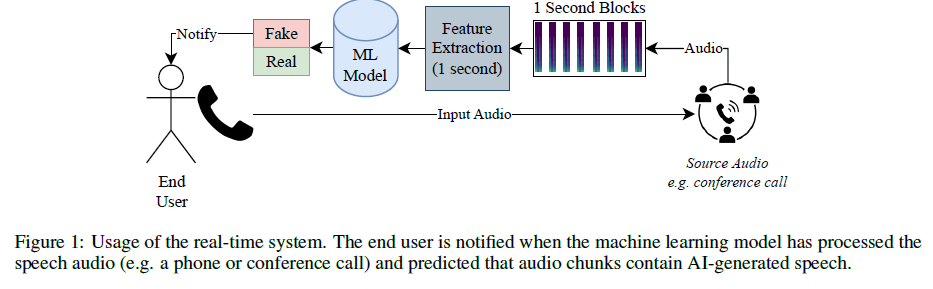
at around 0.004 milliseconds given one second of speech.

# Snippets

The scientific contributions of this work are threefold: first, the provision of an original audio classification dataset comprised of 8 well-known public figures, with real audio collected from the internet and AI-generated speech via Retrieval-based Voice Conversion (RVC). Second, the statistical analysis of extracted audio features to explore which sets of features are statistically significant given the classification of human or AI-generated speech. Third, the hyperparameter optimisation of statistical Machine Learning (ML) models towards improving accuracy and inference time, in order to achieve real-time recognition of AI-generated speech.

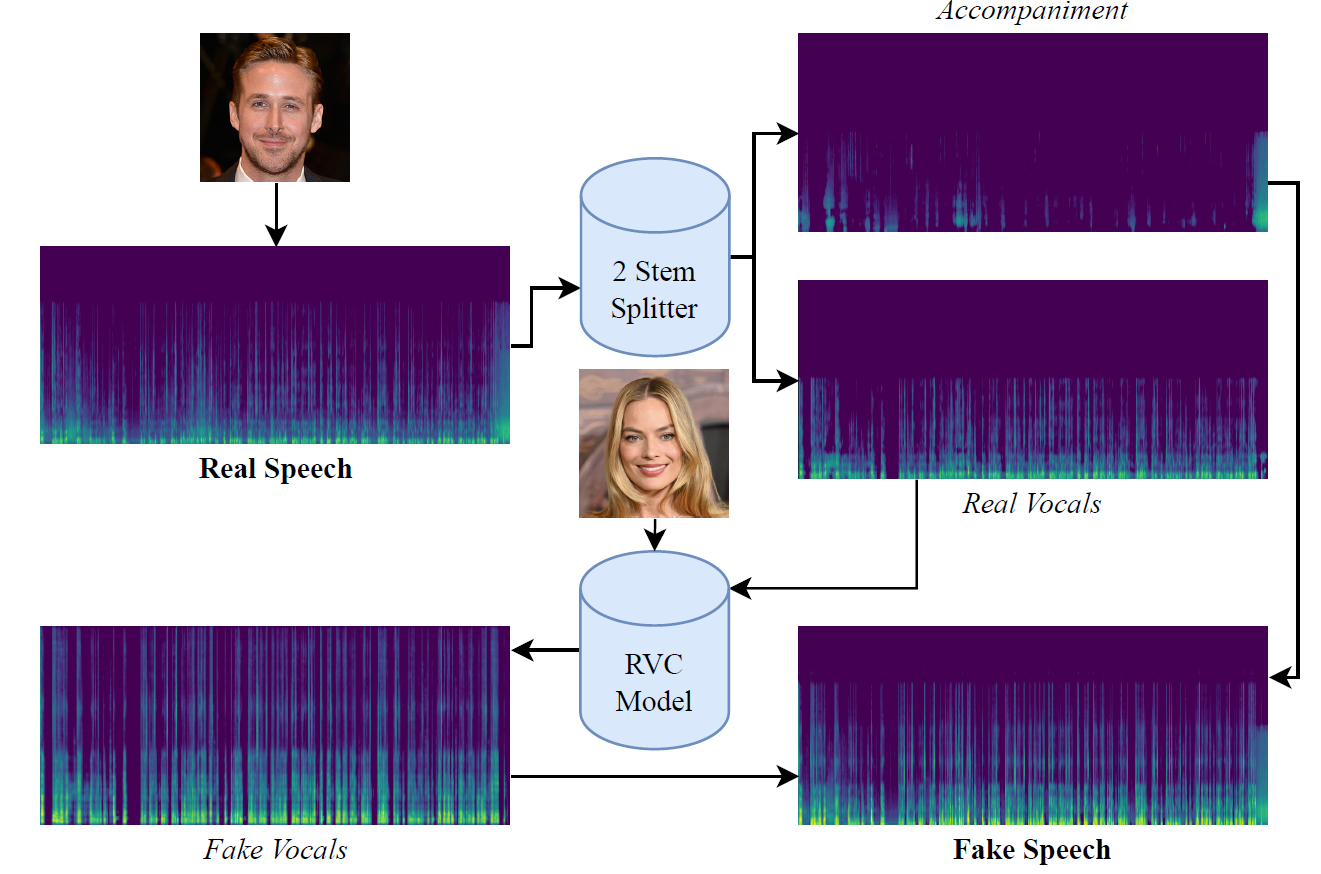
Deep Learning Fakes (DeepFakes) describe a category of algorithms which can generate synthetic media with the purpose of replacing an individual’s likeness with another [1].

In 2022, Lim, Suk-Young and Lee [16] proposed Convolutional and Temporal models for detection. This study showed that the speech generated from Tacotron sometimes showed a relatively flatter spectrogram without randomness as could be observed in human speech. Researchers conjecture that this may be due to the absence of accents in LJSpeech (the base tacotron dataset) rather than an inability of the pipeline itself. ANN and LTSM score around 97-99% accuracy when dealing with recognition.



**Method**

Real speech is split via a two-stem model [21] from Spleeter, which is an encoder-decoder Convolutional Neural Network (CNN) in a U-Net architecture. The model consists of 12 layers, with 6 layers each for the encoder and decoder networks. Following splitting of the real vocals and accompaniment tracks, the vocals are then converted using a Retrieval-based Voice Conversion (RVC) model to another individual. Finally, the original accompaniment and the RVC vocals are combined to form a fake speech track. The reason for splitting the tracks is so the style of the deepfake voice is not converted to any background noise, such as audience cheers or laughter. That is, the aim of this approach is to preserve ambient sounds while converting only the speaker’s voice.



# Data Preprocessing

Chromagram, calculated via the STFT, (Difference between chromagram and spectrogram?)

where is the frame of the signal and is the frequency bin. The spectral centroid is the location of the centre of mass in the spectrum,

Spectral bandwidth is the difference in frequencies around the centroid,

Spectral rolloff (SR) is the frequency below which 85% (or another specified percentage) of the total spectral energy lies. The zero crossing rate (ZCR) is calculated as

where is the sign function and is the number of samples in . The root mean square (RMS) of frame is

Finally, the first 20 Mel-Frequency Cepstral Coefficients (MFCCs) are calculated. The powers of the Short-Time Fourier Transform are mapped to the Mel-scale by applying a triangular window. Each of the log is taken from the power spectrum, and a Discrete Cosine Transform (DCT) is applied.

# DeepFake Voice Conversation

* For converting speech: RVC model, based on VITS architecture.
* Pitch estimation of input speech done using CREPE model, which is a deep CNN model.
* RVC is also used to quickly convert short speech samples.

# ML Models

Following the feature extraction from each 1-second block of audio, this study then implements a range of various ML models. The goal of the models is to perform binary classification of the speech, learning whether the audio is speech spoken naturally by a human being, or has been, tampered with by retrieval-based voice conversion. The models are chosen from a range of different statistical approaches for comparison.

Extreme Gradient Boosting (XGBoost) [25], Random Forests [26], Quadratic and Linear Discriminant analyses [27], Ridge Regression [28] (linear regression with L2 regularisation), Gaussian and Bernoulli Naive Bayes [29], K-Nearest Neighbours [30], Support Vector Machines [31], Stochastic Gradient Descent [32], and Gaussian Process [33].

Matthew Correlation Coefficient (MCC) is considered.

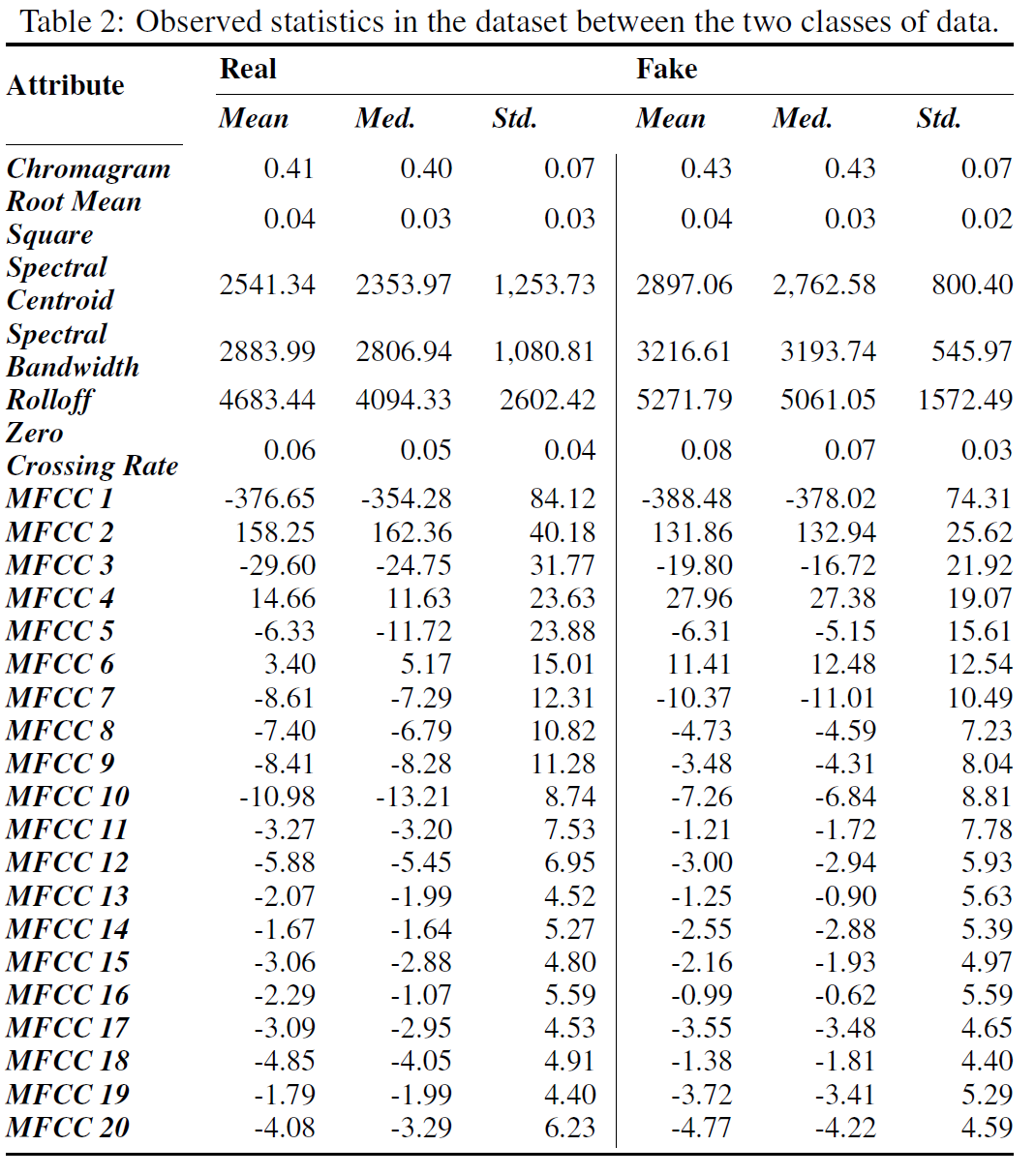
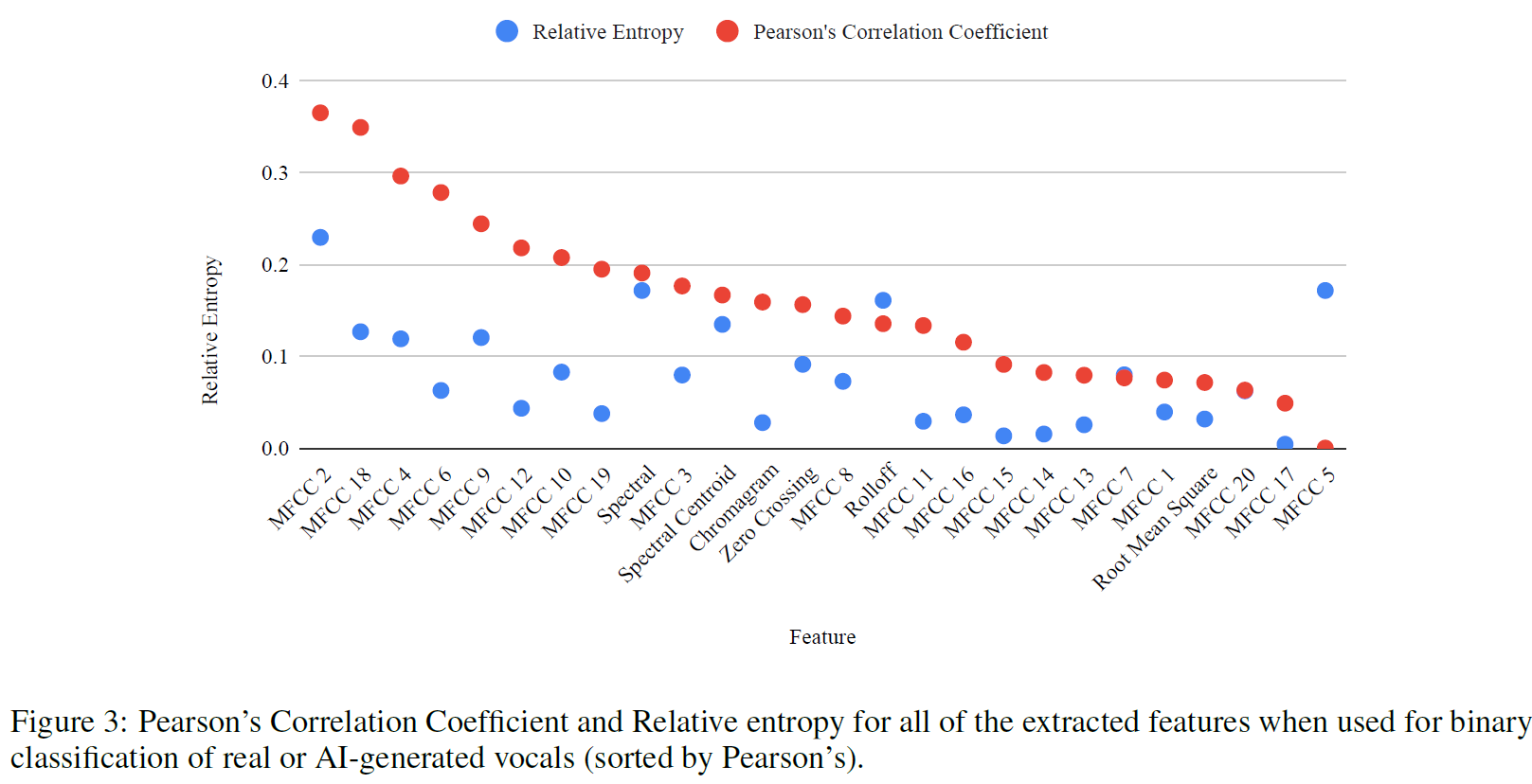


Figure 4 shows the results for the KNN models. Interestingly, the highest performing model was the smallest cluster of 1 nearest neighbour, which scored 81.48% accuracy, 0.827 recall, 0.817 F-Score, an MCC of 0.63, and an area under the ROC Curve of 0.815. However, the highest precision value was 0.886 when using two nearest neighbours as the predictors. As can be observed in Figure 5, this model took, on average, 0.143 milliseconds to classify 1-second of audio data as real or fake.

Unlike the KNN, the Random Forest results were moreso relative to one another given a number of trees in the forest. Figure 6 shows the classification results for each ensemble size. As can be observed, the model containing 310 trees scored an average 98.89% accuracy over the 10-folds of data. Similarly, the recall, precision, F-Score, MCC, and ROC area were 0.995, 0.983, 0.989, and 0.989, respectively. Figure 7 shows the inference time given an ensemble size. As could be expected, inference time increases relatively linearly given the ensemble size, which is shown in Figure 7. The

aforementioned ensemble of 310 random trees took an average of 0.057 milliseconds to classify 1-second of audio data.



A graph of different colored lines

Description automatically generated

A graph with blue lines

Description automatically generated

A graph with different colored lines

Description automatically generated

A graph with a line going up

Description automatically generated

A graph of a number of objects

Description automatically generated with medium confidence

Detecting Deepfake Voice using Explainable Deep Learning Techniques

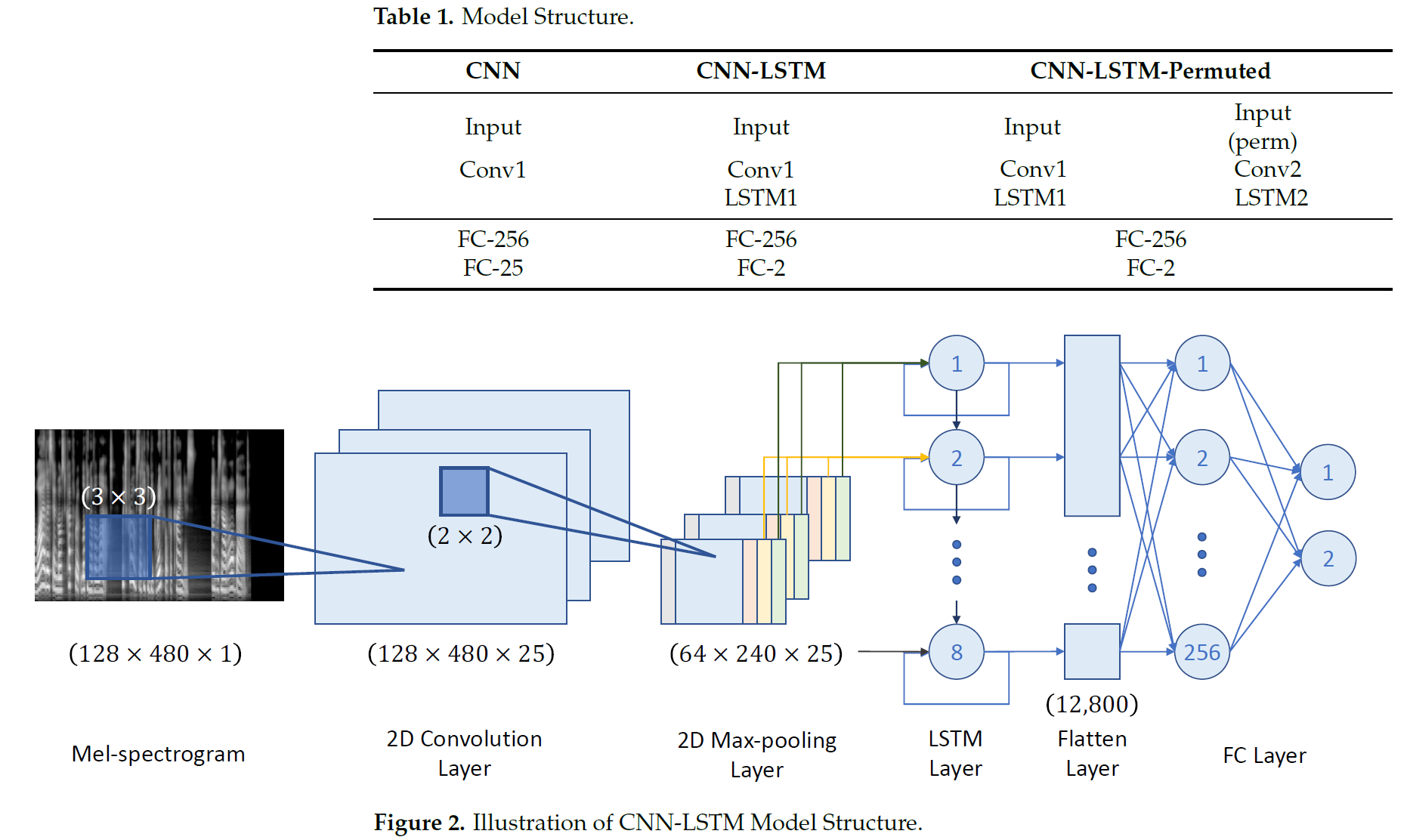
(Suk-Young Lim, Dong-Kyu Chae and Sang-Chul Lee 2022)

# Snippets

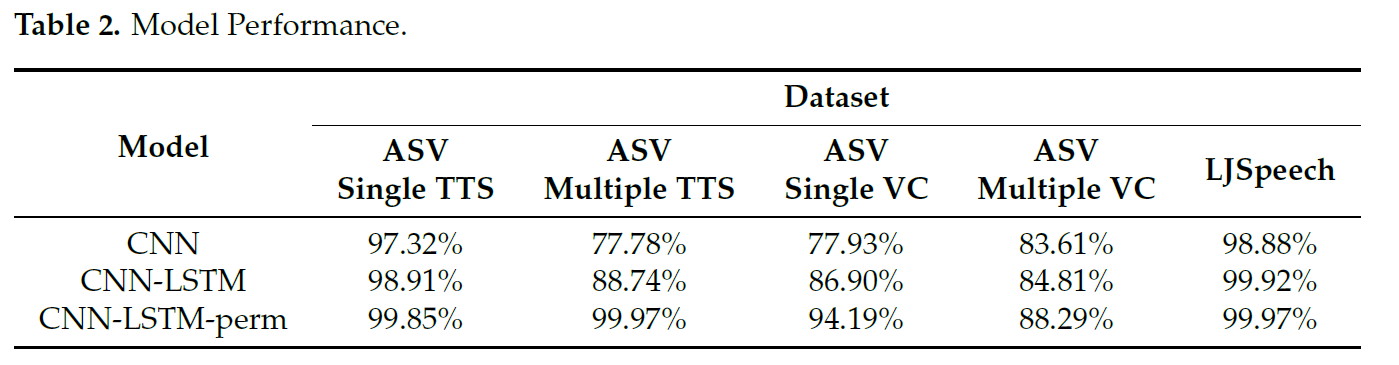
* Datasets: ASVspoof2019 Logical Access: Contains human speech + synthetic from multiple generators. LJSpeech: Contains human speech + Tacotron-generated synthetic speech.
* Three models used in experiment: CNN, CNN-LSTM, CNN-LSTM-permuted.
* XAI (Explainable AI) methods used: Deep Taylor, integrated gradients, layer-wise relevance propagation (LRP)

# Methods

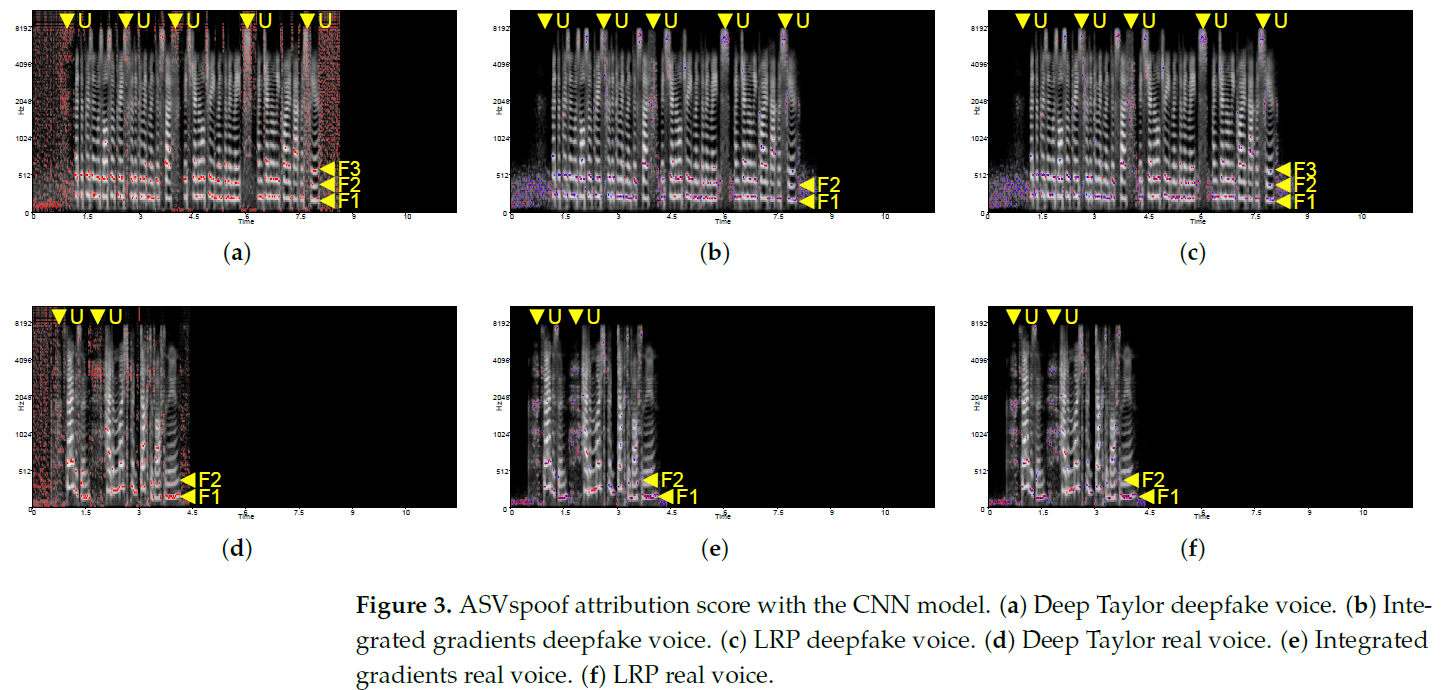
* Uses the STFT to generate spectrogram with Hann window. Mel spectrogram used to emphasise low frequency behavior. Mel filterbank
* Though the audio was converted to visualized data, i.e., a spectrogram, it still had time-series characteristics, unlike other visualized data. Consequently, a sequential model often used for spoken or natural language processing was used. A sequential model, recurrent neural network, before LSTM, had a vanishing gradient problem. Adding cell-state to a hidden state LSTM resolved this issue. For overall sequential data and tasks, LSTM shows good results.



* Each subset of data (ASVspoof and LJSpeech), data randomly split into 80% training data, 10% validation data and 10% test data. Model trained with Adam optimizer for 100 epochs and batch size of 50 speeches. Learning rate is 0.001
* XAI methods: Integrated gradients, Deep Taylor and Speech reconstruction with the Griffin\_lim algorithm.
* Griffin Lim algorithm: used to estimate the complex spectrum (or spectrogram) from the magnitude spectrum. The basic algorithm is an iteration which takes the prior estimate of the complex spectrogram, applies the STFT backward and forward, and the reapplies the desired magnitude. In each back-and-forward step, information leaks between time-frequency bins taking the estimate closer to a consistent estimate.



* Formants are provided for each dataset, using both CNN and CNN-LSTM models.



* Currently, interpretability is focused on a visual method, i.e., a heatmap on a spectrogram, to interpret the model. With the visual approach, diverse characteristics of the model’s tendency, depending on XAI methods and data types, have been found.
* Using the Griffin–Lim algorithm, the attribution scores of each model and XAI method were recomposed back to audio. For the CNN-only model, because of the tendency of XAI methods to show dependencies on energy-like formants, most of the original voice was recovered during the conversion from the attribution score. The transcript of the original voice can be recognized from the recomposed voice as well.
* In the case of the recomposed voice of Deep Taylor and LRP on the CNN-LSTM model, because of the lack of sensitivity on formants 1, 2, and 3, the transcript of the speech was unrecognizable.
* Based on the audio interpretation, i.e., the spectrograms in Figure 8, energy distribution on frequency 2 kHz showed distensibility. The deepfake voice showed a relatively flat pattern compared to a human voice, while the human voice showed randomness. This phenomenon was assumed to be triggered by the absence of training on accents while generating training data from the LJSpeech through Tacotron TTS. This absence resulted in a flat rhythm and accent.

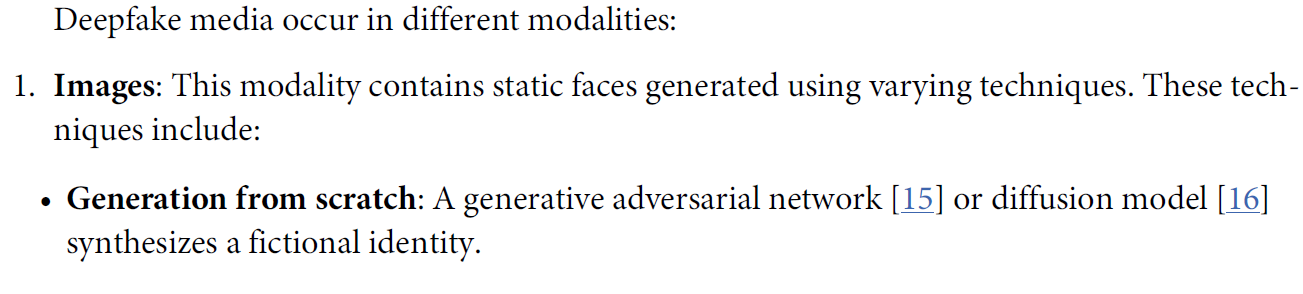
Warning: Humans Cannot Reliably Detect Speech Deepfakes

(Kimberly T. Mai, Sergi Bray, Toby Davies, Lewis D. Griffin 2023)

Adversaries are already using speech deepfakes to commit fraud. In 2020, a bank manager in Hong Kong received a phone call from someone sounding like a company director he had spoken to before [1]. The purported director requested the bank manager to authorize transfers totaling $35 million. Based on their existing relationship, the bank manager transferred $400,000 until he realized something was wrong. The bank manager was a victim of an elaborate hoax: fraudsters had used deepfake technology to clone the director’s voice. This incident is not isolated. In 2019, the CEO of a UK-based firm was swindled by a speech deepfake of his manager into transferring Re220,000 to a Hungarian supplier [2].

Moreover, it is estimated as much as 90% of online content will be synthetically generated by 2026 [14].

We randomly assigned the participants to two configurations. In the first configuration, we presented participants with one audio clip at a time and asked them to decide if the clip was fake. In the second configuration, we presented participants with audio clip pairs containing the same speech (one bona fide and one synthesized) and asked them to identify the synthesized audio. English and Mandarin.



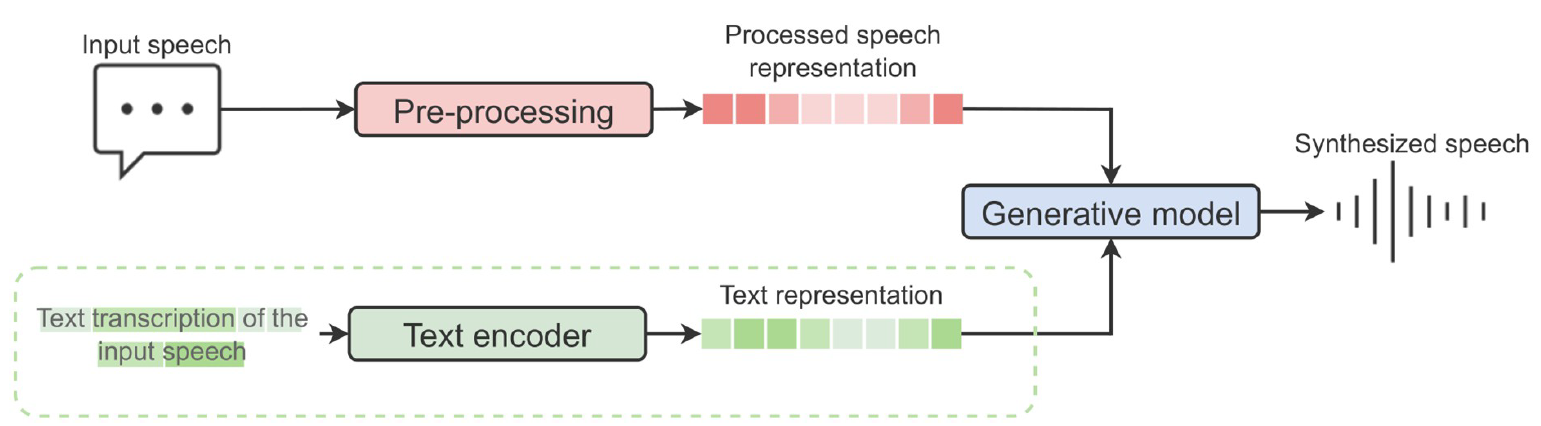
A screenshot of a computer

Description automatically generated

Targeted: Deepfake resembles a real person. Untargeted: deepfakes generated from scratch.

To generate synthetic speech:

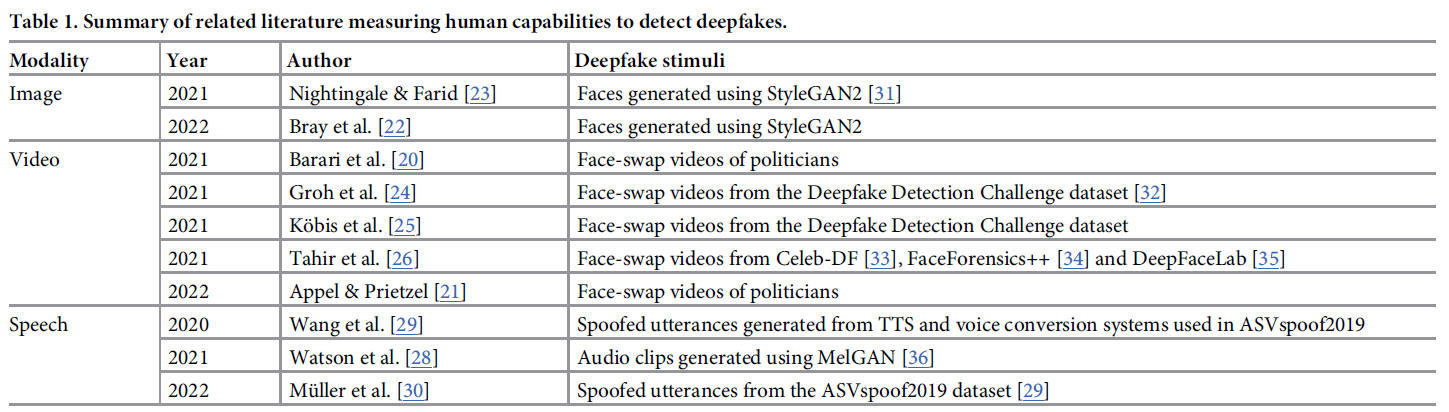
* Data collection: record audio from the speaker
* Pre-processing: Convert to alternative formats for the generative model
* Training: audio recordings are fed to the model to learn patterns/characteristics of the data. Trained model is sometimes called a vocoder.

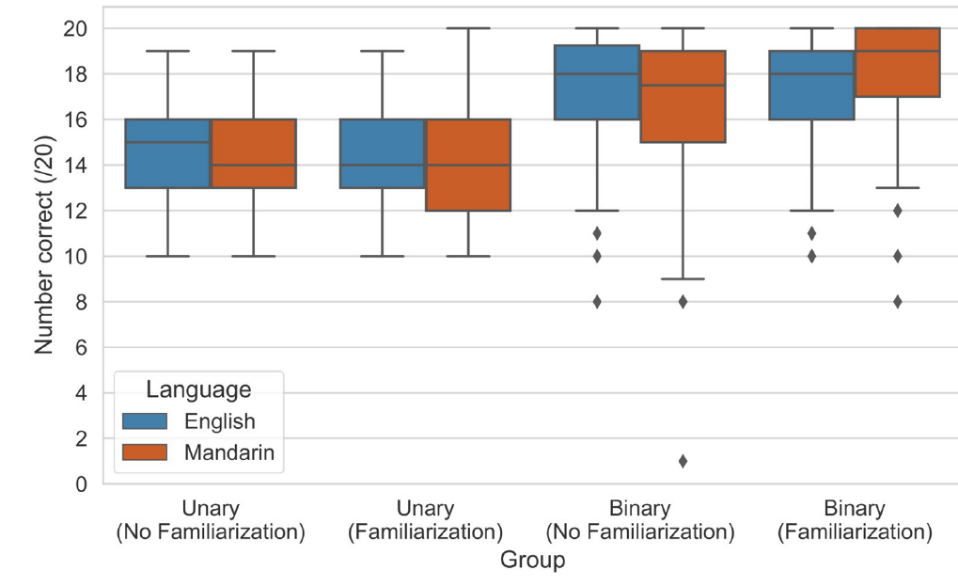


Most deepfake detection studies which examine human performance use visual media. When

faced with deepfake content of politicians, participants rely on contextual knowledge in the

form of political literacy to identify spoofs [20, 21]. Removing such background knowledge makes the detection task more difficult. In the context of images, multiple studies show humans do not perform much better than chance [22, 23]. There is no improvement when evaluating videos either [24–26]. Moreover, these studies suggest humans are overconfident in their deepfake detection abilities [25].





**Reference audio helps with deepfake detection.** The linear regression results indicate the improvement gained from the binary scenario is statistically significant (p < 0.001). Consequently, the results suggest contextual information via reference audio is beneficial for uncovering quirks in synthesized speech.

**Training humans to detect deepfakes only helps slightly.** The familiarization treatment

increases detection accuracy by 3.84% on average (p = 0.001). This effect is also present in the

unary and binary regression results, improving accuracy by 3.76% (p = 0.017) and 3.85%

(p = 0.032), respectively. However, incorporating familiarizations equates only to an accuracy

slightly above chance (52.31%) in the unary setting for the mean clip length (5.76 seconds),

ceteris paribus.

**It is equally challenging to detect deepfakes in Mandarin and English.** Fig 3 shows that

performance in English and Mandarin is comparable across the different treatment groups.

This observation is supported by Table 5, which shows Mandarin-speaking participants only

outperform their English counterparts by 1.79%, and this effect is not statistically significant

(p = 0.202).

**Shorter speech deepfakes are not easier to identify.** As our stimuli varied from 2 to 11

seconds, we included clip length in the regression to verify whether it is easier to discriminate

shorter clips. Our results suggest clip length has a negligible impact on accuracy, improving

performance by only 0.80% for each additional second. Our scatter plot (Fig 4) supports this

and shows no relationship between the two variables. These findings conflict with Watson

et al. [28], who suggest it is easier to identify shorter deepfakes.

**Listening to the clips more frequently does not aid detection.**

**Spending more time on the task does not affect performance.**

**Participants do not get better throughout the task without explicit feedback.**

**Human performance is less sensitive to unknown conditions compared to automated detectors.**

**Crowd speech deepfake detection is comparable to the top-performing automated detectors.**

A graph with different colored lines

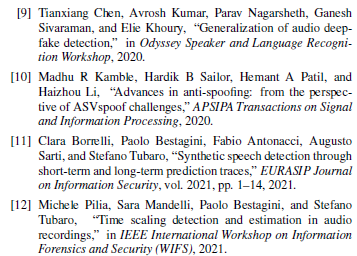
Description automatically generated

Deepfake Speech Detection through Emotion Recognition: A Semantic Approach

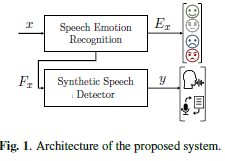
(E. Conti et al. 2022)

Proposes new audio-spoofing detection system leveraging emotional features. Audio deepfake techniques can synthesize low level voice characteristics but cannot correctly synthesize natural emotional behavior.

In the audio field, [9] feeds linear filter banks into a Resnet to generate embeddings used as input of a neural network classifier, and in [10] long-term features are used to discriminate fake and real audio tracks. Recently [11] detected for audio deepfakes based on long-term and short-term predictor features, while [12] exploits the traces left by time scaling to discriminate fake audio signals.



Below diagram is the architecture of the proposed system. The first block is a SER system that exploits the architecture recently proposed in [15]. Starting from an input speech signal x, it estimates the expressed emotion Ex and extracts a set of features Fx. The second block is the Synthetic Speech Detector (SSD) system that associates a class y to the input features Fx.



# Speech Emotion Recognition

Extracts a set of features from the audio signal . Emotional features are computed using a 3D-Convolutional Recurrent Neural Network (CRNN) proposed in [15]. There are possible emotion classes. Output of the network is where is the -th emotion class. Audio signal must be preprocessed by taking the STFT in the mel-frequency domain and applying a logarithmic transformation to the STFT magnitude. This returns a log-mel spectrogram. , where is the number of windows and is the number of mel bins.

Taking first and second derivatives along frequency domain obtains and , from which we define . This matrix is then standardised by means of z-score normalization. The processed input is fed to a set of 3D convolutional layers, followed by a linear layer, a BLSTM and an attention layer. Finally, a sequence of dense layers outputs a probability measure of each emotion class, from which we extract the prediction . Adopting a transfer-learning strategy, we extract a feature vector of dimensionality from an intermediate network layer. is used as an input to a classifier trained for deepfake detection. Random Forest classifier is used.

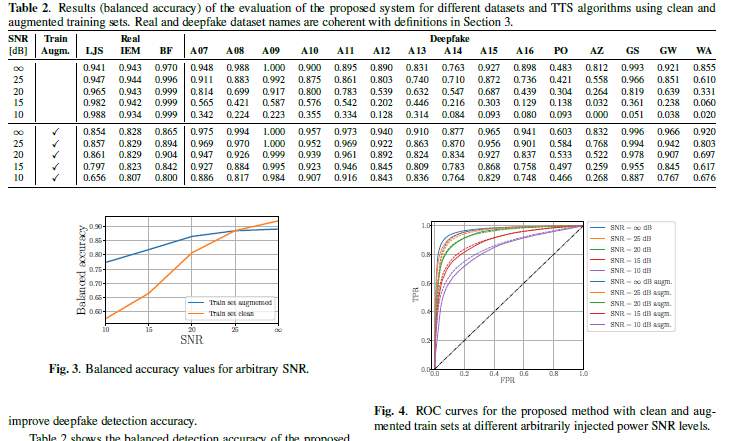
Datasets used in this study: ASVspoof, LibriSpeech, LJSpeech, Cloud2019, IEMOCAP.

# Experiment

Data is preprocessed, which involves converting audio signals to mono and, if necessary, downsampling to 16 kHz. Signals are filtered with a 6th order digital Butterworth bandpass filter and normalised via infinity norm. SER block input is calculated via time-frequency transform and STFT (Hamming window).

A variety of white noise is added to the training data, test data has no noise.

VGGish and RawNet2 used as a baseline. VGGish is a popular CNN architecture .RawNet2 is an end-to-end network aimed at audio-spoofing detection.



Top half of table 2 = trained with clean data. Bottom half = trained with noisy data. Performance is good for all pristine samples and most deepfake generation algorithms. Algorithm A14 from ASVspoof2019 and PO from Cloud2019 are the only cases where the accuracy is below 0:8. We suspect that this is because algorithm A14 is a mixed TTS/VC system that has been built starting from a very efficient VC system [19]. Hence real emotional qualities are probably still present in the audio tracks, affecting the efficiency of the proposed system. For all the other deepfake systems, the balanced accuracy value is close to or greater than 0:9. We can observe in rows 2 to 4 that, as the noise level increases, the performances of the synthetic speech detector degrades more and more. As the noise increases, the classifier tends to label all samples as authentic, as we notice a remarkable increase in the false-negative rate. This behavior encourages the use of data augmentation strategy on the training set.

Time Scaling Detection and Estimation in Audio Recordings

(Michele Pilia, Sara Mandelli, Paolo Bestagini, Stefano Tubaro 2021)

Detect and estimate time scaling applied to an audio signal. Time scaling used to speed up or slow down audio recordings, enabling the creation of natural sounding fake speech compostions.

Levarages the CNN, which analyses the log-mel spectrogram and the phase of the STFT of the audio input signal.

# Preprocessing

is the log-mel spectrogram, is the mel filterbank, is the spectrogram, is a small constant to avoid feeding zero to the logarithm.

# Time scaling

Control the temporal duration of a signal independently from its frequency behaviour. Speed up or slow down the signal without changing its spectral content. Employ a quasi-stationary model introduced by [19], [20],

To implement time scaling,

where is the temporal mapping function and is the so-called time scaling factor. is the ratio of the duration of the original signal to the duration of the time scaled one. For , the signal is slowed down and for the signal is sped up. The time-scaled signal is

Note the frequency and the amplitude of the -th sinusoid of the time scaled signal is equal to the amplitude if the same sinusoid of the original signal at . Time evolution has changed but frequency content has not. This cannot be achieved through resampling, which scales the signal in time but also affects frequency behavior, introducing undesired artifacts and traces which can be easily spotted. Different time scaling algorithms have been proposed, from frequency domain methods (e.g., phase vocoder [21] and sinusoidal spectral modeling [22]) to time-domain methods (e.g., those based on overlap and add [23] and synthesis frames [24]).

# Methodology

* Preprocess the audio data and feed it to a CNN
* Daa processing through the CNN, modifying the final network layer to detect/estimate time scaling

**Preprocessing:** Audio track converted to 2D data derived from the STFT.

* LMS = Log-mel spectrogram (STFT magnitude)
* High pass filtered version of the unwrapped STFT phase. (2nd order Butterworth high pass filter). HPP = High-pass unwrapped phase.
* Residual median-filtered version of the STFT unwrapped phase. RMP=Residual Mean unwrapped phase.
* LHR = Concatenation of LMS, HPP and RMP data

Filtering is done to filter out low frequency information and enhance high-frequency variations. Aim to investigate the possibility that time scaling leaves peculiar but inaudible high-frequency phase variations.

**CNN Data Processing:** EfficientNet-B0 is the backbone for network architecture. From the 2D input, 1280 scalar features are extracted. Linear transformation is applied so that the output of the CNN counts elements. The parameter differentiates the CNN used for time-scaling detection from the CNN used for time-scaling estimation.

**Experiments and Results:**

Datasets: GTZAN genre recognition dataset [12]. Composed of 1000 mono tracks 30 seconds long, equally divided into 10 music genres. IRMAS testing dataset, designed for instrument recognition in musical audio signals [13]. Composed of 3874 excerpts with length between 5 and 20 seconds.

Time scaling implementations:

* Audiotsm library [28]
* TimeStretch algorithm by Torchaudio [29]
* Implementation provided in [30]

Time scaling factor is 0.1 to 1.9.

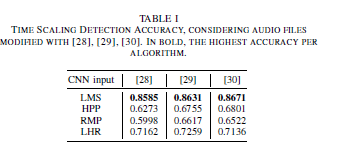
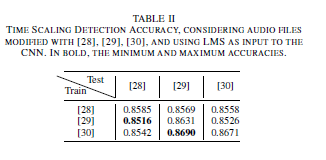
After editing the audio files, the LMS, HPP and RMP is calculated based on the STFT. A Hanning window is used with length 25ms and hop size 10ms. Investigated frequency range is 125Hz to 7.5kHz, sampling rate of 16kHz. STFT computed with 512 frequency samples and variable temporal dimensions depending on the audio length. 64 Mel bins to compute the LMS and HPP/RMP cropped to same size. Then, for each audio track, we select 4 non-overlapped time windows to increase the overall dataset dimensions. The selected windows start from the 4-th to the 7-th temporal bin of the STFT and cound 96 temporal bins each.

Optimisation model is the Adam optimizer with standard parameters. Learning rate is 0.001, reduced by a factor 10 whenever the validation loss does not improve for 10 epochs. We stop the training process if the validation loss does not decrease for 20 consecutive epochs; in any case, we consider a maximum number of epochs equal to 80. After the training, the model providing the best validation loss is selected.

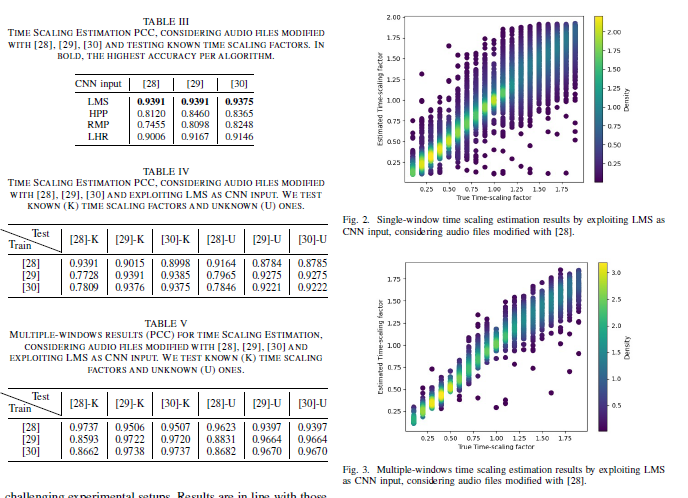
Evaluation metrics:

For time scaling detection: average accuracy of correct predictions

For time scaling estimation: use Pearson Correlation Coefficient (PCC)



LMS has the best accuracy.



Generalisation of Audio Deepfake Detection

(T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury 2020)

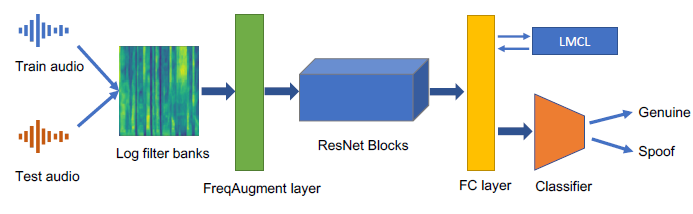
Detecting deepfakes using a large margin cosine loss function (LMCL) and online frequency masking augmentation to force the neural network to learn more robust feature embeddings. Baseline neural network has lowest error rate of 4.04% - improvements in this paper reduce it by a further 1.26%.

In 2015 [1], ASVspoof challenge focused on detecting voice conversion attacks based on hidden Markov models (HMM), Gaussian mixture models (GMM) and unit selection. Since then systems have drastically improved with deep learning.

WaveNet [2], proposed in 2016, was the first end-to-end speech synthesizer that directly uses raw audio for training and showed a mean opinion score (MOS) vey close to human speech. Other speech systems are Deep Voice [3], Tacotron [4], VC systems [5, 6].

ASVspoof challenge in 2019 showed that the current biggest problem in a spoofing detection system is its generalisation ability. Traditionally, signal processing tried to overcome this problem by engineering different low-level spectro-temporal features. Constant-Q cepstral coefficeints (CQCC) were proposed in [8], cosine normalised phase and modified group delay (MGD) was studied in [9, 10]. None of these methods were able to generalize to unknown spoofing technologies.

Large margin cosine loss (LMCL) function is initially used for face recognition. Goal of LMCL is to maximise variance between genuine and spoofed class and, at the same time, minimise intra-class variance. Additionally, inspired by SpecAugment [17], we propose to add FreqAugment, a layer that randomly masks adjacent frequency channels during the DNN training, to further increase the generalization ability of the DNN model. On the ASVspoof 2019 EVAL dataset, we achieve an EER of 1.81% which is significantly better than the baseline. The proposed system is illustrated in Figure 1.



# Low level Features

Linear filter banks (LFBs) are a direct compressed version of the STFT and more adequate for lower computational cost. Lower risk of network overfitting at training time. Used 60-dimensional LFBs extracted on 30ms windows with 10ms frame shift. Mean and variance normalization was performed at the utterance level.

# Frequency Masking

Applied during training to randomly drop out a consecutive frequency band range of . is chosen from a random frequency distribution , where defines the maximum number of frequency channels to be masked. Similarly is chosen from , where is the total number of frequency channels of the input LFB.

# Large Margin Cosine Loss

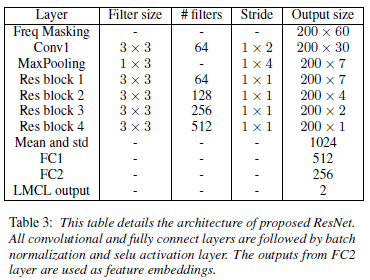
It aims to force the deep neural network to learn the feature embedding that can maximize the inter-class variance and minimize the intra-class variance, by reforming the soft-max loss as a cosine loss and injecting a margin in the cosine space. LMCL can be defined as:

subject to

where is the number of training samples, and denote the normalized -th feature and weight vector corresponding to -th class. denotes the weight vector of -th class. and are the hyper parameters to define the margin in cosine space. In this work, we set s = 10 and m = 0.35.

# Deep Residual Network

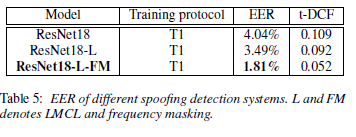
Baseline is the deep residual network (ResNet) based system. Mitigates the gradient vanishing problem in deep neural networks by stacking residual blocks. Short-cut connections added between convolutional layers, enabling gradients to flow any other earlier layer. Scaled exponential linear unit (SELU) activation [27]. Proposed system is an improvement over the baseline – applied frequency masking augmentation before the input layer, removed length normalization layer, replaced softmax loss with LMCL.

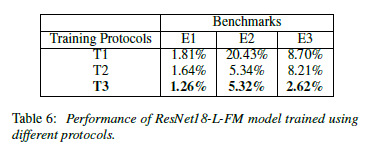


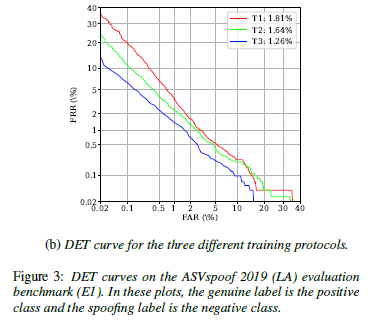
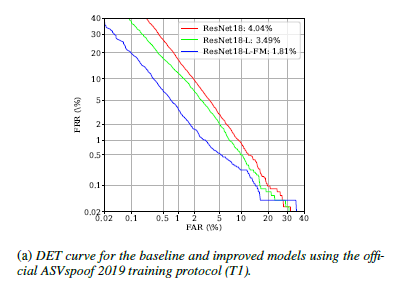
Three different training recipes (T1-T3) and three different evaluation benchmarks (E1-E3). Two key performance metrics – EER represents where false rejection rate (FRR) meets false acceptance rate (FAR). Second metric is minimum normalised tandem detection cost function (t-DCF)

where depends on application parameters, and are countermeasure system miss and false alarm rate at threshold s.

# Results







In this paper, we propose a robust end-to-end deep learning framework for voice spoofing detection, that can detect spoofed audio generated from a wide variety of unknown TTS and VC systems with high accuracy. We successfully demonstrate that we can increase the generalization ability by adding FreqAugment layer and large-margin cosine loss and applying data augmentation. The experimental results show an EER of 1.26% on ASVspoof 2019 evaluation set, which is a remarkable improvement over the state-of-the-art.

Synthetic speech detection through short-term and long-term prediction traces

(Clara Borrelli , Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Stefano Tubaro 2021)

Indeed, synthetic speech can be obtained by simple cut-and-paste techniques performing waveform concatenation [10], in some cases available as open source toolkit. Alternatively, it can be obtained by vocoders exploiting the source-filter model of speech signal [11]. More recently, even multiple convolutional neural networks (CNNs)-based methods for synthetic audio generation have been proposed [12].

In the closed-set scenario, the proposed method detects whether the speech is bona fide or synthetic. In the case of synthetic speech, it also detect which algorithm has been used to generate the speech. In the open-set scenario, the proposed method is also able to highlight whether a fake speech has been generated through an algorithm that has never been seen before.

Methods based on modelling speech as an autoregressive process.

Uses ASVspoof 2019 dataset. This dataset contains synthetic speech tracks generated through 17 different speech synthesis techniques, ranging fromthe older (e.g., waveform concatenation, vocoders, etc) to novel ones based on CNNs approaches.

# Background - Previous Speech Generation

Previous TTS synthesis was based on concatenative waveform synthesis [18-21]. The main drawback of concatenative synthesis is the difficulty of modifying the voice timbral characteristics, e.g., to change speaker or embed emotional content in the voice.

To improve variety of voice qualities/speaking styles HMM-based speech synthesis have ben proposed. These operate with contextual hidden Markov models (HMMs) trained on large datasets of acoustic features extracted from diphones and triphones [22–24].

Parametric TTS synthesis algorithms aims at expanding the variety of generated voices. Take inspiration from the concept of a vocoder, first proposed in 1939 [25]. In this case, starting from a set of speech parameters (e.g., fundamental frequency, spectral envelope and excitation signal), a speech signal is generated, typically as an auto-regressive process. However, parametric TTS synthesis produce results that sound less natural than concatenative one. Nonetheless, in the last years, more sophisticated and high-quality vocoders have been proposed [11, 26, 27]. (H. Dudley, Remaking speech. J. Acoust. Soc. Am. 11, 169–177 (1939)) can’t find paper ;(

In the last few years, CNN and recurrent neural networks (RNN) have enabled to build completely auto-regressive models, hence to synthesize directly raw audio waveforms [12, 28, 29]. These end-to-end speech synthesis architectures stand out with respect to classic methods in terms of timbre, prosody, and general naturalness.

# Fake Speech Detection

Traditional approaches focus on extracting meaningful features from the speech signal and using them to discriminate between fake and real audio tracks. Long-term features preferred to short-term. Examples are constant-Q cepstral coefficients (CQCC) [31], based on a perceptually inspired time-frequency analysis, magnitude-based features like log magnitude spectrum or phase-based features like group delay [32].

It has been noted that traces of synthetic speech algorithms are distributed unevenly across frequency bands. Sub-band analysis was exploited for synthetic speech detection, presenting features such as linear-frequency cepstral coefficients (LFCC), mel-frequency cepstral coefficients (MFCC) [13]. In [15] feature extraction is based on a linear prediction analysis of the signals. Features typically fed to simple supervised classifiers, often based on Gaussian mixture models.

Neural networks:

[8, 14]: CNN is used for feature learning, RNN is able to capture long-term dependencies and is used as a classifier. Several inputs have been tested, from classic spectrograms to complex novel features such as perceptual minimum variance distortionless response (PMVDR).

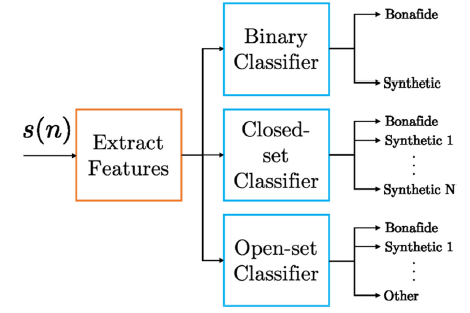
The baseline is the method proposed in [9]. Given a speech signal , split it into windows . Define as the Fourier transform of . The bicoherence is



where \* is the complex conjugate. Finally, the authors extract the first four moments of the bicoherence magnitude and phase and concatenate them in a feature vector which is fed to a simple supervised classifier to distinguish whether a speech is synthetic or bona fide.

# Speech Detection Method

Problem is faced at three granularity levels; binary, open-set and closed-set.



We propose a set of audio descriptors based on short-term and long-term analysis of the signal temporal evolution. Indeed, speech signals can be well modeled as processes with memory. It is therefore possible to extract salient information by studying the relationship between past and current audio samples. Multiple linear prediction orders at once.

# Data Model

Represent speech using source filter model. e(n) is the source excitation signal.

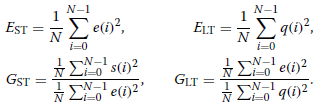


Perform a linear prediction analysis, minimizing the total squared error. Autocorrelation method using the Levinson-Durbin recursive algorithm used to obtain the short-term prediction error

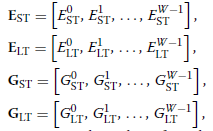
Long term prediction involves estimating parameters , the fundamental pitch period and , the gain factor. Parameter is obtained by minimizing cost function

where is approximated as [35]. Long term prediction error is

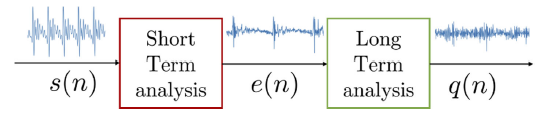
Features employed in proposed method are prediction error energy and prediction gain . ST=short-term analysis and LT=long-term analysis.



These quantities defined for each window makes the vectors



where is the total number of windows. Boxcar window of length 0.025ms.



Vector containing mean, standard deviation, min/max value across the windows,



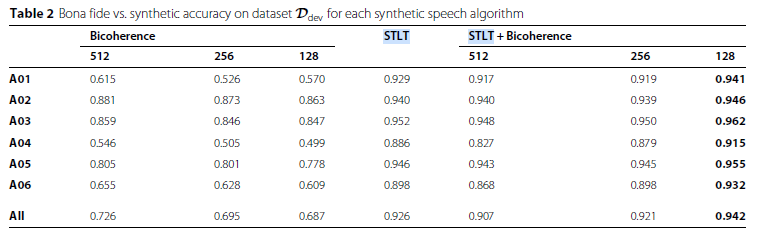
Repeat with different short-time prediction orders,

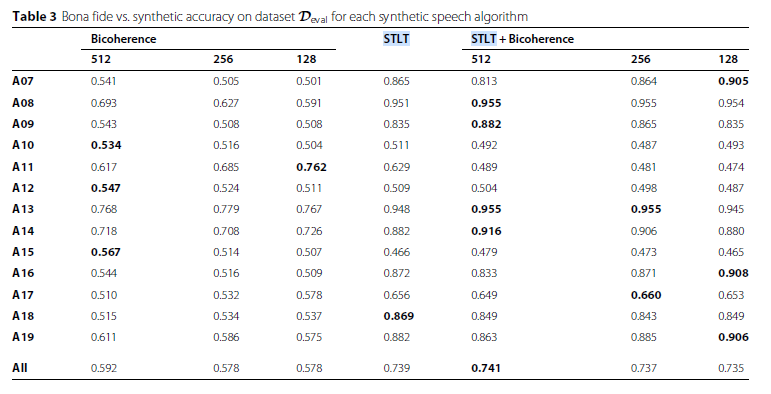


# Experiment

Proposed features can be used with any supervised classifier. In this study, simple/classical classifiers are used: random forest, a linear SVM, a radial basis function (RBF) SVM. Different normalization techniques have been used: min-max normalization, z-score normalization.

# Results





Proposed features are STLT. Bicoherences alone perform well, but are outperformed by STLT. Best results are the STLT+bicoherence case, which achieved average accurace of 0.94 for the development dataset. Evaluation dataset is not so good.

Worst Performance – STLT accuracy < 70%

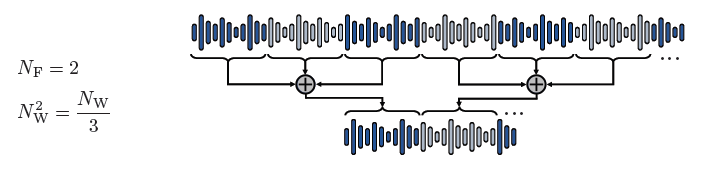
* A10 end-to-end NN based TTS system
* A11 same as A10 but uses Griffin Lin algorithm
* A12 neural TTS system based on wavenet
* A15 combined VS/TTS system
* A17 NN based VC system

Can’t really see a pattern

Synthetic Speech Detection Through Audio Folding

(D. Salvi, P. Bestagini, S. Tubaro 2023)

Speech sampleis ‘folded’ on itself in an attempt to reduce the length of the signal, thereby reducing the amount of required computation for synthetic speech detection. In the below diagram, is the number of folds, is the number of time windows in the original signal and is the number of time windows in the folded speech signal.



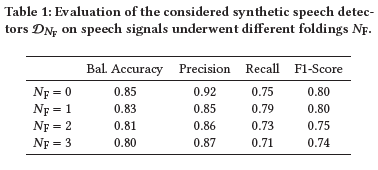
The total computation time for the folded signal is reduced by a factor of . The proposed method works with both speech tracks belonging to a single class (i.e., entirely real or entirely fake) and partially fake tracks.

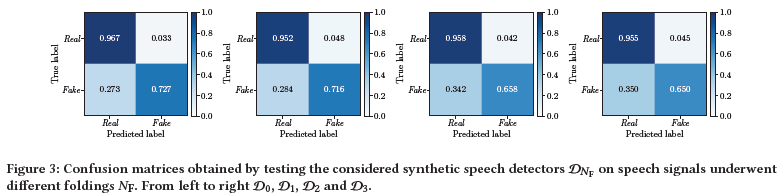
Synthetic speech detector is RawNet2, an end-to-end neural network that operates on raw audio data used to perform a binary classification into synthetic/natural speech. Comprises Sinc filters from SincNet followed by two Residual blocks with skip connections on top of a Gated Recurrent Unit (GRU) layer.

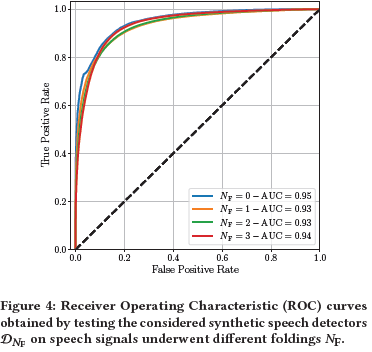
It has been shown that a raw data-based model works better than feature-based models [13]. One of the advantages of using raw data-based models is their high-resolution capacity for making predictions. Compared to feature-based models, which rely on extracting relevant features from the audio signal, raw data-based models process the audio signals directly, capturing more intricate details and complexity that feature-based models may overlook. This greater level of detail can improve the accuracy and robustness of the detection system.

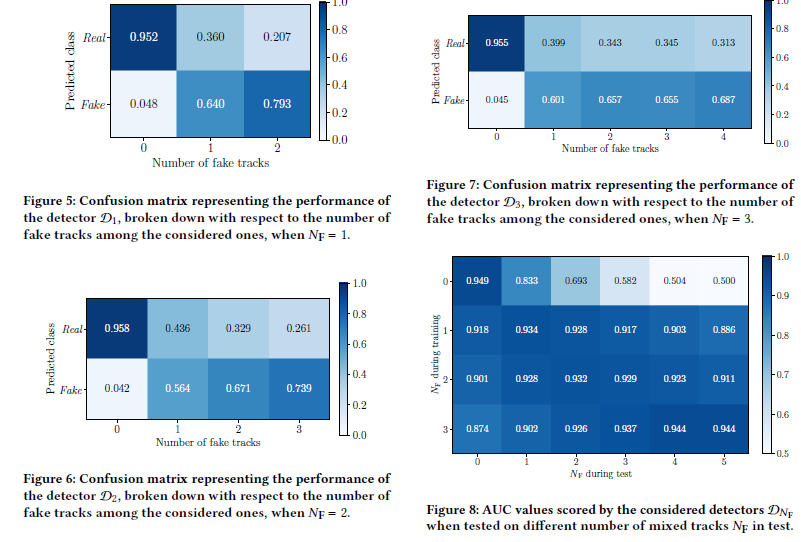
# Results

Evaluated using receiver operating characteristic (ROC) curves, area under curce (AUC), precision, recall and balanced accuracy. Optimal performance is when all values are equal to one.









The Sound of Silence: Efficiency of First Digit Features in Synthetic Audio Detection

(Daniele Mari, Federica Latora, Simone Milani 2022)

Investigating irregularities in silenced parts of the speech sample. Shows that Fermi-Dirac (FD) statistics are effective in discriminating fake audio samples since they allow catching irregularities in silenced parts between the different words of the speech. Tests performed on full audio sequence, just silent parts and just voiced parts.

In this paper:

* We analyzed the role of silent parts in detection showing that most of the classification accuracy derives from the difficulty in synthesizing statistically-realistic silence intervals.
* We evaluated the efficiency of MFCC FD statistics in detecting audio fake samples generated by a set of different heterogeneous algorithms. Such features have proved to be extremely useful in highlighting the statistics of silenced parts.
* We designed a lightweight classifier whose efficiency can cope with more complex detectors.

# First Digit Features for Synthetic Audio Tracks

It is possible to verify that any synthetic signal generated by a set of FIR filters with limited support fits Benford’s law with a different accuracy with respect to a natural signal. The digits computed on the quantised MFCCs where is the frame index, is the quantisation step size and is frequency are computed as . The probability mass function (PMF) is



where denotes an indicator function for digit . Several studies show that this can be approximated by the generalised Benford law,



The Reny, Tsallis and Shannon divergences are calculated, as well as the mean square error.

This paper is complicated ;(

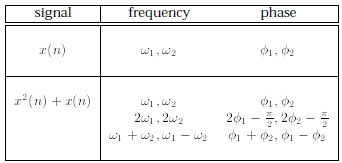
Detecting Digital Forgeries Using Bispectral Analysis

(H. Farid 1999)

Interested in verifying the authenticity of signals without any prior knowledge of the original. Assumption is that natural signals contain weak higher order statistical correlations and forged signals contain strong higher order statistical correlations. This results from the signal originating from a nonlinear processes which can introduce quadratic phase coupling (QPC).

From paper ‘Detecting Nonlinearities in Speech Sounds using the Bicoherence’, speech originating from linear models is intelligible (can be understood) but does not sound very natural.

Input/Output frequency content of a simple nonlinear system :



The power spectrum of a signal with DFFT is . This can be used to detect second order correlations, but is blind to higher order correlations – need to use higher order spectral analysis such as the bispectrum. Normalised bispectrum is the bicoherence which is between 0-1.

Fourier series of a signal can be written as

Similar form to just considering two frequency components.