Voice spoofing detection using a neural networks assembly considering spectrograms and Mel Frequency Cepstral Coefficients

Carlos Alberto Hernández-Nava Corresp., [1], **Eric Alfredo Rincón-García** [2], **Pedro Lara-Velázquez** [2], **Sergio Gerardo de-los-Cobos-Silva** [2], **Miguel Angel Gutiérrez-Andrade** [2], **Roman Anselmo Mora-Gutiérrez** [3]

[1] Posgrado en Ciencias y Tecnologías de la Información, Universidad Autónoma Metropolitana, Ciudad de México, Ciudad de México, Mexico
[2] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Ciudad de México, Ciudad de México, Mexico
[3] Departamento de Sistemas, Universidad Autonoma Metropolitana de Azcapotzalco, Ciudad de México, Ciudad de México, Mexico

Corresponding Author: Carlos Alberto Hernández-Nava
Email address: cahn@xanum.uam.mx

Nowadays, biometric authentication has gained relevance due to the technological advances that have allowed its inclusion in many daily-use devices. However, this same advantage has also brought dangers, as spoofing attacks are now more common. This work addresses the vulnerabilities of automatic speaker verification authentication systems, which are prone to attacks arising from new techniques for the generation of spoofed audio. In this paper, we present a countermeasure for these attacks using an approach that includes easy to implement feature extractors such as spectrograms and Mel Frequency Cepstral Coefficients, as well as a modular architecture based on deep neural networks. Finally, we evaluate our proposal using the ASVspoof 2017 V2 database and find that it outperforms other reported models in the literature.

Voice spoofing detection using a neural networks assembly considering spectrograms and Mel Frequency Cepstral Coefficients

Carlos Alberto Hernández-Nava[1], Eric Alfredo Rincón-García[2], Pedro Lara-Velázquez[2], Sergio Gerardo de-los-Cobos-Silva[2], Miguel Angel Gutiérrez-Andrade[2], Roman Anselmo Mora-Gutiérrez[3]

[1] Postgraduate Studies in Information Sciences and Technologies, Metropolitan Autonomous University - Iztapalapa, Av. San Rafael Atlixco 186, 09340, Mexico City, Mexico.
[2] Department of Electrical Engineering, Metropolitan Autonomous University - Iztapalapa, Av. San Rafael Atlixco 186, 09340, Mexico City, Mexico.
[3] Department of Systems, Metropolitan Autonomous University - Azcapotzalco, Av. San Pablo Xalpa 180, 02200, Mexico City, Mexico.

Corresponding Author:
Carlos Alberto Hernández-Nava[1]
Email address: cahn@xanum.uam.mx

Abstract

Nowadays, biometric authentication has gained relevance due to the technological advances that have allowed its inclusion in many daily-use devices. However, this same advantage has also brought dangers, as spoofing attacks are now more common. This work addresses the vulnerabilities of automatic speaker verification authentication systems, which are prone to attacks arising from new techniques for the generation of spoofed audio. In this paper, we present a countermeasure for these attacks using an approach that includes easy to implement feature extractors such as spectrograms and Mel Frequency Cepstral Coefficients, as well as a modular architecture based on deep neural networks. Finally, we evaluate our proposal using the ASVspoof 2017 V2 database and find that it outperforms other reported models in the literature.

Introduction

The great growth of social networks in recent years is primarily attributed to the widespread accessibility to many different devices that facilitate the exchange of biometric information, such as computers, cell phones and tablets. These devices enable the transmission of images of human faces, full body videos, as well as audio recordings. Such information is used to train different tools capable of generating high-quality audiovisual media, mainly for entertainment purposes, however, due to the vast amount of information and the current power of these techniques, it is very difficult to distinguish between generated and genuine content. This generated material has found beneficial applications in a wide range of fields, including entertainment and, more recently, the generation of diverse digital media on social networks, unfortunately, as mentioned in [1], it is also being exploited for fraudulent activities.

Today, technology has reached a level of maturity that enables biometric authentication across many different applications and devices. However, it is essential to make an effort to safeguard against identity fraud attempts, especially considering the increasing number of generated media as mentioned in [2]. Specifically, automatic speaker verification (ASV) systems, which are frequently used for speaker verification in telephony, are prone to malicious authentication attempts since they rely solely on the received sound as the means of authentication.

Several models based on neural networks have been developed for the detection of generated or manipulated audios intended for identity theft. Initially, basic neural networks were employed for this purpose. However, as technology progressed, more sophisticated architectures were gradually adopted to enhance their performance in fulfilling this task.

Nowadays, the detection of media created with the intention of being used for counterfeiting has garnered significant attention from the community. As a result, applications are being developed to detect these types of counterfeit files. This work specifically focuses on the growing interest in developing countermeasures against identity theft through automatic speaker verification. The remaining sections of this work are organized as follows: the next subsection presents the most important works related to spoof detection; section two describes our proposed methodology for spoof detection; section three outlines the experiments conducted; section four comprises the discussion derived from the findings and section five presents the conclusions.

Related works

Interest in biometric recognition of speech and speakers is not new. In 2007, a liveness verification system, based on lip movement, was proposed in [3] as a form of protection against identity theft attempts, by means of videos generated for this purpose.

Given to the importance of the problem, the establishment of a standardized database became necessary. In 2015, the National Institute of Informatics initiated two challenges, namely the Voice Conversion Challenge and the ASVspoof Challenge, with de aim of providing an evaluation platform and metrics to facilitate a fair comparison among the proposed techniques related to media cloning and detection.

The Voice Conversion Challenge [4] is a biennial event that started in 2016, in this challenge participants are provided with a database and tasked with developing voice converters using their own methods. The organizers then evaluate and classify the transformed speech provided by the participants.

The ASVspoof challenge [5], which is also a biennial event, is highly relevant to this work. The challenge provides a database comprising pairs of genuine and false or generated audios, which participant's models must accurately classify. Since the release of the ASVspoof challenge databases, investigations have yielded remarkably favorable results. The latest version of the databases was published in 2021.

In [6], the authors propose an architecture that combines convolutional neural networks (CNN) and recurrent neural networks (RNN) simultaneously. To evaluate the effectiveness of their method, they utilized the ASVspoof 2015 database, where the input of their model consisted of spectrograms extracted from the audios. Due to the widely varying durations in this database, the authors decided to standardize the duration to four seconds for all audios. Through their experiments, they achieved an equal error rate (EER) of 1.47%.

It is worth highlighting the effort made to create new detection models. For instance, in [7], the authors proposed a simplified version of the Light CNN architecture that utilizes Max Feature Map (MFM) activation. The authors implemented this network to classify audios into two possible outcomes: genuine or false, specifically to prevent spoofing, they reported an equal error rate of 6.73% in the ASVspoof 2017 database.

In addition to proposing novel models, another factor to consider is the level of complexity required for the proposals. In [8], the authors demonstrated that the implementing very deep or complicated neural networks is not necessarily essential for impersonation detection in identity verification. They explained that satisfactory results can be achieved with simple models. Their model consisted of an input layer, two CNN layers, a Gated Recurrent Unit (GRU) layer, and a final layer. Despite its apparent simplicity, this model yielded excellent performance, with an EER of 0.77% in a corpus of 28,000 audios extracted from the APSRD (Authentic and Playback Speaker Recognition Database).

The alternative approach involves employing increasingly deeper neural networks; however, this can give rise to the vanishing gradient problem. To address this difficulty, residual neural networks (ResNet) have emerged. They have proven to be successful in image recognition, as in [9], where their effectiveness for spoofed audio detection was investigated. The evaluation of this research was conducted using the ASVspoof 2017 dataset, revealing that Resnet achieved one of the best performances among the systems employing a single model.

So far, it hasn't been mentioned whether high-quality audio is truly essential to carrying out an audio attack. However, [10] explored the potential of utilizing solely low-quality data to train models against spoofing. For this purpose, they developed a system based on generative adversarial networks (GAN), to enhance the quality of audio files accessible on the internet.

Understanding the significance of audio quality is crucial to comprehend the nature of the challenges at hand. This is particularly relevant because voice-controlled devices (VCDs) including popular examples like Alexa, Siri, and others are increasingly prevalent. These devices are primarily utilized for automating home appliances and other entertainment tools. Since voice attacks do not necessitate high-quality audio, it can be inferred that these devices might be vulnerable to such attacks.

In an analysis conducted by [11], multi-hop replay attacks were shown to be a vulnerability since they are carried out using VCDs with the intention of accessing other VCDs connected to the internet. For example, a device could be used to replicate the voice of the speaker, giving an order or command to a second VCD, and the latter would fulfill its function without verifying whether the instruction truly originates from the speaker or is simply a repetition of the voice of the user.

After establishing the vulnerability of VCDs, [12] presents another concern regarding the number of channels employed in attacks on these devices. They present a model based on neural networks that specifically targets multichannel audio for detection purposes. The proposed model allows for an arbitrary number of input channels, in addition, what can be highlighted is that their model is fully developed in a neural networks framework, enabling potential integration with other neural network-based models in the future.

Combinations of neural networks are frequently employed in recent and advanced studies. In [13], the authors examined the performance of various architectures, incorporating deep neural networks (DNNs), long short-term memory (LSTM) layers, temporal convolution (TC), and spatial convolution (SC). To evaluate the performance of their proposal, they used ASVspoof 2015 and ASVspoof 2019 databases, achieving good results in both, with particularly impressive results in the latter.

It is worth mentioning that there are two main research approaches for spoof detection. The first approach involves using diverse architectures and classification models, including the Gaussian Mixture Model (GMM), support-vector machines (SVM), neural networks such as CNN, RNN, LSTM, GAN, ResNet, and even autoencoders.

The second research approach involves using diverse techniques for audio feature extraction, including different types of spectrograms or coefficients such as Mel Frequency Cepstral Coefficients (MFCC), Inverse Mel Frequency Cepstral Coefficients (IMFCC), Complex Cepstral Coefficients (CCC), Linear Frequency Cepstral Coefficients (LFC), Constant Q Cepstral

Coefficients (CQCC), Teager Energy Cepstral Coefficients (TECC), Linear Predictive Cepstral Coefficients (LPCC), as well as various combinations thereof.

Although some of the extraction techniques and proposed architectures are highly efficient, they can be complex to comprehend and replicate. Therefore, we believe it is crucial to propose a technique that is both easily understandable and reproducible, while maintaining high precision in the specific task of audio spoof detection.

Materials & Methods

To accurately classify genuine and spoofed audio, it is necessary to identify and extract useful information from them. In this study, we found that combining the use of spectrograms and Mel frequency cepstral coefficients is sufficient to achieve higher accuracy that the state-of-the-art methods.

Spectrograms

To obtain the spectrograms, samples are taken through a time window to calculate the frequency content of the samples using the short time Fourier transform (STFT). This process involves extracting and analyzing several frames of a signal at each window displacement over time. Each frame is added to a matrix that represents the variation in the spectrum and energy of the signal. As new frames are obtained, they are consecutively added to the first position in the array. In this way, the variation of the signal's spectrum and energy can be represented as a function of time.

After conducting some tests, it became clear that the linear spectrograms were not capturing all the necessary information to distinguish between a spoof audio and a genuine one. Therefore, we decided to use spectrograms with a logarithmic scale to better capture audio information. As we expected, after performing additional experiments, we found that both linear and logarithmic spectrograms were necessary to achieve high accuracy. In the following paragraphs, we provide more details on these findings.

Although there are many representations available, for this work, we considered the time on the abscissa axis as consecutive sequences of Fourier transforms, the frequency expressed in Hz on the ordinate axis, and finally the representation of the energy expressed in dB represented with a color palette, Fig. 1 includes a time versus frequency linear spectrogram.

Figure 1: Spectrogram.

The main modification for logarithmic spectrograms involves changing the scale of the ordinate axis from a linear to a logarithmic scale. For this study, we considered two different areas of interest. The first area includes values between 10^2 and 10^4, as shown in Fig. 2a. This interval was selected because it represents the range where the greatest amount of sound wave energy is concentrated for human voice. The second zone was reduced to values between 10^3 and 10^4 to focus on the zone that corresponds to the highest frequencies and exclude the lowest ones, as shown in Fig. 2b.

Figure 2: Spectrograms with logarithmic scale.

MFCC

The most commonly reported feature extraction technique for spoof detection in the specialized literature is the Mel-frequency cepstral coefficients, originally proposed by [14]. These are the widely used features to represent the human voice and have shown good results in various environments.

Before extracting MFCCs, an analog signal is converted to a digital signal by sampling and at a specific sample rate. The digital signal is subjected to a series of processes, as shown in Fig. 3, to extract the MFCC features.

Figure 3: MFCC feature extraction process.

After dividing the analog signal into overlapping frames, the Discrete Fourier Transform (DFT) of the signal is calculated. Next, the signal is filtered using the Mel filter bank, and the output is log compressed. It is then transformed to the cepstral domain using the Discrete Cosine Transform (DCT), preserving the first 13 coefficients while discarding the higher ones. As mentioned in [15], this is because 13 coefficients are sufficient for representing the speech signal.

In this work, MFCCs are used in two different ways. First, 13 coefficients are extracted for each window, resulting in a 13 x 298 matrix. Subsequently, the matrix is reduced to a vector of 13 coordinates, by calculating the average of all the windows. Both the matrix and the vector are used to classify the audios.

Models for spectrograms

To extract information from the spectrograms and properly classify the audios, we propose using two models based in convolutional neural networks.

The first model comprises four 2-D convolutional layers with 32, 64, 96, and 96 filters, respectively. After each convolution, a batch normalization and a max-pooling are performed. Finally, a flatten layer and two dense layers are included, as shown in Fig. 4. We trained two copies of model 1: one with linear spectrograms and the other with logarithmic spectrograms with values between 10^2 and 10^4.

Figure 4: Model 1, CNN with 2-D layers and batch normalization.

The second model was trained using logarithmic spectrograms with values between 10^3 and 10^4. The network comprises three time-distributed 2-D convolutional layers with 32, 64, and 96 filters respectively. After each convolution, a time-distributed max-pooling is performed. Next, a flatten layer, a dense layer and a dropout layer are included. Finally, a dense layer completes the model, as shown in Fig. 5.

Figure 5: Model 2, CNN with tD 2-D layers and dropout.

Models for MFCCs

For the MFCCs coefficients, we used two additional models.

Model 3 takes the vector representation of the MFCCs as input and passes the 13 coefficients through three time distributed 1-D dense layers with 24, 13 and 10 units, respectively, along with a dropout layer. Next, we added three LSTM layers with 10, 15 and 30 units, followed by a second dropout layer. Finally, a dense layer with 30 units was included, as can be seen in Fig. 6.

Figure 6: Model 3, CNN with tD and convolutional 1-D layers.

Model 4 receives the two-dimensional MFCCs as input. The network comprises a time distributed 2-D convolutional layer with 36 filters, followed by a batch normalization layer, a second 2-D convolutional layer with 64 filters, and a max pooling layer. We then included two dense layers with 24 units each, followed by a flatten layer and a dense layer, as can be seen in Fig. 7.

Figure 7: Model 4, CNN with tD 2-D layers to work with MFCCs.

All models were trained separately and then assembled to form the final architecture, as shown in Fig. 8. For each audio, the linear spectrogram, the two logarithmic spectrograms described above, and the MFCC coefficients are calculated and introduced into their respective models. The outputs of these models are then introduced into two dense layers, and the final prediction for classification is made.

Figure 8: Final architecture of the assembly of the proposed models.

Results

To evaluate the performance of our proposed methodology, we used the ASVspoof 2017 V2 database, which is the second version of the database used in the ASVspoof 2017 challenge [16], [17]. This database is focused on replay attacks, which are generated by recording the voice of a genuine speaker and then replaying it to an ASV system instead of using the genuine speech of the person.

The corpus consists of 13,306 audio files of varying durations, divided into three datasets as shown in Table 1. The first dataset contains 1,507 genuine and 1,507 generated audios for training. The second dataset consists of 760 genuine and 950 generated audios for development. Finally, the third dataset consists of 1,298 genuine and 12,008 generated audio files for evaluation.

Table 1: Description of the ASVspoof 2017 V2 database.

To homogenize the audios, we decided to make them all have a duration of 3 seconds. This was based on the findings of [6], where it was mentioned that if the audios have a duration of less than 2.5 seconds, favorable results are not obtained. Furthermore, after analyzing the database, we found that most of the audios have a duration close to three seconds, therefore, we determine that a duration of 3 seconds is already sufficient to achieve a good performance.

In this work, when an audio was shorter, silence was added, and for longer audios only the initial three seconds were used. This strategy was as a preprocessing step for all the feature extraction techniques used in this document. A similar strategy is used by [8], where if the sample is too long, they simply cut the excess part, and if the sample is too short, they concatenate the original sample with itself to obtain a desired length.

The performance of the proposed methodology was evaluated using the Equal Error Rate or Crossover Error Rate (CER), which represents the point at which the false rejection rate (FRR) and the false acceptance rate (FAR) are equal. It is expected that higher accuracy will result in a lower EER, as optimal performance is achieved with 100% accuracy or an EER equal to 0.

As previously mentioned, all models were individually trained and tested before being assembled into the final architecture. Table 2 presents the accuracy and ERR (expressed as a percentage) achieved by each proposed model.

Table 2: Accuracy and EER for each model.

The next step involves assembling the five feature extraction processes and their corresponding models to establish a correspondence rule between the audios and their respective classification. As expected, the final architecture outperforms the individual models when they work independently, as evidenced by the results in Table 3, where an accuracy of 96.46% and an EER of 6.66% were achieved.

Table 3: Accuracy and EER for assembly.

Discussion

To properly evaluate our proposal, we compared it with other models reported in literature that used the ASVspoof 2017 V2 database. Table 4 presents the EER reported by different authors, along with the extraction techniques and the classifiers used by them. Our proposal achieved the lowest EER among all the models compared. Thus, our methodology, which uses Mel Frequency Cepstral Coefficients and linear and logarithmic spectrograms, along with CNNs, exceeds state-of-the-art results.

(References in Table 4: [17], [18], [19], [20], [21], [22], [23], [24], [25], [26])

Table 4: Comparison of proposed approach with existing techniques, in the ASVspoof 2017 V2 database.

Individually, the proposed models exhibit adequate but not outstanding behavior. Most of the models in this research extract an image directly from the audio files, namely spectrograms. Although MFCCs are not images, they generate a matrix that can be treated as an image, allowing us to leverage the power of CNNs for tasks involving image processing, such as segmentation and recognition.

It is worth highlighting the importance of preprocessing the audio files to obtain a more homogeneous database to perform the feature extraction. This ensures that the features serve their intended purpose in training the models, even when dealing with the unbalanced ASVspoof 2017 V2 database used in this work.

Conclusions

As artificial intelligence continues to improve, it becomes increasingly easier to generate spoofed audio that is harder to detect. In this paper, we describe a methodology that can be easily implemented and achieves high accuracy in detecting spoofed audio. Our findings suggest that MFCCs and linear and logarithmic spectrograms are sufficient to achieve outstanding performance, and the advantage is that these features can be easily calculated using established libraries. Finally, the information is processed by either CNN or DNN, and the classification is completed with only a small number of misclassified audio files.

The strength of this proposal is achieved by not only combining various techniques for extracting useful information from audio, but also by proposing different neural network architectures to be used with each technique, which highlights the importance of using a tailored approach for each technique.

Indeed, the truly outstanding result is achieved by the architecture of the final assembly, which merges the predictions given by each proposed model and generates a final classification. This assembly has not only proven to have an adequate behavior but also is above the cutting-edge results.

Data availability

The Version 2 of the database used for the second Automatic Speaker Verification Spoofing and Countermeasures Challenge, for short, ASVspoof 2017 V2 was used in this study, this dataset is online available in: https://doi.org/10.7488/ds/2332

References

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *SPEECH Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[2] I. Echizen *et al.*, "Generation and Detection of Media Clones," *IEICE Trans. Inf. Syst.*, vol. E104D, no. 1, pp. 12–23, Jan. 2021.

[3] M.-I. Faraj and J. Bigun, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1169–1175, Sep. 2007.

[4] T. Toda *et al.*, "The Voice Conversion Challenge 2016," in *Interspeech 2016*, 2016, pp. 1632–1636.

[5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, and M. Sahidullah, *ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge*. 2015.

[6] C. Zhang, C. Yu, and J. H. L. Hansen, "An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 684–694, Jun. 2017.

[7] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *18th Annual Conference Of The International Speech Communication Association (Interspeech 2017), Vols 1-6: Situated Interaction*, 2017, pp. 82–86.

[8] W. Pang and Q. He, "A Simple Neural Network Based Countermeasure for Replay Attack," in *Proceedings Of 2017 2nd International Conference On Communication And Information Systems*, 2017, pp. 234–238.

[9] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection," in *18th Annual Conference Of The International Speech Communication Association (Interspeech 2017), Vols 1-6: Situated Interaction*, 2017, pp. 102–106.

[10] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, “Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama’s voice using GAN, WaveNet and low-quality found data,” *Odyssey 2018 Speak. Lang. Recognit. Work.*, Jun. 2018.

[11] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, “A Light-Weight Replay Detection Framework For Voice Controlled IoT Devices,” *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 982–996, Aug. 2020.

[12] Y. Gong, J. Yang, and C. Poellabauer, “Detecting Replay Attacks Using Multi-Channel Audio: A Neural Network-Based Method,” *IEEE Signal Process. Lett.*, vol. 27, pp. 920–924, 2020.

[13] M. Dua, C. Jain, and S. Kumar, “LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems,” *J. Ambient Intell. Humaniz. Comput.*, vol. 13, pp. 1985–2000, 2021.

[14] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[15] A. Kuamr, M. Dua, and T. Choudhary, “Continuous Hindi Speech Recognition Using Gaussian Mixture HMM,” in *2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2014.

[16] T. Kinnunen *et al.*, *The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection*. 2017.

[17] H. Delgado *et al.*, “ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements,” in *Odyssey 2018 - The Speaker and Language Recognition Workshop*, 2018.

[18] B. Wickramasinghe, E. Ambikairajah, J. Epps, V. Sethu, and H. Li, *Auditory Inspired Spatial Differentiation for Replay Spoofing Attack Detection*. 2019.

[19] R. K. Das and H. Li, “Instantaneous Phase and Excitation Source Features for Detection of Replay Attacks,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1030–1037.

[20] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, *Phoneme Specific Modelling and Scoring Techniques for Anti Spoofing System*. 2019.

[21] S. Jelil, S. Kalita, S. R. M. Prasanna, and R. Sinha, “Exploration of Compressed ILPR Features for Replay Attack Detection. BT - Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.” pp. 631–635, 2018.

[22] M. R. Kamble and H. A. Patil, “Detection of replay spoof speech using teager energy feature cues,” *Comput. Speech Lang.*, vol. 65, p. 101140, 2021.

[23] B. B T, L. Kin Wah Edward, S. Lui, J.-M. Chen, and D. Herremans, “Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features,” *IEEE Access*, vol. PP, p. 1, 2019.

[24] M. R. Kamble, H. Tak, and H. A. Patil, “Amplitude and Frequency Modulation-based features for detection of replay Spoof Speech,” *Speech Commun.*, vol. 125, pp. 114–127, 2020.

[25] P. A. Tapkir, M. R. Kamble, H. A. Patil, and M. Madhavi, “Replay Spoof Detection using Power Function Based Features,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1019–1023.

[26] J. Yang, R. K. Das, and H. Li, “Extended Constant-Q Cepstral Coefficients for Detection of Spoofing Attacks,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1024–1029.

Figure 1

Spectrogram.

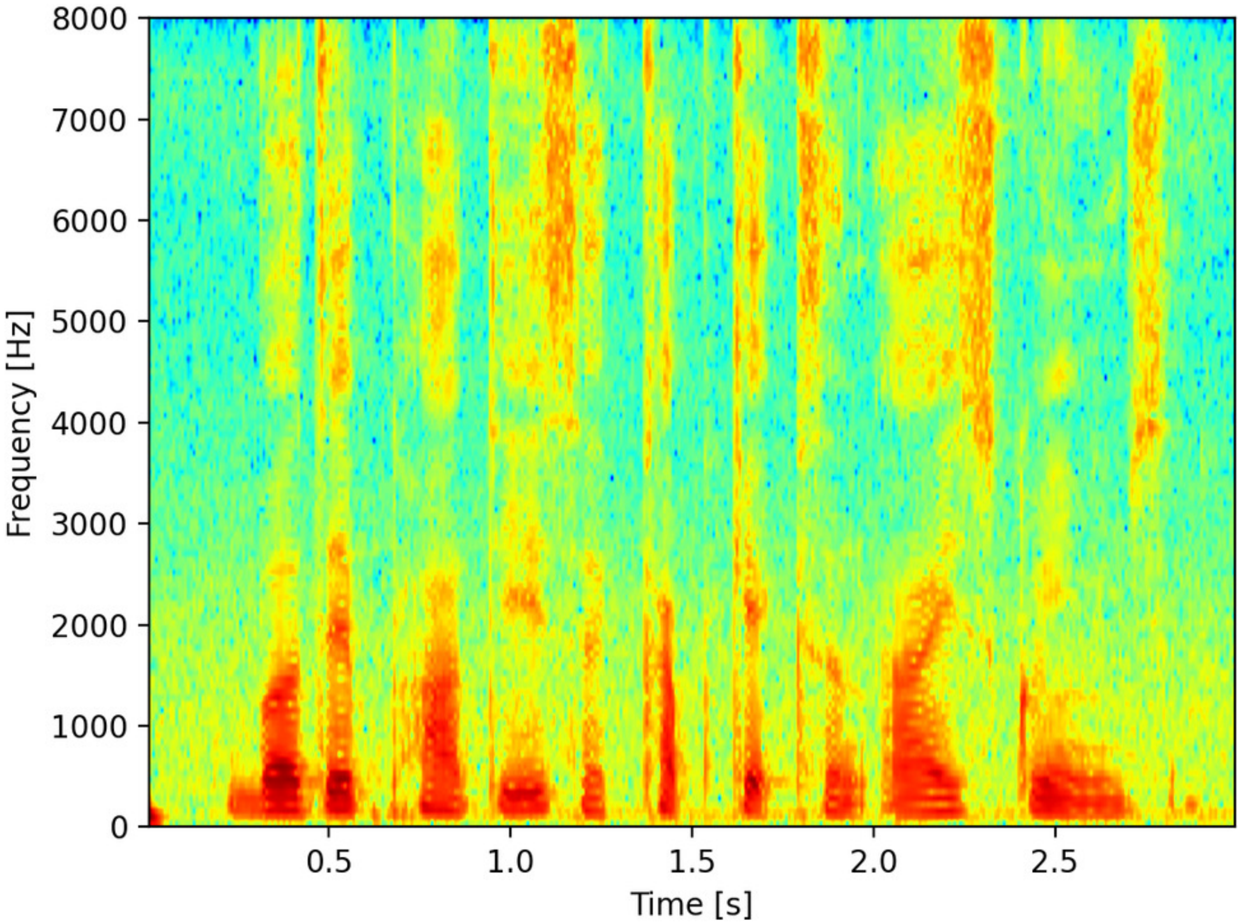

Figure 2

Spectrograms with logarithmic scale.

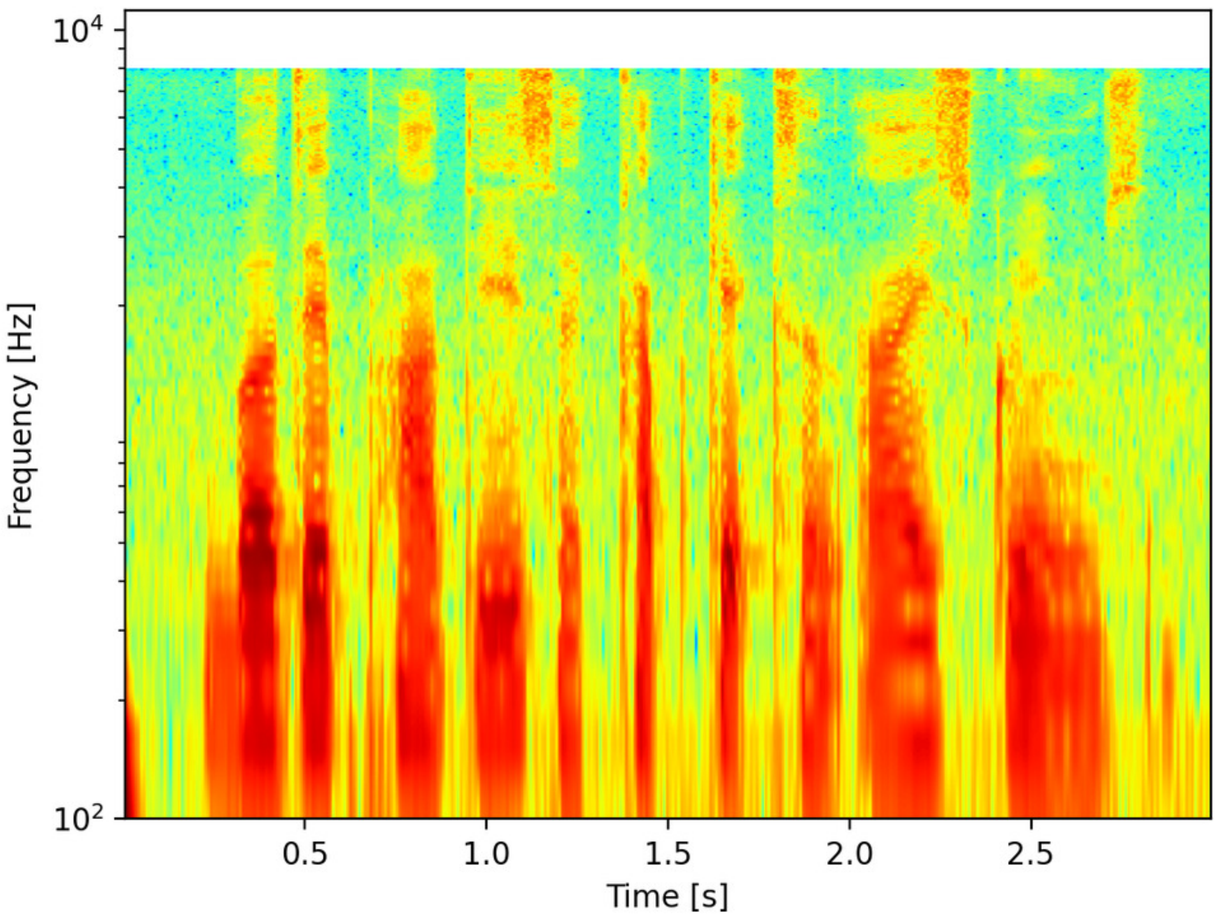

Figure 3

Spectrograms with logarithmic scale.

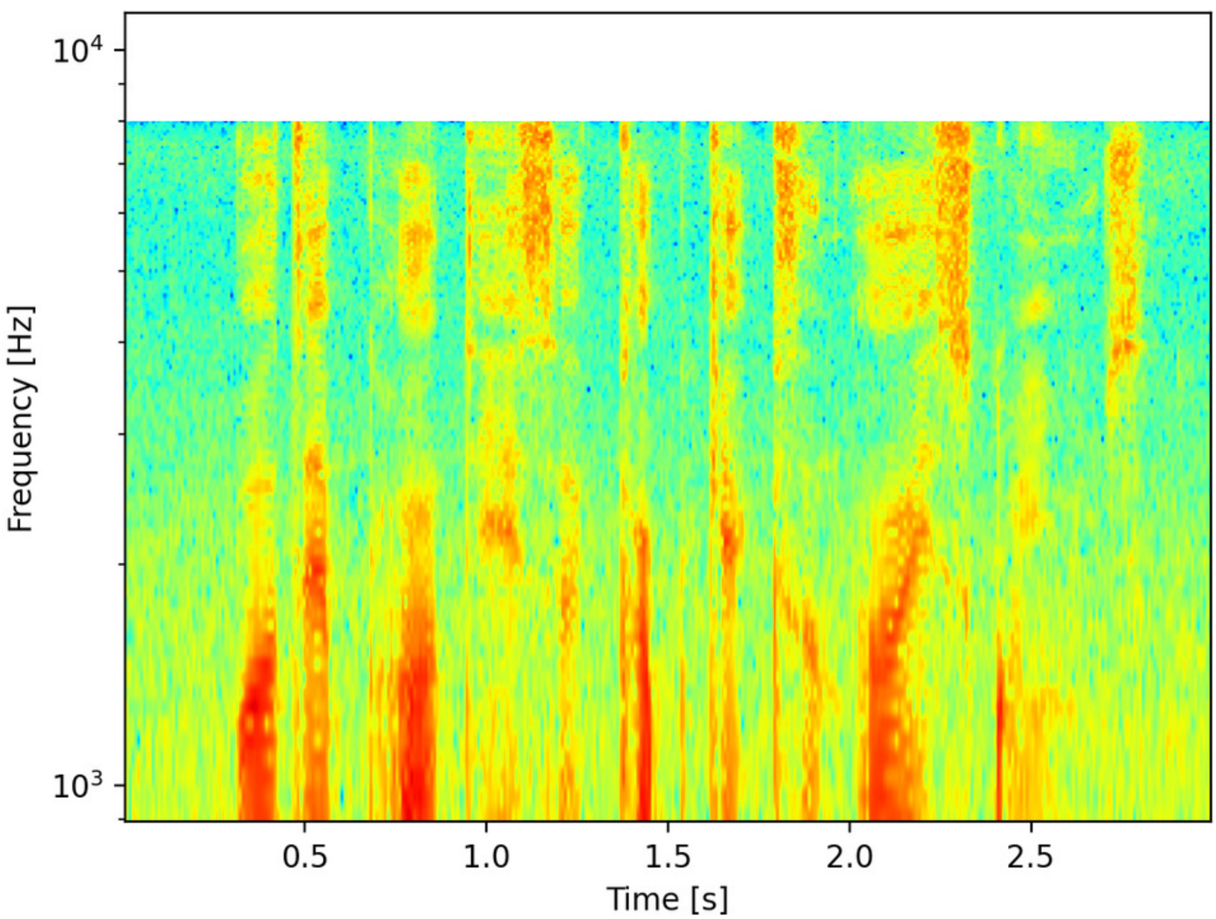

Figure 4

MFCC feature extraction process.

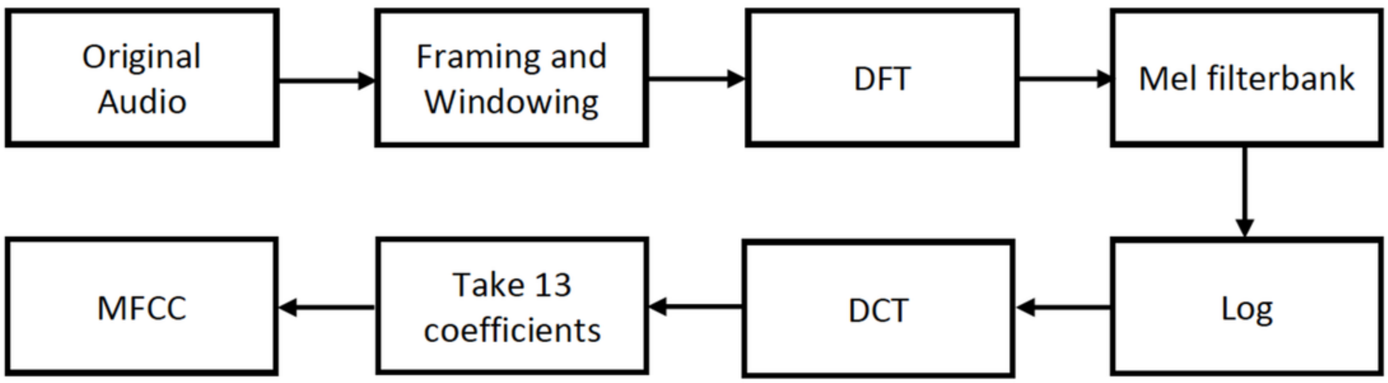

Figure 5

Model 1, CNN with 2-D layers and batch normalization.

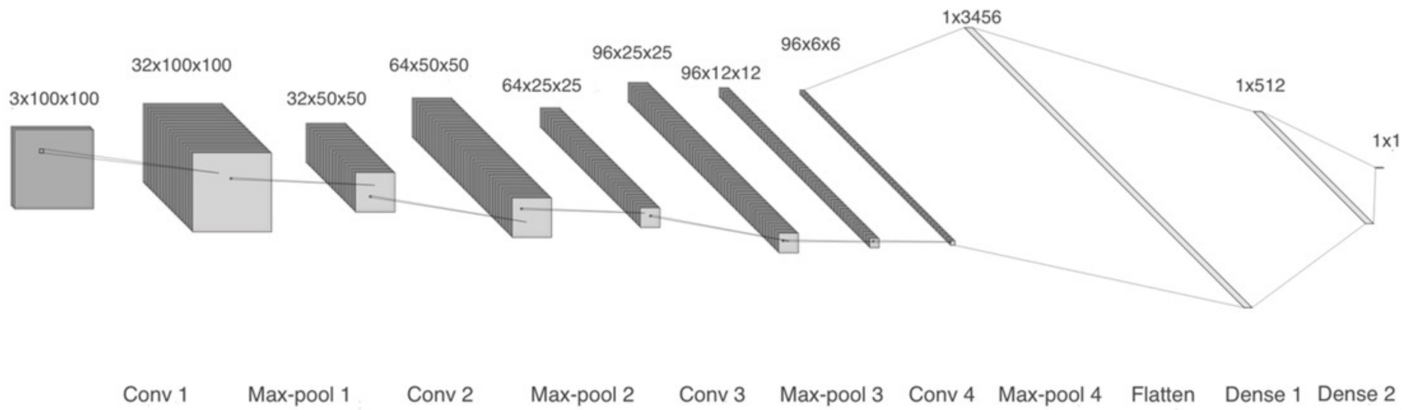

Figure 6

Model 2, CNN with tD 2-D layers and dropout.

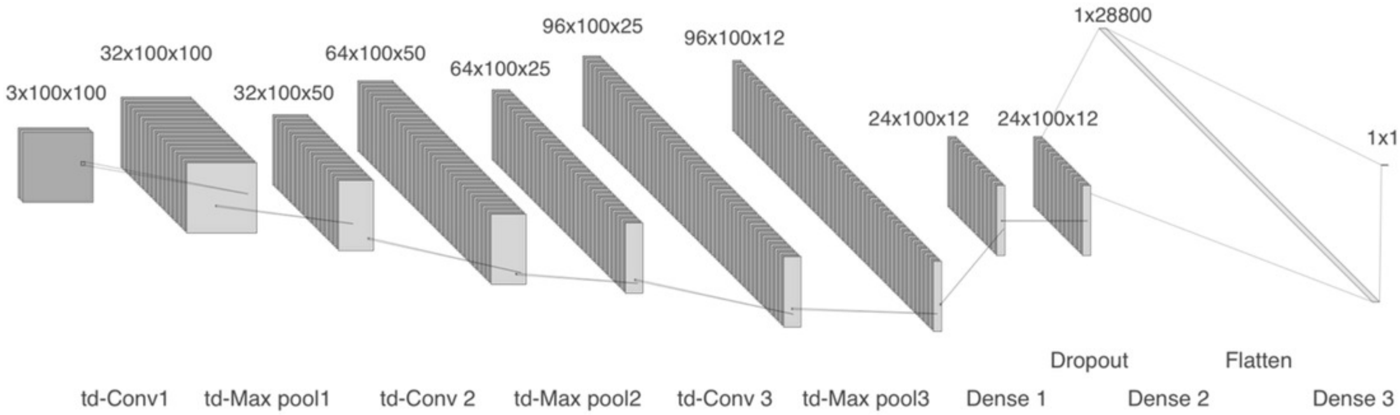

Figure 7

Model 3, CNN with tD and convolutional 1-D layers.

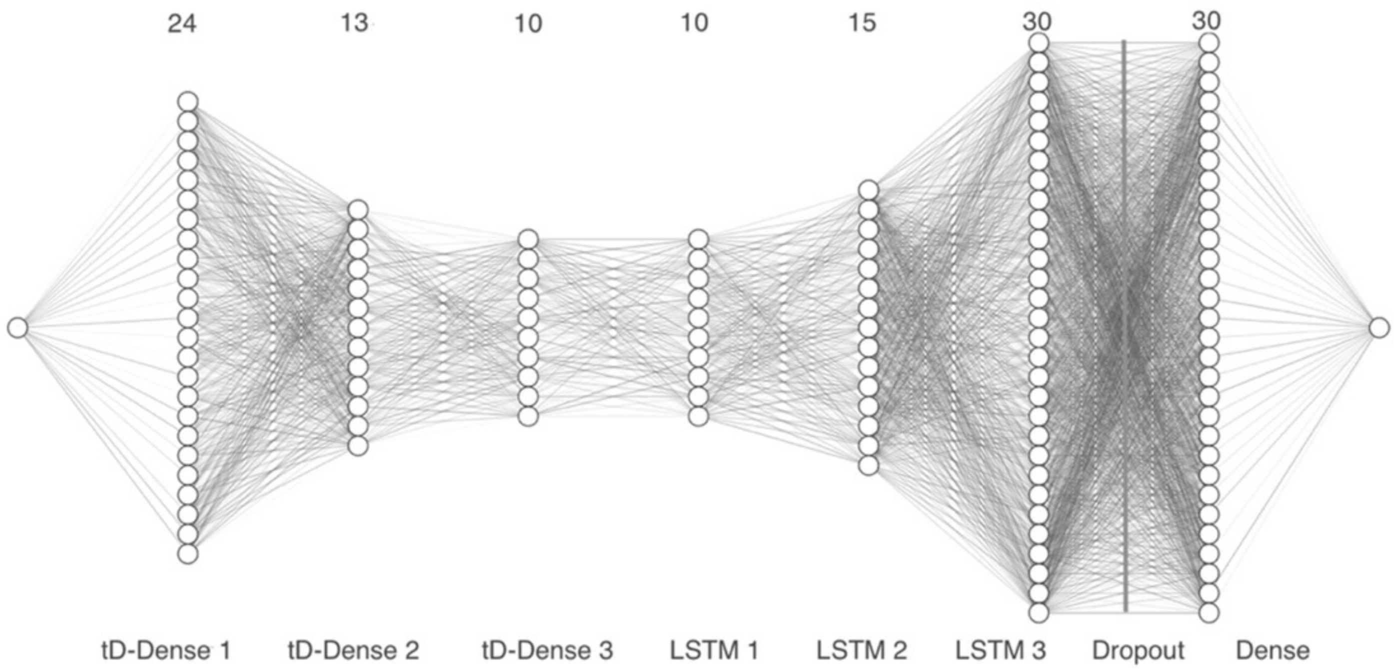

Figure 8

Model 4, CNN with tD 2-D layers to work with MFCCs.

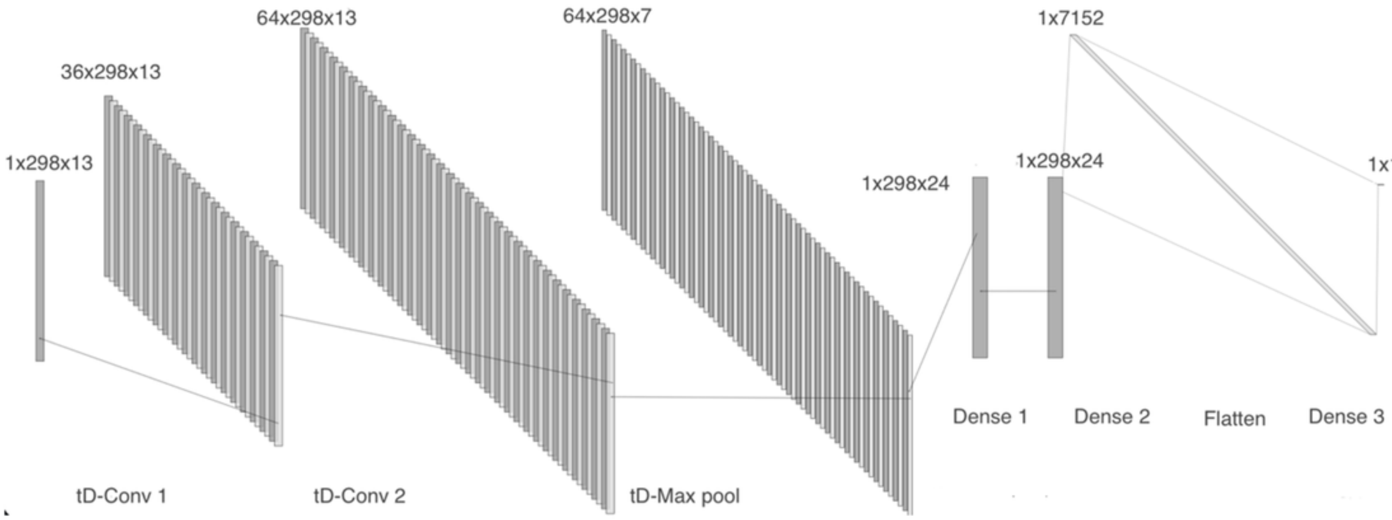

Figure 9

Final architecture of the assembly of the proposed models.

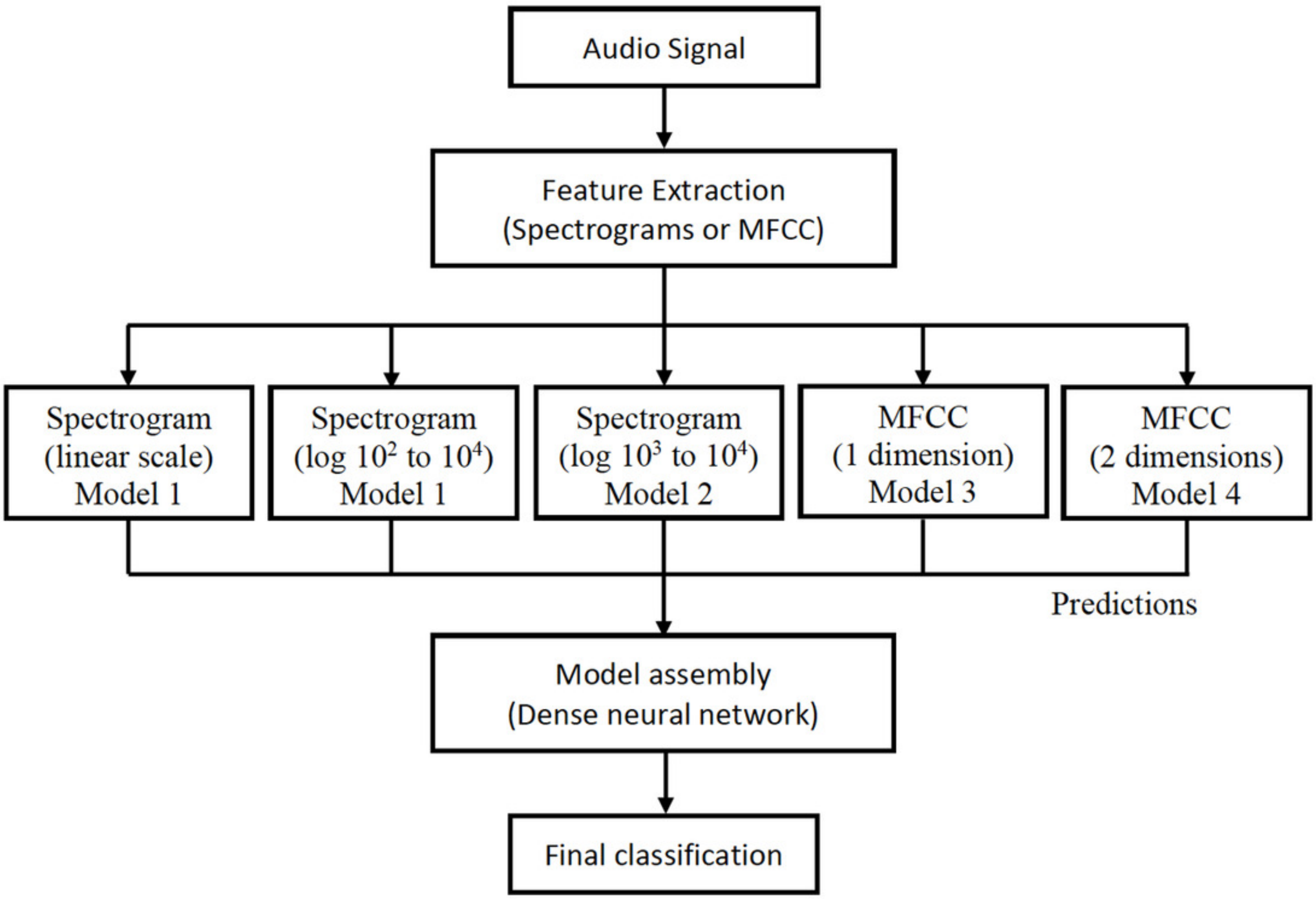

Table 1 (on next page)

Description of the ASVspoof 2017 V2 database.

Table 1: Description of the ASVspoof 2017 V2 database.

Dataset	Genuine audio	Spoof audio
Training	1507	1507
Development	760	950
Evaluation	1298	12008

Table 2 (on next page)

Accuracy and EER for each model.

Table 2: Accuracy and EER for each model.

Feature extraction	Training accuracy (%)	Evaluation accuracy (%)	EER (%)
Linear Spectrograms	99.41	72.67	22.67
Log spectrograms (10^2, 10^4)	93.20	64.02	23.13
Log spectrograms (10^3, 10^4)	92.20	70.71	26.12
MFCC (vector)	98.11	68.98	21.06
MFCC (matrix)	87.93	81.42	22.80

Table 3(on next page)

Accuracy and EER for assembly.

Table 3: Accuracy and EER for assembly.

Model	Evaluation accuracy (%)	EER (%)
Assembly	96.46	6.66

Table 4 (on next page)

Comparison of proposed approach with existing techniques, in the ASVspoof 2017 V2 database.

Table 4: Comparison of proposed approach with existing techniques, in ASVspoof 2017 V2 database.

Approach	Feature extraction	Classifier	Eval EER (%)
Proposed Assembly	Spectrograms + MFCC	DNN	**6.66**
Wickramasinghe, et al., 2019 [18]	CF + CM	GMM	8.58
Das, et al., 2018 [19]	CQCC+IFCC, DCTILPR+RMFCC	GMM	9.01
Suthokumar, et al., 2019 [20]	PPRFWS_LR	GMM	9.28
Jelil, et al., 2018 [21]	CQCC + CILPR	GMM	9.41
Kamble, et al., 2021 [22]	CQCC + LFCC + MFCC + TECC	GMM	10.45
Balamurali, et al., 2019 [23]	MFCC + IMFCC + CQCC + CCC + RFCC + LFCC + LPCC + Spectrogram + Autoencoder features	GMM	10.8
Delgado, et al., 2018 [17]	CQCC (Baseline)	GMM	12.24
Kamble, et al., 2020 [24]	CQCC + ESA-IFCC	GMM	12.93
Tapkir, et al., 2018 [25]	CQCC + PNCC	GMM	12.98
Yang, et al., 2018 [26]	eCQCC-DA	DNN	13.38