Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure

Bence Mark Halpern[123]*, Finnian Kelly*[4]*, Rob van Son*[12]*, Anil Alexander*[4]

[1]University of Amsterdam, ACLC, The Netherlands
[2]Netherlands Cancer Institute, Amsterdam, The Netherlands
[3]TU Delft, Delft, The Netherlands
[4]Oxford Wave Research Ltd, Oxford, United Kingdom

b.halpern@nki.nl, finnian@oxfordwaveresearch.com
r.v.son@nki.nl, anil@oxfordwaveresearch.com

Abstract

In this paper we propose a spoofing countermeasure based on Constant Q-transform (CQT) features with a ResNet embeddings extractor and a Gaussian Mixture Model (GMM) classifier. We present a detailed analysis of this approach using the Logical Access portion of the ASVspoof2019 evaluation database, and demonstrate that it provides complementary information to the baseline evaluation systems. We additionally evaluate the CQT-ResNet approach in the presence of various types of real noise, and show that it is more robust than the baseline systems. Finally, we explore some explainable audio approaches to offer the human listener insight into the types of information exploited by the network in discriminating spoofed speech from real speech.

1. Introduction

As artificially generated (or manipulated) speech becomes more naturalistic to the human listener, it is important to consider the impact of 'spoofed' speech on automatic speaker verification (ASV) systems. In response to this need, the development of spoofing countermeasures to detect spoofed speech has become an active area of research. The ASVspoof initiative[1] [1] was established to encourage research into spoofing countermeasures via a series of evaluations focused on different classes of spoofed speech. In this paper, we focus on the development and analysis of a spoofing countermeasure using the Logical Access portion of the ASVspoof2019 evaluation dataset [1], which consists of a diverse collection of text-to-speech (TTS) and voice conversion (VC) spoofed speech samples.

The recent wave of Deep Neural Network (DNN) based approaches to speech synthesis have contributed to a rise in naturalness of spoofed speech; it is logical therefore that detecting spoofed speech is increasingly done with such architectures, e.g. [2, 3]. We consider the use of a Dilated ResNet for spoofing detection. A Dilated ResNet [4] is a DNN with dilated convolutional layers, enabling larger time-dependencies to be captured while keeping model complexity constant.

Gaussian Mixture Models (GMMs) have featured prominently in voice conversion approaches, and also as classifiers for spoofing detection [5]. Here we propose a spoofing countermeasure that uses a GMM to classify an embedding layer of a Dilated ResNet. We also evaluate the performance of the ASVspoof2019 baseline GMM systems for comparison.

[1]www.asvspoof.org

Constant Q-transform (CQT) [6] and constant Q-transform based cepstral coefficients (CQCCs) have become popular feature extraction approaches for spoofing detection, having been shown to outperform conventional log spectrogram or log Mel spectrogram based features [7]. The CQT is closely related to the Fast Fourier Transform (FFT), but with logarithmically rather than linearly spaced frequency bins. This results in higher temporal resolution, but lower frequency resolution in high frequency bins. A reason for the effectiveness of the CQT may be that speech synthesisers are precisely optimised for psychoacoustic quality with the Mel based representations, however they can still contain perceptible artifacts from a machine's point of view. The constant Q-transform was originally developed for music processing, which indicates that not only psychoacoustic qualities, but also musical quality of speech, i.e. harmonies might have a role in the design of successful spoofing countermeasures. In this paper, we consider the use of CQT features within a dilated ResNet architecture to bring in complementary information for noise robustness. We will explore how robust these techniques are and show their trade-offs.

The motivation behind developing a spoofing countermeasure (CM) in a speaker recognition pipeline is to distinguish genuine (*bonafide*) speech from spoofed speech. Thus, a successful CM should learn the characteristics of realistic speech (i.e. naturalness). Many systems in the ASVspoof2019 challenge performed poorly with voice conversion (VC) attacks [8]. This is likely to be because these utterances have substantial difference in their naturalness compared to their text-to-speech (TTS) and hybrid (VC-TTS) counterparts. One way to verify this would be to ask human listeners to rate these utterances for their naturalness, and then find correlations between a CM's decision and the utterance ratings. This would require a substantial time investment. However, quite recently MOSNet [9] has been introduced to provide an objective way to evaluate utterances on a large-scale, which allows us to automatically rate naturalness of ASVspoof2019 utterances and perform such experiments. This enables us to explore the extent to which a CM models the naturalness of speech.

There have been several empirical studies on the noise robustness of spoofing countermeasures [10, 11, 12] using earlier ASVspoof evaluation datasets. All of these works agree that noisy environments affect the spoofing detection performance significantly. In this work, we test the robustness of the CQT-ResNet architectures in the presence of real noise. For comparison, we assess the robustness of linear frequency cepstral

coefficient (LFCC) and CQCC based GMM countermeasures in parallel.

A common objection to DNN-based classifiers is an inability to explain the classification decision. Explainable machine learning techniques are readily available for computer vision applications [13], but are currently lacking in audio processing. In this paper, we demonstrate two explainable audio based methods to give an idea about the acoustic cues that the spoofing countermeasure neural network uses to discriminate spoofed from bonafide speech.

2. Dataset

We base our analysis in this paper on the ASVspoof2019 logical access dataset [1], which includes three main classes of attacks: text-to-speech synthesis (TTS), voice conversion (VC), and hybrid TTS-VC speech, along with bonafide speech. Physical access attacks such as replay and mimicry are not considered in this paper.

An observation from the ASVspoof2019 evaluation results [1] was that the spoofing detection performance across different logical access attacks was more variable than in the case of physical access replay attacks; we therefore consider it a more challenging task for a single spoofing detection system. Furthermore, we expect the greater variability in the logical access audio to be more interesting from an explainable audio perspective.

We follow the ASVspoof2019 protocols for training and testing our proposed systems (i.e. using the training and evaluation lists respectively). Additionally, we select a subset of the development and evaluation data to create a noisy test set, which is explained in Section 3.4.

3. Methods

3.1. Feature extraction

All audio files were preprocessed by applying a CQT with 250 samples between successive frames (hop length), with 240 bins in total and 48 bins per octave using librosa's CQT-implementation [14]. The LFCCs were calculated using librosa, and the CQCCs were directly extracted from the published baseline feature extractor in MATLAB.

3.2. Classifier architectures

We used two different types of neural network classifiers, which have a Dilated ResNet architecture based on [15] as their underlying classifier. A Dilated ResNet [4] is a deep neural network using dilated convolutional layers. Dilated convolutional layers are expansions of traditional convolutional layers: i.e. with a dilation of (2,2), a 3x3 kernel is padded with zeroes, resulting in a 5x5 kernel, see Figure 1. This enables longer spatiotemporal relationships to be learned with the same number of parameters. We are therefore able to learn higher resolution feature maps with reduced model complexity.

We believe it is advantageous for a CM to mimic the spoof generators' architecture. WaveNet, a state-of-the-art speech waveform generator [17] is used as part of two different spoofing attacks in this dataset. This architecture also uses dilated convolutions, thus we think a Dilated ResNet is a suitable choice as a CM neural network.

The first network we consider has a simple softmax layer at the end, which we call **CQT-DNN**, see Figure 2. The second network we consider uses the embeddings learnt by the neural

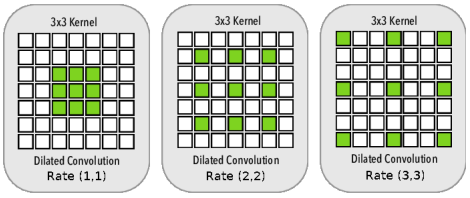

Figure 1: Illustration of dilated convolution reproduced from [16]. Increasing the dilation rate does not change the size of the kernel (green boxes). This means that the number of parameters learned does not change, while longer spatiotemporal relationships can be modelled.

network to train two Gaussian mixture models (**CQT-GMM-DNN**). This second approach is motivated by the ability to reject classification decisions. As the last (softmax) layer of a neural network outputs a normalised probability mass function for the classes, we are unable to get a degree of confidence from the neural networks. With GMMs, we obtain log likelihoods, which can be thresholded to reject classification decisions. If desired, this can be used to include a human in the loop of CMs, or to reason about classification errors in general (i.e. by analysing samples with low evidence) [18]. We note also that the ability of embeddings to provide a compact and generalised description of the audio has been shown to be effective for other audio classification tasks like speaker recognition [19].

We trained the network with the full training set using an Adam optimiser [20] with a learning rate of $\alpha = 0.001$. The input was zero padded during training to a fixed 400 frame input size. This kind of padding is justified with ResNet, because the max pooling should do away with filter activations related to the zero padding of the sequence. The size of the embedding layer was 100. The neural network was implemented in Keras [21].

To provide a baseline for comparison, we reproduced the **LFCC-GMM** and **CQCC-GMM** baseline systems from the ASVspoof2019 evaluation in Python. We trained our baseline systems with a 10% random subset of the training set. We verified that our Python implemented baselines only differ marginally from the published results in [7], i.e. our CQCC-GMM EER is 9.52%, while the official CQCC-GMM baseline EER is 9.57%. For the LFCC-GMM, our reimplementation's EER is 9.06%, the official LFCC-GMM baseline being 8.09%.

3.3. Performance measure

The performance for the neural networks are reported in terms of the equal error rate (EER):

$$\text{EER} = P_{\text{fa}}(s_{\text{EER}}) = P_{\text{miss}}(s_{\text{EER}}), \quad (1)$$

where $P_{\text{fa}}(\cdot)$ refers to the probability of false acceptance, $P_{\text{miss}}(\cdot)$ indicates the probability of false rejection, and s_{EER} is the predicted score of the CM.

To analyse which classes contribute most to the equal error rate, we introduce the notion of classwise EERs. Calculation of classwise EERs is identical to normal EERs, except not all spoofed examples are considered as targets, but only the selected class.

In the ASVspoof2019 challenge, the t-DCF score was also introduced as an evaluation metric. [22]. The t-DCF is a cost function taking into account (1) the risk associated with misclassification within a Bayesian framework along with (2) the

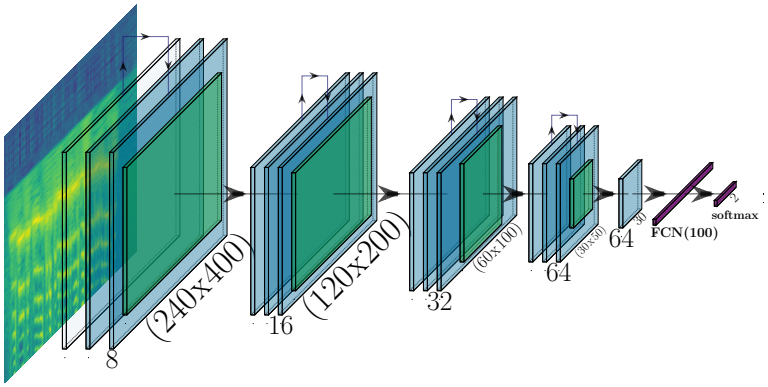

Figure 2: The Dilated ResNet architecture, consisting of four Dilated ResNet blocks. The green layers correspond to max pooling layers (1/2 pooling). This means that the maximum of the activations are taken in a (2,2) window with a stride of 1 to halve (downsample) the feature map. The blue layers correspond to dilated convolutional layers. The blocks have dilation rates: (2,2), (4,4), (4,4) and (8,8). The two-way arrows above the layers represent the skip connections. In the last pooling layer only the vertical axis is pooled. The numbers below the layers indicate the number of filters and the size of the convolutional layers. The kernel size is always (3x3). FCN stands for fully connected layer. Note that aspect ratio is not conserved.

performance of an ASV system. The advantage of this metric compared to CM EER is that it takes into account the vulnerabilities of an ASV system. For example, if an ASV is vulnerable to attacks from a particular spoofing type, a CM should prioritise detecting those spoofing types against other, easier attacks. However, there are other applications of CM systems, so for this reason, we will only use spoof detection EERs in this paper.

3.4. Noise analysis

To assess the robustness of the systems to noise, we added several types of realistic noise to the original ASVspoof2019 audio files using MUSAN [23] and RIR [2] datasets. A subset of the development and the evaluation sets was selected for this purpose; from each set we extracted a balanced number of bonafide and spoof samples, and sampled evenly across the spoofing types. Samples in the subset were additionally balanced across speakers, with all 10 development speakers and a random set of 10 evaluation speakers represented evenly. The resulting subset contained 380 files (190 bonafide, 190 spoof). Note that no noised data was incorporated into system training. The following types of noise were added to the subset:

- **Reverberation**: the original files were convolved with random selections of the simulated room impulses from the RIR database
- **Speech**: audio samples from 3 to 6 speakers were randomly selected from the MUSAN data set and mixed into the original file at a signal to noise ratio (SNR) of 5 dB.
- **Music**: a music file was randomly selected from the MUSAN data set and then mixed into the original file at an SNR of 5dB.
- **Noise**: a noise file was randomly selected from the MUSAN data set and then mixed into the original file at an SNR of 5dB.
- **Pink**: a randomly generated pink noise sample was added to each original sample at an SNR of 10 dB.

[2] http://openslr.org/28

3.5. Explainable audio

Several approaches to explain classification decisions with audio examples were considered in a pilot study; we found the following two the most promising:

GradCAM-Binary map: The first approach is based on asking what part of the spectrogram the neural network focuses on to make its classification decision for a given audio sample. We use the GradCAM technique [13] to obtain a saliency map for the audio sample, using a publicly available GradCAM library [24]. The saliency map shows which parts of the spectrogram are the most sensitive to the class activation decision. In other words, this shows which parts are the most important. This saliency map can be used to threshold the spectrogram for its salient parts, as it is just a "2D array of importance". Finally, the new spectrogram can be resynthesised to generate audio using a Griffin-Lim vocoder [25]. This process is visualised in Figure 3.

Mean audio: Another way to emphasise the different acoustic cues the neural network fits on is to generate a 'mean' audio sample. A limiting factor when listening to individual audio samples (to assess naturalness, for example), is that our brains inevitably focus on the semantic content instead of any acoustic anomalies. By playing back multiple audio samples simultaneously, we can simulate a cocktail party scenario, where the listener is forced to listen to the acoustics. In our setup, we created mean audio samples by grouping individual samples based on the CM scores. For each spoof type, we collect the 100 closest files to each side of the CM decision boundary (i.e. bonafide and spoof) and generate a mean audio sample.

3.6. Dimensionality reduction on the the embedding space

In order to get a different perspective on the problematic classes, and the shapes of the transformed probability distributions by the neural network, we visualise the embedding space. This is done with two different methods. First, by performing a principal components analysis (PCA) on the activations of the penultimate layer (the embeddings) and plotting the first two principal components. Then, we use the first 50 principal components to calculate the t-Stochastic Neighbour Embedding (t-SNE) [26]. We then visualise the first two resulting embeddings.

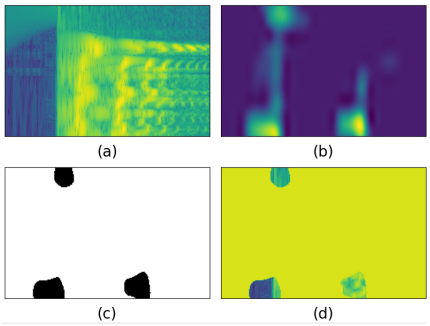

Figure 3: Example showing the process of audio reconstruction. From left to right and top to bottom: (a) CQT-spectrogram (b) GradCAM saliency map, (c) binary threshold on saliency map, (d) map applied on CQT-spectrogram, which is passed to the Griffin-Lim vocoder to get the reconstructed waveform.

3.7. Naturalness calculations

We used a publicly available implementation of MOSNet [9] [27], to estimate the mean opinion scores (MOS) for the naturalness of the utterances. Mean opinion scores are typically determined by multiple raters in a subjective listening test; MOSNet is an automatic system trained on subjective ratings of utterances, which outputs the estimated MOS between 1 (low naturalness) and 5 (high level of naturalness). After that, mean of the utterance-level MOS values were calculated for each spoofing category and a linear regression was performed between the first principal component and the utterance-level MOS values.

4. Results and discussion

In Figure 4 we can see the classwise equal error rates of the CQT and CQT-GMM-DNN architectures. Overall, the performance of these models is better than the baseline LFCC-GMM and CQCC-GMM systems. We note here again that our Python implemented baselines only differ marginally from the published results (see Section 3.2).

Although not the focus of this paper, we observe that fusion of multiple systems has been demonstrated to be effective on the logical access data [1]. To demonstrate the potential of fusion here, we consider a simple sum fusion of the two baseline systems (LFCC-GMM and CQCC-GMM) and the CQT-GMM-DNN system, which results in a large performance improvement.

From Figure 5 we can observe that most spoof examples are perfectly separated from bonafide, but A17-A19 have substantial overlap.

4.1. Audio examples

The generated audio examples can be found on the website[3].

4.2. Voice conversion is challenging to detect

The classwise equal error rate results support that classes A17-A19 are very challenging for the spoofing detectors. The CQT-GMM-DNN system achieves slightly better overall performance, by outperforming the CQT-DNN in difficult cases.

[3] https://karkirowle.github.io/publication/odyssey-2020

4.3. Embeddings show separation and confirm difficulties

The neural network was able to generalise on a variety of examples not present in the training data, which is confirmed by both the projections and the classwise equal error rates. The projections of the embeddings tell us that the distribution of the A17-A19 classes are likely very different from the other spoof classes. The shape and size of this class distribution is very similiar to the bonafide class cluster. This indicates that the neural network would benefit a lot from learning some discriminative features between these classes and the bonafide classes.

The first two principal components explain 22.87% and 15.20% of the variance respectively. Observing that the overlap is larger than the reported classwise EERs on Figure 4, we can conclude that discriminating A17-A19 classes require more than two degrees of freedom.

The t-SNE projections tell a very similar story, the main difference being that other classes' point cloud includes four or five clusters. On the other hand, a striking similarity is that the bonafide and difficult classes are not perfectly overlapping.

4.4. Proposed countermeasures are more robust to realistic noise

As seen in Table 1, the performance of all of the evaluated spoofing countermeasures are affected negatively by noise. The proposed CQT-GMM-DNN and CQT-DNN systems demonstrate more robustness than the baseline systems however. Interestingly, the CQT-GMM-DNN is slightly less robust to noise and it has a better EER in the evaluation set. In addition, we find that the CQT-DNN's performance hardly changes in reverberation. The 10% decrease on EER compared to the LFCC-GMM in both developed architectures, also justifies the usage of CQT. The embedding visualised in Figure 7 shows on the example of reverberation how VC attacks and robustness are related. It shows that performance on VC attacks is related to noise robustness, which is also supported by the evidence in Table 1.

The performance in noise is summarised by the DET (Detection Error Trade-off) curves in Figure 8, which shows the improvement offered by CQT-GMM-DNN and CQT-DNN over the baseline systems in both clean and noisy conditions.

4.5. Naturalness is a key decision factor for CM

Regression of MOS with the first principal axis results in an $R^2 = 0.315$, which indicates that naturalness of signals plays significant role in the outcome. However, the influence is counterintuitive as it is illustrated by Table 4. Surprisingly, the framework deems bonafide signals less natural than several spoof types, and we can see that the most challenging classes are actually suffering from low naturalness. This is in agreement with the results in [8], where in general, classes that were more challenging to subjective listeners, were not necessarily those which were more challenging for the neural networks.

4.6. Minor acoustic cues in explainable audio

The explainable audio gives a qualitative idea what kind of acoustic cues the neural network fit on. The GradCAM-Binary examples show buzziness and rhythm of speech to be the determining factors, which are also the most apparent differences between generated and natural speech.

The mean audio examples give a better idea of what acoustic cues are present in these signals. Especially in class A19, there is a very audible noise. Spoof audio examples in general

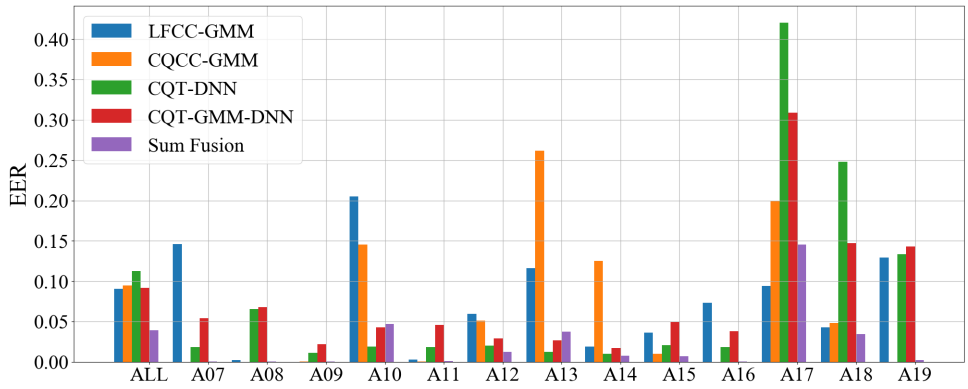

Figure 4: EERs on the ASVspoof2019 evaluation set for the two proposed CQT networks, the two baseline systems, and a score fusion. The class labels on the x-axis indicate the official spoof type category labels.

Model	Reverb	Speech	Music	Noise	Pink	All noise	Noiseless
LFCC-GMM	21.05	38.94	35.79	38.42	**23.68**	41.37	5.78
CQCC-GMM	16.31	28.42	31.05	40.00	24.74	36.32	8.42
CQT-DNN	**10.52**	**20.52**	22.10	**26.84**	24.74	**21.68**	5.26
CQT-GMM-DNN	13.15	28.42	**21.57**	27.89	30.00	23.68	**4.21**
Sum Fusion	**8.42**	23.15	**20.52**	32.11	**21.05**	28.95	**2.63**

Table 1: EERs (%) of the classifiers exposed to different kinds of realistic noise scenarios, and to the combined set of all noisy files ('All noise'). The best performing variant of each condition is emphasised with a **bold** typeface, with fusion considered separately.

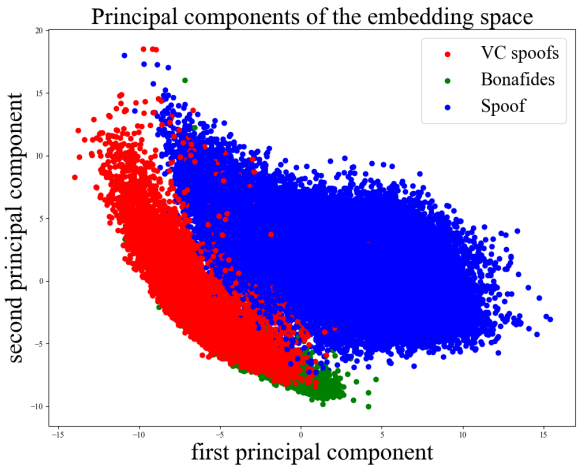

Figure 5: The projected principal components of the embeddings learnt by the neural network. Each dot corresponds to an activation of an example from the evaluation set. The activations are linearly transformed in such a way to maximise the spread of the dots.

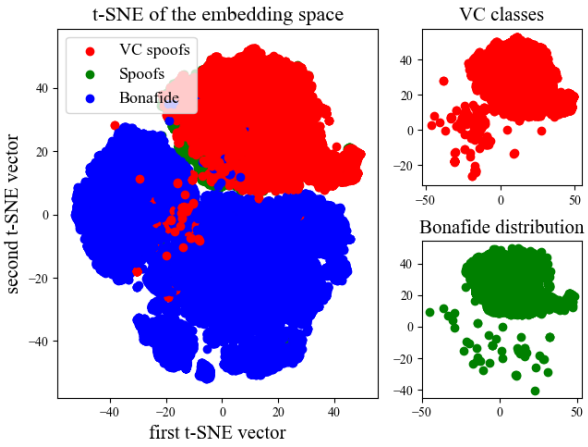

Figure 6: The figure shows the t-SNE projections in a similar fashion as Figure 5. On the right, the difficult classes and the bonafide classes distribution are shown too.

A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	Bonafide
3.60	3.56	3.10	3.69	3.58	3.52	2.80	3.54	3.74	3.31	2.62	2.57	2.75	3.02

Table 2: Mean of predicted mean opinion scores by MOSNet.

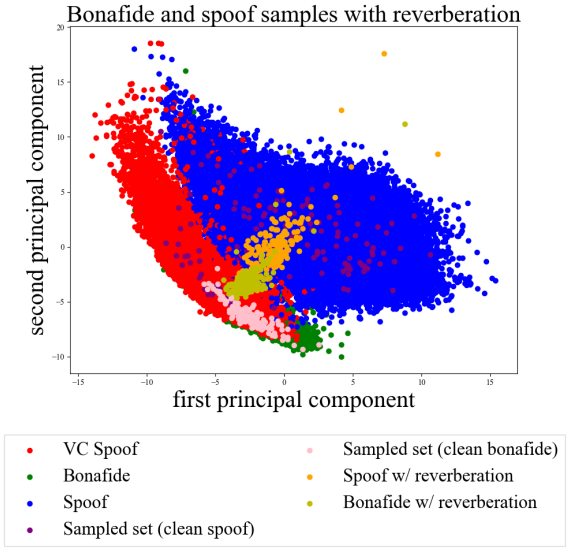

Figure 7: The figure shows the internal PCA representation for the samples which were noised with reverberation.

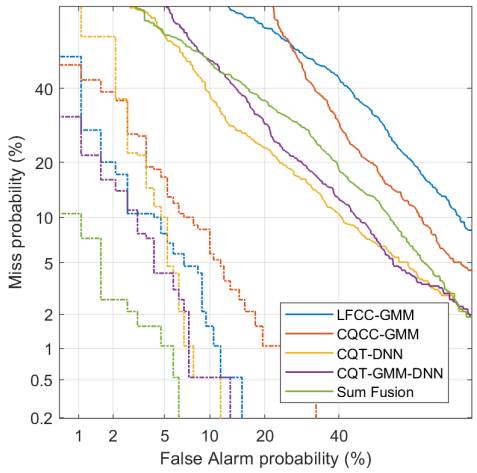

Figure 8: DET curves for each system for all noisy recordings (solid lines) and all clean recordings (dashed lines).

seem to have a more rapid, louder onset of speech, which is readily detected by our spoofing countermeasure.

4.7. Limitations of CQT/FFT based explainable audio

Because phase information is neglected both in constant Q-transform and Fast Fourier Transform (FFT), the synthesised explainable audio quality has some limitations. Our experience is that in the case of FFT based audio synthesis (not used in this paper) this is not a problem, while it certainly affects reconstruction quality based on CQT features.

4.8. Voice activity detection

Previous studies [2, 28] have used voice activity detection (VAD) to focus on the speech regions of the audio signal for spoofing detection; here we have not applied VAD, as we believe that some useful artefacts are present in the non-speech regions of the signal, particularly in LA samples. However, we acknowledge that the application of VAD is worthy of future investigation.

5. Conclusion

In this paper we have proposed a novel combined CQT-ResNet and GMM architecture for spoofing detection. We have demonstrated it to perform effectively relative to the baseline ASVspoof2019 countermeasures and have shown that it can be fused with these systems to further improve performance. We have shown that neural network based spoofing countermeasures are generally more robust to noise than their GMM-based baseline counterparts. Investigation using explainable audio techniques enabled us to tap into "*the black box of neural networks*" in order to understand their behaviour by listening to audio examples. Finally, we have shown that the embeddings of the CQT-ResNet significantly correlate with an objective naturalness function, providing evidence that it statistically models the perceptual quality of utterances.

6. Acknowledgements

The first author would like to thank Oxford Wave Research Ltd for the internship opportunity and Mathew-Magimai Doss for the helpful discussions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

7. References

[1] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and

Kong Aik Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Interspeech 2019*, ISCA, sep 2019, pp. 1008–1012, ISCA.

[2] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov, "STC Antispoofing Systems for the ASVspoof2019 Challenge," in *Proc. Interspeech 2019*, 2019, pp. 1033–1037.

[3] Weicheng Cai, Haiwei Wu, Danwei Cai, and Ming Li, "The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion," in *Proc. Interspeech 2019*, 2019, pp. 1023–1027.

[4] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser, "Dilated residual networks," *CoRR*, vol. abs/1705.09914, 2017.

[5] "ASVspoof 2019 Interspeech Evaluation Session," `https://www.asvspoof.org/interspeech2019\_slides.pdf`, Accessed: 2020-01-08.

[6] Christian Schörkhuber and Anssi Klapuri, "Constant-q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[7] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients.," in *Odyssey*, 2016, vol. 45, pp. 283–290.

[8] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling, "The ASVspoof 2019 database," 2019.

[9] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning based Objective Assessment for Voice Conversion," in *Interspeech*, ISCA, 2019, pp. 1542–1545.

[10] Cemal Hanilci, Tomi Kinnunen, Md Sahidullah, and Aleksandr Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83–97, 2016.

[11] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "Spoofing detection under noisy conditions: a preliminary investigation and an initial database," *arXiv preprint arXiv:1602.02950*, 2016.

[12] Hong Yu, Achintya Sarkar, Dennis Alexander Lehmann Thomsen, Zheng-Hua Tan, Zhanyu Ma, and Jun Guo, "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*. IEEE, 2016, pp. 1–5.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[14] Brian McFee, Vincent Lostanlen, Matt McVicar, Alexandros Metsai, Stefan Balke, Carl Thomé, Colin Raffel, Dana Lee, Frank Zalkow, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Ryuichi Yamamoto, Eric Battenberg, Rachel Bittner, Keunwoo Choi, Josh Moore, Ziyao Wei, Scott Seyfarth, nullmightybofo, Pius Friesch, Fabian-Robert Stöter, Darío Hereñú, Thassilo, Taewoon Kim, Matt Vollrath, Adam Weiss, and Adam Weiss, "librosa/librosa: 0.7.1," Oct. 2019.

[15] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *Interspeech 2019*, ISCA, sep 2019, pp. 1013–1017, ISCA.

[16] Christian Perone, Evan Calabrese, and Julien Cohen-Adad, "Spinal cord gray matter segmentation using deep dilated convolutions," *Scientific Reports*, vol. 8, 10 2017.

[17] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.03499, 2016.

[18] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP 2018*, 2018.

[20] Diederik P Kingma and Jimmy Lei Ba, "Adam: A method for stochastic gradient descent," *ICLR: International Conference on Learning Representations*, 2015.

[21] François Chollet, "keras," `https://github.com/fchollet/keras`, 2015.

[22] Tomi Kinnunen, Kong Aik Lee, Hector Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," 2018.

[23] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015.

[24] "keras-vis library," `https://github.com/raghakot/keras-vis`, Accessed: 2020-01-31.

[25] Daniel W. Griffin and Jae S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[26] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] "MOSNet implementation," `https://github.com/lochenchou/MOSNet`, Accessed: 2020-01-31.

[28] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 335–341.