Improving the Robustness of Deepfake Audio Detection through Confidence Calibration

Yuxiang Zhang[1,2,†], Jingze Lu[1,2,†], Zhuo Li[1,2], Zengqiang Shang[1,2], Wenchao Wang[1,*] and Pengyuan Zhang[1,2,*]

[1] *Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Haidian District Beijing, 100190, China*
[2] *University of Chinese Academy of Sciences, No.1 Yanqihu East Rd, Huairou District, Beijing, 101408, China*

Abstract

The issue of overconfidence in out-of-distribution data in current deepfake audio detection models has resulted in poor robustness, making it difficult for models to achieve good results in the test set with large distribution differences, such as Audio Deepfake Detection Challenge (ADD 2023) Track 1.2 audio fake game detection task. In this paper, the Energy-based Open-World Softmax (EOW-Softmax) is introduced to calibrate model confidence and achieves good results in the challenge. Additionally, the paper presents a range of data augmentation methods, including vocoder-generated training data, to effectively improve the performance of the deepfake audio detection models. By fusing the scores of a variety of single systems based on Squeeze-and-Excitation residual neural network (SENet), light convolutional neural network (LCNN) and Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks (AASIST), the proposed system achieves the second place in the challenge.

Keywords

fake detection, deepfake audio, ADD Challenge, confidence calibration

1. Introduction

In recent decades, the progress of deep learning and artificial intelligence generated content (AIGC) has led to significant advancements in speech synthesis and voice conversion technologies. Deep learning models have the ability to generate incredibly realistic audio. However, the malicious use of deepfake has also raised concerns. As a result, the detection of deepfake audio has become an important topic of interest. Four ASVspoof Challenges [1, 2, 3, 4] have contributed to the development of countermeasures (CMs) against deepfake audio. However, ASVspoof only has English datasets and lacks a focus on Chinese spoof speech detection. The Audio Deep Synthesis Detection Challenge (ADD 2022) [5] is a challenge for Chinese deepfake audio detection and proposes new tracks such as noisy scenes and fake game. The second Audio Deepfake Detection Challenge (ADD 2023) [6] has been launched to encourage researchers to develop innovative technologies that can help identify and analyze deepfake audio. ADD 2023 is distinct from previous challenges as it aims to go beyond binary classification. Track 1 of ADD 2023, which is comprised of two tasks: the audio generation task and the fake audio detection task. For the Track 1.1 generation task (FG-G), participants aim to create fake audio that can deceive the fake detection model of Track 1.2. The Track 1.2 detection task (FG-D) requires participants to identify fake utterances, with a particular focus on identifying fake samples generated in Track 1.1. Both tasks have two rounds of evaluations. Track 2 focuses on localizing manipulated intervals within a partially fake speech. While Track 3 intends to identify the algorithms that generate the deepfake audio.

According to the description of the Challenge, the evaluation set contains not only fake audio generated by the organizers but also fake audio generated by participants in Track 1.1. In addition, there is noise interference in the test set. However, the training and development datasets are easily discriminated. Therefore, improving the robustness in the face of unknown deepfake algorithms and complex noise is an important issue in ADD 2023 Track 1.2. Some work has explored the reasons for the poor robustness of spoof speech detection algorithms. One of the possible causes is that the CMs have high but incorrect confidence in unknown algorithms [7, 8]. To address this issue, Energy-based Open-World Softmax (EOW Softmax) based confidence calibration is introduced to improve robustness in Track 1.2.

In addition to the confidence calibration, the CMs based on light convolutional neural network (LCNN) [9], Squeeze-and-Excitation residual neural network

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R

[*] Corresponding author.

[†] These authors contributed equally.

zhangyuxiang@hccl.ioa.ac.cn (Y. Zhang); lujingze@hccl.ioa.ac.cn (J. Lu); li1_zhuo@foxmail.com (Z. Li); shangzengqiang@hccl.ioa.ac.cn (Z. Shang); wangwenchao@hccl.ioa.ac.cn (W. Wang); zhangpengyuan@hccl.ioa.ac.cn (P. Zhang)
0000-0003-4624-5663 (Y. Zhang)

(SENet) [10] and Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks (AASIST) [11] are proposed. Loss function including cross-entropy loss, center loss [12] and Angular Softmax (A-Softmax) Loss [13]. Different features, including short-time Fourier transform (STFT) spectrograms and raw wave are used as front-ends. Various data augmentation methods including vocoder data creation [14], RawBoost [15] and noise addition effectively improve the robustness of the deepfake audio detection systems. Finally, score fusion further improves performance.

2. Methods

In this section, data augmentation methods and confidence calibration is presented. In both rounds, large performance gains are achieved by creating training data with vocoders. In the second round, the confidence calibration further improve the performence effectively.

2.1. Data Augmentation

Data augmentation is a common training method that can be effective in improving model robustness in the absence of sufficient training data. Three data augmentation methods are used to enhance the performance.

There is obvious noise and reverberation in the speech of the evaluation sets. Therefore, the most commonly used methods for speech data angmentation, noise and reverberation addition from the MUSAN [16] and RIRs [17] datasets are performed online in a similar way to Kaldi [18] during training.

RawBoost [15] is a method of data boosting and augmentation that operates directly on the raw waveform and has achieved good results for complex channels in speech anti-spoofing. Rawboost does not require additional data. Through a combination of various filters, RawBoost models nuisance variability originating from coding, transmission, microphones and amplifiers as well as linear and non-linear distortion. All methods of RawBoost data augmentation[1] are applied in model training.

In order to enhance the robustness of CMs to unknown deepfake algorithms, data augmentation can be performed by doing copy-synthesis on genuine speech through various vocoders [14]. Four statistical parametric speech synthesis vocoders: WORLD [19], STRAIGHT [20, 21], Griffin_lim [22] and HMPD [23] are used to create spoof audio. All genuine speech in the training and development sets is transformed into spoof speech by these four vocoders.

2.2. Confidence Calibration

Discriminators based on deep neural networks often suffer from overfitting problems, producing overconfident predictions [24]. In the speech anti-spoofing task faced with a large amount of out-of-distribution (OOD) data, overconfident false predictions seriously affect the performance [7, 8]. Confidence calibration is therefore important for the reliability of decisions made by deep learning based CMs. To address the issue of overconfidence, Energy-based Open-World Softmax (EOW Softmax) [25] is introduced into deepfake audio detection, which models open-world uncertainty as an additional dimension via a $K + 1$-way softmax formulation. The extra dimension can be forced to be negatively correlated with the marginal data distribution through an energy-based objective function. In this way, confidence is automatically calibrated to reduce confidence in inputs that fall beyond the distribution of the training data.

The purpose of the energy function E_θ is to map a D-dimensional data point to a scalar. With E_θ it is possible to assign low energy to the observed variable configuration and high energy to the unobserved variable. The probability density $p(x)$ for $x \in \mathbb{R}^D$in an energy based model (EBM) can be written as

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)} \tag{1}$$

where $Z(\theta) = \int_x \exp(-E_\theta(x))$ represents the normalizing constant, also known as the partition function

In order to make the $K + 1$-th score to represent open-world uncertainty, the classifier should be able to generate high uncertainty scores for anomalous inputs, which in turn reduces the confidence in the prediction of the original K categories. Let $f_\theta : \mathbb{R}^D \to \mathbb{R}^{K+1}$ be the neural network that generates $K + 1$ logits, and $f_\theta(x)[i]$ denotes the i-th logit given x, with $i \in \{1, \ldots, K, K + 1\}$. The probabilities of the outputs can be obtained by:

$$h_\theta(x)[i] = \frac{\exp(f_\theta(x)[i])}{\sum_{j=1}^{K+1} \exp(f_\theta(x)[j])} \tag{2}$$

where h_θ is the concatenation of network f_θ and the softmax normalization layer.

To allow $h_\theta(x)[K+1]$ to encode uncertainty, the score of $h_\theta(x)[K+1]$ needs to be correlated with the marginal data distribution. When the input comes from within the training data distribution $p(x)$, the model should be confident in its decisions. Therefore, $h_\theta(x)[K + 1]$ should be low, while $\sum_{i=1}^{K} h_\theta(x)[i]$ should be high. If the input features deviate from the distribution of the training data, the model should remain uncertain about its decision. As a result, $h_\theta(x)[K + 1]$ should be high to indicate greater uncertainty, which naturally yields to low $\sum_{i=1}^{K} h_\theta(x)[i]$. By designing learning objectives

[1] https://github.com/TakHemlata/RawBoost-antispoofing

through EBM, $h_\theta(x)[K+1]$ can capture the marginal distribution $p(x)$.

Firstly, define the energy function as

$$E_\theta(x) = \log h_\theta(x)[K+1] \tag{3}$$

Then, the objective function is defined as

$$\begin{aligned}\min_\theta & \mathbb{E}_{p(x)}\left[-\log h_\theta(x)[y]\right] \\ & + \lambda \mathbb{E}_{p_{\bar{\theta}}(x)}\left[-\log h_\theta(x)[K+1]\right]\end{aligned} \tag{4}$$

where $\lambda > 0$ is a hyper-parameter. The first term is the maximum log-likelihood objective for the K-way classification task using the true label y; the second term can be viewed as the maximum log-likelihood objective for identifying data sampled from $p_{\bar{\theta}}(x)$. $p_{\bar{\theta}}(x)$ denotes the model distribution of the frozen parameters in this iteration. Optimizing Eq.4 can make the sum of the K softmax scores of the original class proportional to the marginal density $p(x)$, which in turn makes the $K+1$-th softmax score negatively correlated with $p(x)$.

3. Experiments

This section introduces the datasets and evaluation metrics in ADD 2023 Track 1.2. And a detailed system description is presented. Two systems are implemented in the first round and four are applied in the second round.

3.1. Dataset and Evaluation Metrics

The ADD 2023 Challenge datasets are utilized in all experiments. There are 27,084 pieces of audio in the training set, 3,012 of which are genuine. And the development set has 28,324 pieces of audio, of which 2,307 are genuine. Solely the training set is employed to train all systems, while the development set is used for performance validation during training. All data augmentation methods described in Section 2.1 are utilized. 21,276 pieces of fake audio are generated through four vocoders.

The objective of Track 1.2 is to develop an algorithm or method capable of distinguishing between genuine and deepfake audio. The weighted equal error rate (WEER) serves as the evaluation metric for this task. Specifically, WEER is defined as follows[2]:

$$WEER = \alpha EER_{R1} + \beta EER_{R2}$$

where $\alpha = 0.4$ and $\beta = 0.6$ represents the weights for the two equal error rates (EERs) obtained in round 1 and round 2 of Track 1.2, respectively.

3.2. Model and Training strategy

The main acoustic features used in our experiments are STFT log power magnitude spectrograms. The input features are extracted with Blackman window of 400 frame length and 512 Fast Fourier Transform (FFT) points. Only low-frequency part ($0-4$ kHz) of spectrograms are fed into neural networks [26].

The SENet is a combination of residual neural network (ResNet) with SE block [10], which is one of the commonly used classifiers for spoof speech detection. The SENet implemented here is SENet34, and the number of channels is (16, 32, 64, 128). SENet can assign weights to the features of different channels through attention mechanisms, thereby enhancing features that are more important for fake audio detection. The A-softmax is used as the loss function with $m = 4$. The features are extracted with torchaudio [27].

The end-to-end anti-spoofing CM AASIST[3] is the same as [11], where the encoder is based on six residual blocks. The backend is based on graph attention layer and graph pooling layer. The loss function is cross-entropy loss with weights of (0.1, 0.9) for spoof and bonafide class.

The features fed in the STFT-LCNN system are extracted with librosa [28], and the loss function is cross-entropy loss and center loss. The factor of center loss combined with the cross-entropy loss is 0.05. The proposed system based on LCNN is similar to the LCNN baseline system [29] for the ASVspoof 2021 Challenge and the system presented in the first place of ASVspoof 2021 [9]. The LCNN model comprises of nine convolutional layers and two bidirectional long short-term memory (BLSTM) layers. The architecture of the LCNN model is characterized by the use of Max-Feature-Map (MFM) [30] activation, which is based on the Max-Out activation function. This activation function allows for the selection of critical features, reduces the number of parameters, and improves the robustness of model. The features are extracted with convolutional layers and subsequently fed into BLSTM layers for sequential pooling. The inputs and outputs of BLSTMs are then averaged and summed in the time domain, with the resulting mean vector then reduced to a 128-dimensional embedding using fully connected layers. The 128-dimensional embeddings are used to derive the center loss.

The wide ResNet used for EOW Softmax [4] is modified based on the [25], with depth of 22 and wide factor of 2.

All systems are optimized with Adam optimizer. The Adam optimizer is adopted with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and weigth decay 10^{-4}. The learning rate is initialized as 0.0003. For ResNet based systems, the learning rate increases linearly for the first 1000 warm-up

[2] http://addchallenge.cn/add2023

[3] https://github.com/TakHemlata/aasist

[4] https://github.com/BIGKnight/Energy-Based-Open-World-\ Uncertainty-Modeling-for-Confidence-Calibration

steps and then decreases proportionally to the inverse square root of the step number. StepLR is used as scheduler for LCNN with step size of 10 epochs and coefficient 0.5. Adam optimizer with learning rate of 10^{-4} and cosine annealing learning rate decay are utilized in AASIST. All models are trained with 100 epochs, and the model with the lowest loss on the development set is selected as the final model for evaluation.

In the first round, two systems, STFT-SENet and AASIST, are implemented and the scores are fused with weights of 0.75 and 0.25 . In the second round, four systems of STFT-SENet, AASIST, STFT-LCNN and STFT-ResNet-EOW are implemented, and the score fusion weights are 0.06375, 0.02125, 0.732, 0.183. The fusion weights are manually selected based on the performance.

3.3. Results

3.3.1. Results of data augmentation

In the first round, LCNN and ResNet are tested as baseline systems in different configurations. The experimental results of data augmentation on the first round of ADD 2023 Track 1.2 evaluation set are shown in Table 1.

Table 1
EER% comparison of data augmentation experimental results in the first round evaluation.

Augmentation	LCNN	SENet
Noise & Reverb	38.45	57.83
RawBoost	57.78	51.88
+ Vocoder	27.31	**23.29**

The effects of two data augmentation methods vary significantly. Rawboost exhibits minimal impact on the evaluation set in both systems. In contrast, the addition of noise and reverb online is more effective for LCNN, but less effective than RawBoost for SENet. A plausible explanation for this discrepancy can be attributed to the pooling layers and MFM operations in LCNN, which may result in the model paying less attention to input feature details than SENet. Additionally, the presence of noise in the test set, which differs from the noise used for data augmentation, may cause performance degradation of SENet, while LCNN may benefit from it.

After identifying the most effective data augmentation approach for each system, the fake audio generated by vocoders is incorporated into the training process. The use of vocoders to generate training data proved to be highly effective for both systems, resulting in a significant relative reduction in the EER of LCNN and SENet by 29% and 53%, respectively. This suggests that when faced with limited training data, transforming genuine audio into fake audio using vocoders can be an extremely effective data augmentation technique.

3.3.2. Results of EOW Softmax

EOW-Softmax is introduced in the second round to reduce the confidence of the model and alleviate the overfitting problem. Compared to vanilla trained LCNN and SENet, EOW-Softmax demonstrates great performance on the evaluation set of round 2 with a significant amount of OOD data. But in the second round EER is slightly higher than LCNN. The effectiveness of this approach is also validated on the evaluation set from the first round after the challenge. The results are presented in Tabel 2.

Table 2
EER% comparison between EOW-Softmax and other systems in ADD 2023 track1.2.

System	1st round	2nd round
SENet	23.29	23.24
LCNN	27.31	**17.69**
EOW-Softmax	**18.60**	19.34

A comparison of score distribution histograms, as depicted in Figure 1, reveals that EOW-Softmax successfully calibrates the confidence of model predictions. The scores obtained by the SENet are more extreme than those obtained using EOW-Softmax. However, the scores obtained by EOW-Softmax seem to be difficult to clearly distinguish between genuine and fake. And the effect of EOW-Softamx on other models is still worth exploring.

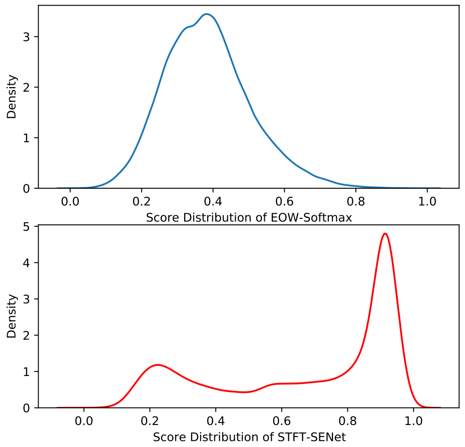

Figure 1: Comparision of score histograms in the first round between EOW-Softmax and STFT-SENet. The difference in genuine and fake is not reflected due to the lack of labels.

3.3.3. Results of submitted systems

Table 3 shows the performance of the submitted systems for ADD 2023 Track 1.2. Compared with the best single systems, the score fusion reduces the EER by about 2%.

Table 3
EER% of final submitted systems for Track 1.2 in ADD 2023.

System	1st round		2nd round	
	Weight	EER	Weight	EER
AASIST	0.25	38.45	0.02	35.00
SENet	0.75	23.29	0.06	23.24
LCNN	-	-	0.73	17.69
EOW-Softmax	-	-	0.18	19.34
Fusion	-	21.11	-	15.82

Table 4 presents the EERs of the top 5 performing systems in ADD 2023 Track 1.2. While there is a significant gap between the first and second place, the differences between the remaining teams are relatively small. Notably, with the exception of the first place team, the EERs of all other teams are lower in the second round than in the first round. Our final submission achieves the first runner-up in ADD 2023 Track 1.2.

Table 4
EER% of the top-performancing systems in track1.2.

ID	EER_{R1}	EER_{R2}	WEER
B01	11.56	13.05	12.45
B02 (ours)	21.11	15.82	17.93
B03	23.44	21.26	22.13
B04	23.51	21.75	22.45
B05	24.06	22.59	23.17

4. Conclusion

This paper describes the system developed for ADD 2023 Track 1.2. Several single systems that effectively leverage various data augmentation methods to achieve strong performance are presented. Of particular note is the introduction of EOW-Softmax, which addresses the challenge of deepfake audio detection systems exhibiting overconfidence in OOD data. The EOW-Softmax based system successfully calibrates confidence, mitigates overfitting, and improves overall robustness. The final system submitted achieves the second place in the challenge.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2021YFC3320103).

References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Proc. Interspeech 2015, 2015, pp. 2037–2041.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K. A. Lee, The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection, in: Proc. Interspeech 2017, 2017, pp. 2–6.

[3] M. Todisco, X. Wang, V. Vestman, et al., Asvspoof 2019: Future horizons in spoofed and fake audio detection, in: Proc. Interspeech 2019, 2019, pp. 1008–1012.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado, ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 47–54. doi:10.21437/ASVSPOOF.2021-8.

[5] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, et al., Add 2022: the first audio deep synthesis detection challenge, in: Proc. ICASSP 2022, IEEE, 2022, pp. 9216–9220.

[6] J. Yi, J. Tao, R. Fu, X. Yan, T. Wang, Chenglong ang Wang, C. Y. Zhang, X. Zhang, Y. Zhao, Y. Ren, L. Xu, J. Zhou, H. Gu, Z. Wen, S. Liang, Z. Lian, H. Li, Add 2023: the second audio deepfake detection challenge, in: accepted by IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), 2023.

[7] X. Wang, J. Yamagishi, Estimating the confidence of speech spoofing countermeasure, in: Proc. ICASSP 2022, IEEE, 2022, pp. 6372–6376.

[8] Y. Zhang, J. Lu, X. Wang, Z. Li, R. Xiao, W. Wang, M. Li, P. Zhang, Deepfake detection system for the add challenge track 3.2 based on score fusion, in: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, DDAM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 43–52. URL: https://doi.org/10.1145/3552466.3556528. doi:10.1145/3552466.3556528.

[9] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, G. Lavrentyeva, STC Antispoofing Systems for the ASVspoof2021 Challenge, in: Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 61–67. doi:10.21437/ASVSPOOF.2021-10.

[10] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation net-

works, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[11] J.-w. Jung, H.-S. Heo, H. Tak, et al., Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: Proc. ICASSP 2022, 2022, pp. 6367–6371. doi:10.1109/ICASSP43922.2022.9747766.

[12] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on computer vision, Springer, 2016, pp. 499–515.

[13] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[14] X. Wang, J. Yamagishi, Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders, in: Proc. ICASSP 2023, 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10094779.

[15] H. Tak, M. Kamble, J. Patino, M. Todisco, N. Evans, Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing, in: Proc. ICASSP 2022, IEEE, 2022, pp. 6382–6386.

[16] D. Snyder, G. Chen, D. Povey, MUSAN: A Music, Speech, and Noise Corpus, 2015. arXiv:1510.08484, arXiv:1510.08484v1.

[17] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, S. Khudanpur, A study on data augmentation of reverberant speech for robust speech recognition, in: Proc. ICASSP 2017, IEEE, 2017, pp. 5220–5224.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011.

[19] M. Morise, F. Yokomori, K. Ozawa, World: A vocoder-based high-quality speech synthesis system for real-time applications, IEICE Transactions on Information and Systems E99D (2016) 1877–1884. doi:10.1587/transinf.2015EDP7457, publisher Copyright: © 2016 The Institute of Electronics, Information and Communication Engineers.

[20] H. Kawahara, Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited, in: Proc. ICASSP 1997, volume 2, 1997, pp. 1303–1306 vol.2. doi:10.1109/ICASSP.1997.596185.

[21] H. Kawahara, Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds, Acoustical Science and Technology 27 (2006) 349–353. doi:10.1250/ast.27.349.

[22] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (1984) 236–243. doi:10.1109/TASSP.1984.1164317.

[23] G. Degottex, D. Erro, A uniform phase representation for the harmonic model in speech synthesis applications, EURASIP Journal on Audio, Speech, and Music Processing 2014 (2014) 1–16. doi:10.1186/s13636-014-0038-1, gilles Degottex and Daniel Erro are equal contributors.

[24] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

[25] Y. Wang, B. Li, T. Che, K. Zhou, Z. Liu, D. Li, Energy-based open-world uncertainty modeling for confidence calibration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9302–9311.

[26] Y. Zhang, W. Wang, P. Zhang, The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System, in: Proc. Interspeech 2021, 2021, pp. 4279–4283. doi:10.21437/Interspeech.2021-1281.

[27] Y.-Y. Yang, M. Hira, Z. Ni, A. Astafurov, C. Chen, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Hwang, J. Chen, P. Goldsborough, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, Torchaudio: Building blocks for audio and speech processing, in: Proc. ICASSP 2022, 2022, pp. 6982–6986. doi:10.1109/ICASSP43922.2022.9747236.

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, volume 8, Citeseer, 2015, pp. 18–25.

[29] X. Wang, J. Yamagishi, A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection, in: Proc. Interspeech 2021, 2021, pp. 4259–4263. doi:10.21437/Interspeech.2021-702.

[30] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: International conference on machine learning, PMLR, 2013, pp. 1319–1327.