VIPER
Video and Image Processing Laboratory

ASSD: Synthetic Speech Detection in the AAC Compressed Domain

Amit Kumar Singh Yadav[1], Ziyue Xiang[1], Emily R. Bartusiak[1], Paolo Bestagini[2], Stefano Tubaro[2] and Edward J. Delp[1]

[1]Video and Image Processing Laboratory (*VIPER),*
School of Electrical and Computer Engineering , Purdue University, USA

[2]Image and Sound Processing Lab (*ISPL*),
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

Acknowledgements

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL or the U.S. Government

Motivation

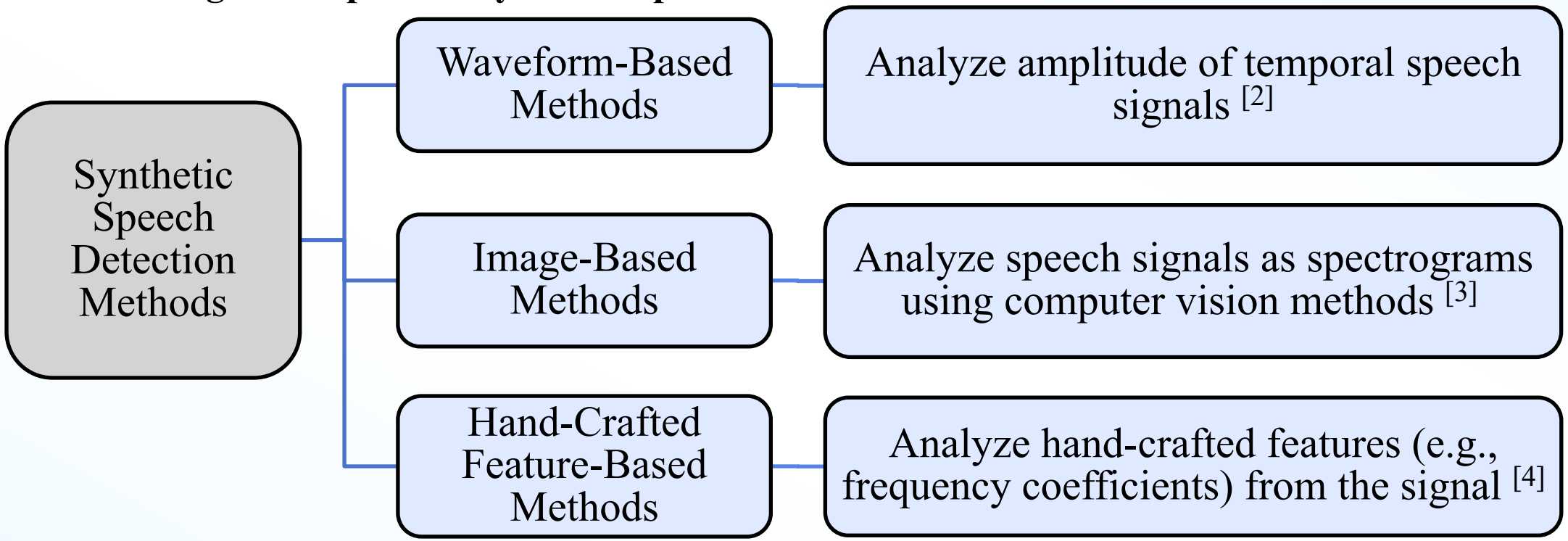
- With recent developments in deep-learning, one can generate high quality, semantically consistent speech
  - perceptually indistinguishable from bona fide speech (recorded by human)

- Boon for application in entertainment industry and voice-based applications
- However, several incidents report misuse of such high-quality synthetic speech
- Synthetic speech is generated using Grad-TTS[1]

[1] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," *Proceedings of the International Conference on Machine Learning*, vol. 139, pp. 8599–8608, July 2021, Virtual. [Online]. Available: http://proceedings.mlr.press/v139/popov21a/popov21a.pdf.

Synthetic Speech Detection and Challenges

Objective: to determine if a given speech signal is bona fide or synthetic? **Existing work is mainly on detecting uncompressed synthetic speech**



[2] G. Hua, *et al.*, "Towards End-to-End Synthetic Speech Detection," *SPL*, vol. 28, June 2021. DOI : 10.1109/LSP.2021.3089437

[3] E. R. Bartusiak, *et al.*, "Synthesized Speech Detection Using Convolutional Transformer-Based Spectrogram Analysis," *ACSSC*, October 2021. DOI : 10.1109/IEEECONF53345.2021.9723142

[4] F. Hassan, *et al*, "Voice Spoofing Countermeasure for Synthetic Speech Detection," *ICAI*, April 2021. DOI : 10.1109/ICAI52203.2021.9445238

AAC Compressed Synthetic Speech Detection

- Speech generated for malicious purposes are often shared on social platforms such as YouTube
- These platforms compress the speech signal using lossy compression standards such as Advance Audio Coding (AAC)[5]
- **Question:** Do existing methods proposed for uncompressed synthetic speech detection work for detecting compressed synthetic speech?

[5] J. Herre and H. Purnhagen, "General Audio Coding," in *The MPEG-4 Book*, F. C. Pereira and T. Ebrahimi, Eds., Upper Saddle River, NJ, USA: Prentice Hall PTR, 2002, pp. 487–544. [Online]. Available: https://sound.media.mit.edu/resources/mpeg4/ audio/general/vancouver-general-audio.pdf .

Experiment 1: Performance on Compressed Speech

We investigated performance of two existing methods:

1. Time-Domain Synthetic Speech Detection Network (TSSDNet)[2]:
   - Processes time domain speech signal using recurrent neural network
   - outperforms all the hand-crafted features-based approaches
   - representative of time domain and hand-crafted feature-based approach
2. Compact Convolutional Transformer (CCT)[3]:
   - converts time domain speech signal to a spectrogram using Short Term Fourier Transform and processes spectrogram using transformer
   - outperforms other machine learning techniques on spectrogram such as KNN, CNN, SVM, Logistic Regression
   - representative for spectrogram-based methods

[2] G. Hua, *et al.*, "Towards End-to-End Synthetic Speech Detection," *SPL*, vol. 28, June 2021. DOI : 10.1109/LSP.2021.3089437
[3] E. R. Bartusiak, *et al.*, "Synthesized Speech Detection Using Convolutional Transformer-Based Spectrogram Analysis," *ACSSC*, October 2021. DOI : 10.1109/IEEECONF53345.2021.9723142

Experiment 1: Dataset

- TSSDNet and CCT have reported performance on the Logical Access (LA) part of ASVspoof2019 dataset[6]

- ASVspoof2019 dataset contains 121,461 bona fide and synthetic speech signals (all uncompressed in FLAC format)
  - split into the training set D_{tr} , validation set D_{dev} , and evaluation set D_{eval} with an approximate ratio of 1:1:3

- There are in total 63.9k synthesized speech samples in D_{eval} , where approximately 61.5k samples are generated from synthetic speech generation methods that do not coincide with D_{dev} or D_{tr}

[6] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *Proceedings of the Interspeech*, pp. 1008–1012, September 2019, Graz, Austria. DOI : 10.21437/Interspeech.2019-2249.

ASVspoof2019 Dataset

- A01-A19: different speech synthesizers
- NN: Neural Network Synthesizer
- WC: Waveform Concatenation Synthesizer
- VC: Vocoder (source modeling)
- A04 is same as A16 but trained on different data
- A06 is same as A19 but trained on different data
- the training set D_{tr} has A01 to A06
- validation set D_{dev} has A01 to A06
- evaluation set D_{eval} has A07 to A19

ASVspoof2019 Dataset					
		D_{tr}	D_{dev}	D_{eval}	Category
Samples	Bona fide	2580	2548	7355	
	Synthetic	22800	22296	63882	
Speakers	Bona fide	20	10	48	
Synthetic Methods	A01	✓	✓	×	NN
	A02	✓	✓	×	VC
	A03	✓	✓	×	VC
	A04 = A16	✓	✓	✓	WC
	A05	✓	✓	×	VC
	A06 = A19	✓	✓	✓	VC
	A07	×	×	✓	NN
	A08	×	×	✓	NN
	A09	×	×	✓	VC
	A10	×	×	✓	NN
	A11	×	×	✓	NN
	A12	×	×	✓	NN
	A13	×	×	✓	NN
	A14	×	×	✓	VC
	A15	×	×	✓	VC
	A17	×	×	✓	VC
	A18	×	×	✓	VC

Experiment 1: Result on Compressed Speech

- We AAC compressed the D_{eval} set at a data rate of 128kbps, typically used by YouTube[7]
- The TSSDNet and CCT were trained on uncompressed training set of ASVspoof19 dataset

Method	Evaluation Dataset	Balanced Accuracy (%)	Accuracy (%)	AUPRC (%)
TSSDNet	Uncompressed D_{eval}	96.27	95.17	99.84
	Compressed D_{eval}	73.82	54.29	97.36
CCT	Uncompressed D_{eval}	94.13	94.98	68.76
	Compressed D_{eval}	64.72	88.62	36.30

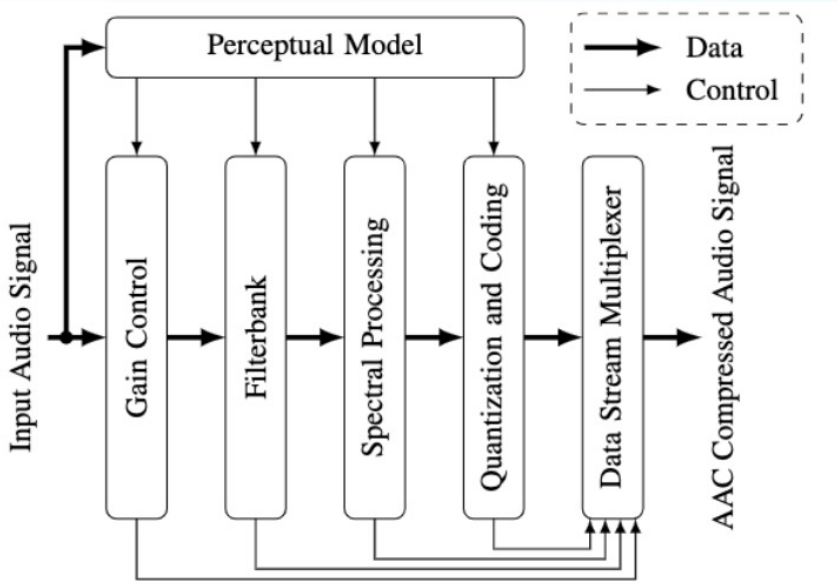
[7] Google Inc., *YouTube Recommended Upload Encoding Settings*, 2022. [Online]. Available: https://support.google.com/youtube/answer/1722171.

Our Proposed Approach: ASSD

- AAC Synthetic Speech Detection (ASSD) is proposed for detecting AAC compressed synthetic speech
- ASSD does not use time-domain speech signal, spectrogram, or hand-crafted features
- ASSD extracts information from the AAC compressed bit stream without decoding the speech signal
- AAC compression is done block-wise, ASSD extract as low as 1000 bits per speech block from the AAC compressed bit stream to detect synthetic speech

AAC Compression

- Advanced Audio Coding (AAC)[5] is an audio compression standard that is the successor of the MPEG-1/2 Audio Layer 3 (MP3) standard
- Used by social platforms such as YouTube and Facebook
- There are many configurations in AAC, known as profiles, and each profile can introduce new coding tools
- Some AAC profiles are AAC-LC, AAC-Main, AAC-SSR, HE-AAC, and HE-AAC v2



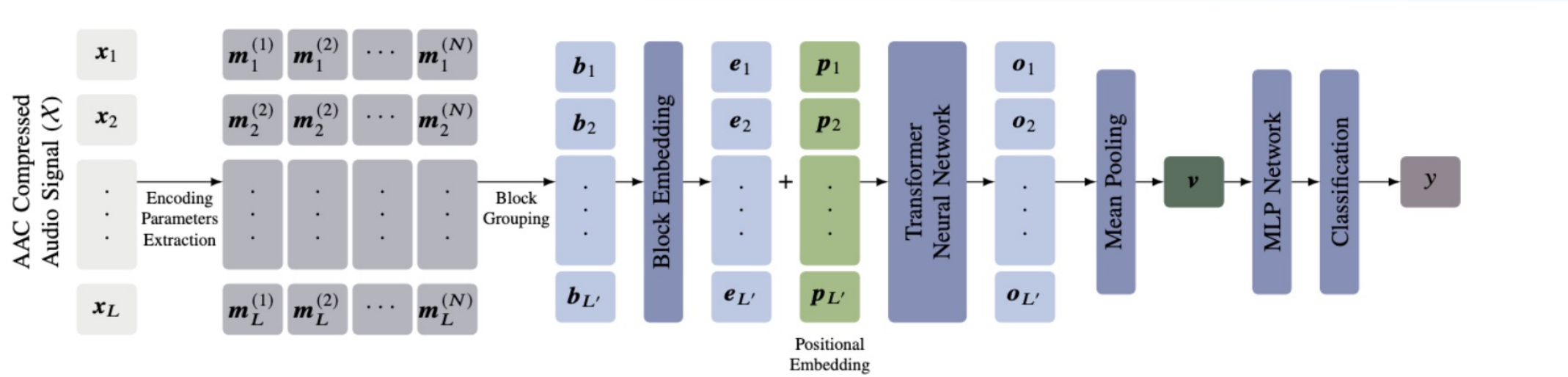
[5] J. Herre and H. Purnhagen, "General Audio Coding," in *The MPEG-4 Book*, F. C. Pereira and T. Ebrahimi, Eds., Upper Saddle River, NJ, USA: Prentice Hall PTR, 2002, pp. 487–544. [Online]. Available: https://sound.media.mit.edu/resources/mpeg4/ audio/general/vancouver-general-audio.pdf .

AAC Compressed Bit Stream

- Our proposed ASSD purposely does not extract information from the compressed bit stream that exists only in a particular profile of AAC
  - For example, Gain Control is not extracted as it is only used in the AAC-SSR profile
- We extract information related to location of block, spectral coefficients, scale factors, sampling rate
- Scale factor
  - group spectral coefficients
  - before quantization
- These are general features:
  - found in several standards

Parameter	Description
`frame_indx`	Temporal location of the block
`WindowSequence`	Windowing Scheme (WS)
`pSpec`	Contains MDCT coefficients
`L`	Number of MDCT coefficients in each window
`GlobalGain`, `pScaleFactor`	SF info (size 41 or 61)
`pCodeBook`	Huffman table selection info
`aacdec_sample_rate`	Sampling rate

Architecture of ASSD



- X = AAC compressed bit stream, consist of L blocks i.e., $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_L}\}$
- From each compressed block $\boldsymbol{x_i}$, extract N-dimensional information $\mathbf{m}_i \in \mathrm{R}^N$
- Group information from several blocks to get $\boldsymbol{b_i}$
- Convert each $\boldsymbol{b_i}$ to vector representation $\boldsymbol{e_i}$, include position representation $\boldsymbol{p_i}$ and process using a transformer encoder to obtain $\boldsymbol{o_i}$
- Mean of all $\boldsymbol{o_i}$ to obtain $\boldsymbol{v}$, assign a detection label y to each speech signal

Ablation Study

- Train on AAC compressed ASVspoof2019 training set D_{tr} at 128kbps using HE profile
- Training set is unbalanced
  - used balanced sampling
- L: # of blocks used for making decision
  - Higher L better performance
  - 30 blocks approx. 3.84 seconds of speech
  - average speech length is 3.3 seconds
- bshape: # of blocks features grouped
- Information from compressed bit stream
  - All: all information extracted was used
  - MDCT: only spectral coefficients used
  - SF: only scale factor used
  - depending on window type used in compression SF can be 41 or 61 dimensional

Hyperparameters/ Configuration		Balanced Accuracy (%)	AUPRC (%)
Sampling	w/o Balanced Sampling	75.82	99.47
	w/ Balanced Sampling	**77.89**	**99.60**
L	20	76.56	99.55
	25	75.94	99.52
	30	**77.89**	**99.60**
	35	73.83	99.52
Loss	$\mathcal{L}_{BCE}$	77.89	99.60
	$\mathcal{L}_{BCE} + \mathcal{L}_{center}$	**79.13**	**99.63**
bshape	1	79.13	99.63
	2	82.54	99.69
	4	**82.98**	**99.67**
	8	80.93	99.65
Information from Compressed Bit stream	All	82.98	99.67
	WS+MDCT+scale factors (SF)(41)	83.67	**99.72**
	WS	50.00	90.90
	MDCT	53.46	92.35
	SF(41)	82.07	99.70
	SF(61)	**84.10**	99.70

Experiment 2: Performance of ASSD

- Trained on AAC compressed D_{tr} at data rate = 128kbps
- Retrained TSSDNet, and CCT on AAC compressed D_{tr}
- Input Vector is size of data extracted and processed by each method
- Class 0: Bona fide speech class, Class 1: Synthetic Speech

Data Rate	Method	Class 0 Accuracy (%)	Class 1 Accuracy (%)	Balanced Accuracy (%)	Accuracy (%)	AUPRC (%)	Input Vector Size
128kbps	TSSDNet	93.32	72.70	83.01	74.83	98.94	96,000
	CCT	59.00	96.80	77.90	92.90	68.70	65,536
	ASSD	80.86	84.96	82.91	84.53	98.94	1,860
64kbps	TSSDNet	93.32	72.70	83.01	74.83	98.94	96,000
	CCT	59.00	96.80	77.90	92.90	68.70	65,536
	ASSD	80.86	84.96	82.91	84.53	98.94	1,860
32kbps	TSSDNet	93.22	72.83	83.02	74.94	98.94	96,000
	CCT	59.07	96.81	77.94	92.91	68.39	65,536
	ASSD	80.69	84.10	82.40	83.75	98.89	1,860
16kbps	TSSDNet	90.27	73.35	81.81	75.10	98.77	96,000
	CCT	50.30	97.32	73.80	92.47	63.85	65,536
	ASSD	75.91	86.98	81.44	85.84	98.82	1,860

Observations From Experiment 2

- TSSDNet performance is low on majority class i.e., detecting synthetic speech, it is 72.70%
  - ASSD detect ~83% of synthetic speech
- CCT performance is significantly low on minority class i.e., bona fide speech, it is ~59%
  - ASSD detect ~81% of bona fide speech signal
- To make decision, TSSDNet use 96K time domain speech signal, CCT use spectrogram of size ~65K
  - ASSD comparatively use 33 times less information from bit stream 1860

Conclusion

- We investigated performance of existing work on compressed speech
- We proposed ASSD, a synthetic speech detector for detecting AAC compressed synthetic speech
- ASSD performs well on AAC compressed ASVspoof2019 dataset
- ASSD uses information directly from the compressed bit stream

References

[1] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," *Proceedings of the International Conference on Machine Learning*, vol. 139, pp. 8599–8608, July 2021, Virtual. [Online]. Available: http://proceedings.mlr.press/v139/popov21a/popov21a.pdf.

[2] G. Hua, *et al.*, "Towards End-to-End Synthetic Speech Detection," *SPL*, vol. 28, June 2021. DOI : 10.1109/LSP.2021.3089437

[3] E. R. Bartusiak, *et al.*, "Synthesized Speech Detection Using Convolutional Transformer-Based Spectrogram Analysis," *ACSSC*, October 2021. DOI : 10.1109/IEEECONF53345.2021.9723142

[4] F. Hassan, *et al*, "Voice Spoofing Countermeasure for Synthetic Speech Detection," *ICAI*, April 2021. DOI : 10.1109/ICAI52203.2021.9445238

[5] J. Herre and H. Purnhagen, "General Audio Coding," in *The MPEG-4 Book*, F. C. Pereira and T. Ebrahimi, Eds., Upper Saddle River, NJ, USA: Prentice Hall PTR, 2002, pp. 487–544. [Online]. Available: https://sound.media.mit.edu/resources/mpeg4/ audio/general/vancouver-general-audio.pdf.

[6] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *Proceedings of the Interspeech*, pp. 1008–1012, September 2019, Graz, Austria. DOI : 10.21437/Interspeech.2019-2249.

[7] Google Inc., *YouTube Recommended Upload Encoding Settings*, 2022. [Online]. Available: https://support.google.com/youtube/answer/1722171.

VIPER
Video and Image Processing Laboratory

ASSD: Synthetic Speech Detection in the AAC Compressed Domain

Amit Kumar Singh Yadav[1], Ziyue Xiang[1], Emily R. Bartusiak[1], Paolo Bestagini[2], Stefano Tubaro[2] and Edward J. Delp[1]

[1]Video and Image Processing Laboratory (*VIPER),*
School of Electrical and Computer Engineering , Purdue University, USA

[2]Image and Sound Processing Lab (ISPL),
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy