Modeling Audio Attributes of Deepfakes for Detecting Tampered Speech

Ewald Enzinger, Ph.D.
Eduworks Corporation
Corvallis, OR
ewald.enzinger@eduworks.com

Benjamin Bell, Ph.D.
Eduworks Corporation
Corvallis, OR
benjamin.bell@eduworks.com

ABSTRACT

Alarming advances in the sophistication of "deepfakes" pose numerous substantive threats as transnational criminal organizations, hostile intelligence agencies, terrorist groups, and other bad actors become more skilled at manipulating media for conducting fraudulent financial transactions, nefariously influencing elections, promoting deliberate misinformation, or distracting law enforcement and counter-terrorist organizations. Modeling techniques offer opportunities to apply machine learning to detect media that has been synthetically generated or tampered, even when such manipulation itself employs state of the art machine learning algorithms. Eduworks has developed a technology, Realtime Enhanced Voice Authentication (REVA), that employs machine learning models trained to detect audio deepfakes. In this paper, we present our approach for creating the machine learning models and describe the architecture for making REVA services available through an API. We demonstrate REVA's capability for real-time analysis of audio samples and of the audio portions for video content and applications of REVA's bulk processing in non-realtime. We conclude with a discussion of how REVA's fundamental approach can be generalized for application in adjacent domains, summarizing recent work helping the U.S. Coast Guard detect hoax distress calls.

ABOUT THE AUTHORS

Dr. Ewald Enzinger is a Senior Research Engineer at Eduworks, where he served as PI on the DARPA *Real-time Enhanced Voice Authentication (REVA)* project and as Co-PI on the DARPA *Semantic Forensics (SemaFor)* project and led the research effort on the *Distress Evaluation: Situational Cueing, Alerting, and Monitoring (DE-SCAM)* project with the U.S. Coast Guard. He has previously held research appointments at the Acoustics Research Institute of the Austrian Academy of Sciences and has published more than 30 peer-reviewed articles in scientific journals and at international academic conferences on innovative techniques in automatic speaker recognition and forensic voice analysis. He holds a PhD from the University of New South Wales, Sydney, Australia, and a MagPhil in Computational Linguistics from the University of Vienna.

Dr. Benjamin Bell is the COO of Eduworks, where he leads simulation, training, and decision support development. His research has addressed the use of simulation for training and education across a spectrum of applications, including K-12, higher education, military, and national security training. He has held faculty positions, chief executive positions in industry, and leadership roles for several international conferences. He holds a PhD from Northwestern, and BSE from the University of Pennsylvania.

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

Modeling Audio Attributes of Deepfakes for Detecting Tampered Speech

Ewald Enzinger, Ph.D.
Eduworks Corporation
Corvallis, OR
ewald.enzinger@eduworks.com

Benjamin Bell, Ph.D.
Eduworks Corporation
Corvallis, OR
benjamin.bell@eduworks.com

INTRODUCTION

Deepfakes is a term that generally refers to digital media content (images, audio, video) that has been altered to depict a person or group speaking or engaging in activity that differs from what the person or group actually said or did, or that changes where and when the event occurred, or all of the above. The deepfake content could be a change to an actual occurrence or could be entirely invented. Sophisticated applications and tools using Machine Learning technology are enabling a growing number of content manipulators to generate convincingly real media. And although there are multiple legitimate applications of these techniques (public service and education, advertising, entertainment), there is an alarming growth in the nefarious use of deepfakes for hoaxes and disinformation.

Deepfake attacks thus pose numerous substantive threats as transnational criminal organizations, hostile intelligence agencies, terrorist groups, and other bad actors become more skilled at manipulating media for conducting fraudulent financial transactions, nefariously influencing elections, promoting deliberate misinformation, or distracting law enforcement and counter-terrorist organizations.

In the next section, we introduce a technology Eduworks has developed, called Realtime Enhanced Voice Authentication (REVA), that employs machine learning models trained to detect audio deepfakes. We then present our approach for creating the machine learning models and describe the architecture for making REVA services available through an API. We describe REVA's capability for real-time analysis of audio samples and of the audio portions for video content and applications of REVA's bulk processing in non-realtime. We conclude with a discussion of how REVA's fundamental approach can be generalized for application in adjacent domains, and summarize recent work helping the U.S. Coast Guard detect hoax distress calls.

PROBLEM CONTEXT

Because of the growing danger to national security and global stability this emerging threat presents, Eduworks and its U.S. Government client are exploring modeling techniques that offer opportunities to apply machine learning to detect media that has been synthetically generated or tampered, even when such manipulation itself employs state of the art machine learning algorithms. Deepfakes have generally commanded global media attention in high-profile attacks that resulted in considerable financial loss, or for particularly convincing or entertaining depictions of celebrities or world leaders. Of greater interest in our work are deepfakes that create disinformation about election integrity, military and leaders and activities, diplomacy and diplomats, terrorists and counter-terrorism, and other targets relevant to national security.

The REVA project has been evolving along with the ML techniques employed by attackers. The project initially focused on technology-assisted telephonic social engineering, called voice phishing (or vishing) attacks, and we developed a mobile app for identifying suspicious callers or content. As the sophistication of ML approaches (specifically, generative adversarial networks, or GANs) expanded the opportunities available to attackers, we pivoted from protecting against vishing attacks to a more general capability to identify deepfake audio content; content that may have been tampered or synthesized. To make this capability available, we created a web platform to expose REVA services in several ways. In the next section, we summarize our technical approach to detecting possible deepfake content. We present our web-based services approach and a preliminary interface for providing content for REVA to analyze and for reviewing REVA's findings. We then summarize our results to-date, and discuss plans for the remainder of the program.

TECHNICAL APPROACH

Audio deepfake detection

REVA's audio deepfake detection component is based on a deep neural network model consisting of multiple sub-models. The initial feature extraction component is based on the Wav2vec 2.0 neural network architecture (Baevski et al., 2020), which is based on convolutional and transformer layers that is trained using self-supervised learning (see Figure 1).

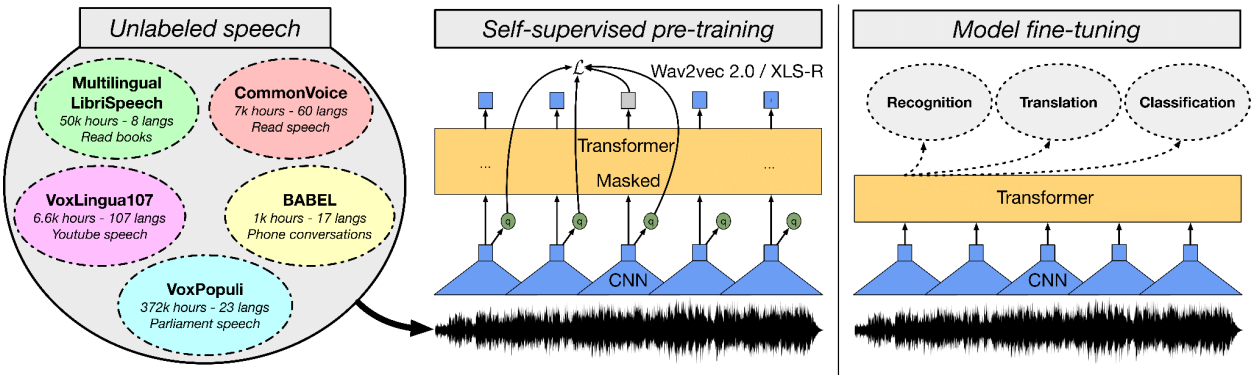

Figure 1. Overview of the Wav2vec 2.0 architecture (taken from Figure 1 from Babu et al., 2021).

The model uses a convolutional feature encoder to encode raw audio waveforms into latent speech representations. These latent representations are input to a transformer-masked network which initially quantizes the continuous representations to obtain a discrete set of outputs. These form the targets in the self-supervised learning objective and are in turn contextualized using the attention blocks from the transformer module, obtaining a set of discrete contextual representations. The feature encoder consists of seven convolutional blocks with 512 channels, strides of {5, 2, 2, 2, 2, 2, 2} and kernel widths of {10, 3, 3, 3, 3, 2, 2}. The transformer network consists of 24 blocks, an inner dimension of 4096, a model dimension of 1024, and a total of 16 attention heads. We considered a pre-trained multilingual acoustic model based on the Wav2vec 2.0 architecture named XLS-R-300m (Babu et al., 2021). This model was developed and published by Facebook Research and was trained on over 436,000 hours of speech from multiple publicly available speech datasets covering 128 languages. We used this wav2vec 2.0 model as the initial feature extraction component in our overall detection model and combined it with a ResNet-type encoder model and a binary classification head.

We trained this combined model on the task of classifying generated speech and pristine human speech. We collected a dataset consisting of over a million samples of speech samples generated by a large number of different speech synthesis and voice conversion algorithms, covering over 125 different algorithms and trained models capable of producing speech to impersonate different voices. These include speech synthesis acoustic models such as Tacotron (1/2), FastSpeech (1/2), FastPitch, and neural vocoders such as WaveNet, WaveRNN, MelGAN, and HifiGAN, as well as state-of-the-art diffusion approaches. We used data augmentation techniques to simulate the deteriorating effects of transmission and recording channels such as mobile telephones, VoIP, and compression using lossy compression algorithms such as MP3 and Opus. This was done to improve the robustness of the model under real world conditions, as the effects caused by reencoding could obscure artifacts left by the speech generation approaches.

Voice authentication

We combine our audio deepfake detection component with a voice authentication component. The aim is to verify that the voice of a speaker is consistent with a stored voice profile obtained during an enrollment phase or from a prior contact with the speaker. REVA's model uses a deep neural network consisting of an encoder network with 1-D depth-wise separable convolutions with Squeeze-and-Excitation (SE) layers with global context followed by channel attention based statistics pooling layer (Koluguri et al., 2022). See Figure 2 for a schematic overview of the neural network architecture of the encoder and decoder networks. The model was trained end-to-end with additive angular margin (AAM) loss (Deng et al., 2019). The dataset used to train the model consists of several million audio samples of over 10,000 speakers. REVA's voice authentication component uses this model to compute fixed-length vector representations from continuous speech input. These representations are used as voice profiles and are used for comparison with other voice profiles using cosine distance.

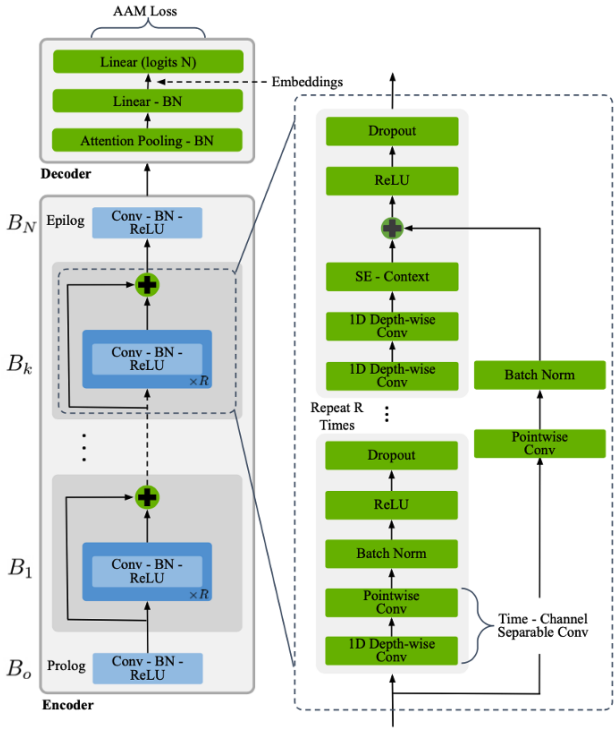

Figure 2: Voice authentication deep neural encoder and decoder network architecture (taken from Figure 1 from Koluguri et al., 2022).

The voice authentication model can also be used in an identification scenario where there is no claimed ID associated for verification but rather the aim is to identify who out of potentially thousands of enrolled speakers is speaking on an audio recording. This approach conceptually suffers from several drawbacks, such as the fact that the higher the number of enrolled speakers, the higher the chance of a random speaker whose voice is not enrolled in the system sounding sufficiently similar to an enrolled speaker that it is considered a match by the system. Another consideration is the runtime performance of performing identification among large numbers of enrolled speakers. Through techniques for fast Approximate Nearest Neighbor Search (ANNS) search, REVA can enable low-latency scoring of a voice profile against a potentially large list of voices. These include random projection trees and product quantization (Johnson et al., 2019).

EVALUATION RESULTS

Deepfake detection

In this section we present the results of the evaluation of REVA's deepfake detection model. For this evaluation we collected a number of datasets. We used the evaluation portions of the datasets created as part of the *Automatic Speaker Verification and Spoofing Countermeasures Challenges (ASVspoof)* in 2019 and 2021, specifically the Logical Access (LA) and Deepfake (DF) tasks. These datasets contain both genuine and synthetic speech of more than 48 speakers (21 male, 27 female) in many challenging conditions, such as having different transmission and recording conditions applied to them to obscure any potential artifacts left by the speech generation process. Over 100 different speech generation algorithms were used in the creation of the synthetic speech samples in the ASVspoof datasets. In addition, we collaborated with *VocaliD, Inc.*, a company specialized in creating personalized synthetic voices, to create a dataset of very natural synthetic speech samples of over 40 voices created using VocaliD's proprietary technology matched with genuine speech samples of the speakers whose voices were used to create the synthetic voices. This dataset is very challenging due to the high quality of the synthetic speech samples. Finally, we use a custom dataset consisting of a held-out portion of the data collected as part of the training data that we used to develop our model ("Eduworks Varied").

We compare REVA's model with that of four recently published baseline detection models:

1. RawNet2 (Tak et al, 2021)
2. AASIST (Jung et al., 2022)
3. AASIST-L (Jung et al., 2022)
4. RawBoost (Tak et al., 2022)

Results for Equal Error Rates (EERs) and classification accuracy are presented in Table 1. EER results are shown graphically in Figure 3, where shorter bars indicate better performance, and classification accuracy (%) in Figure 4, where taller bars indicate better performance. Both metrics indicate that REVA's detection model outperforms all four baseline detection models.

Table 1. Results of deepfake detection on multiple datasets.

Models	**ASVspoof 2019 Logical Access**		**ASVspoof2021 Logical Access**		**ASVspoof2021 DeepFake**		**VocaliD**		**Eduworks Varied**	
	Acc%	EER%	Acc%	EER%	Acc%	EER%	Acc%	EER%	Acc%	EER%
RawNet2	93.1	5.00	94.1	8.00	95.7	21.0	57.7	33.4	58.1	30.0
AASIST	95.8	1.00	96.7	10.0	88.5	17.0	63.4	35.7	62.9	28.9
AASIST-L	95.8	1.00	96.1	10.1	91.9	17.2	60.6	34.7	59.6	30.2
RawBoost	95.5	4.00	95.9	4.00	97.4	23	55.1	35.0	56.3	27.7
REVA	**99.7**	**0.14**	**98.8**	**0.94**	**99.4**	**1.50**	**94.8**	**4.10**	**99.6**	**0.30**

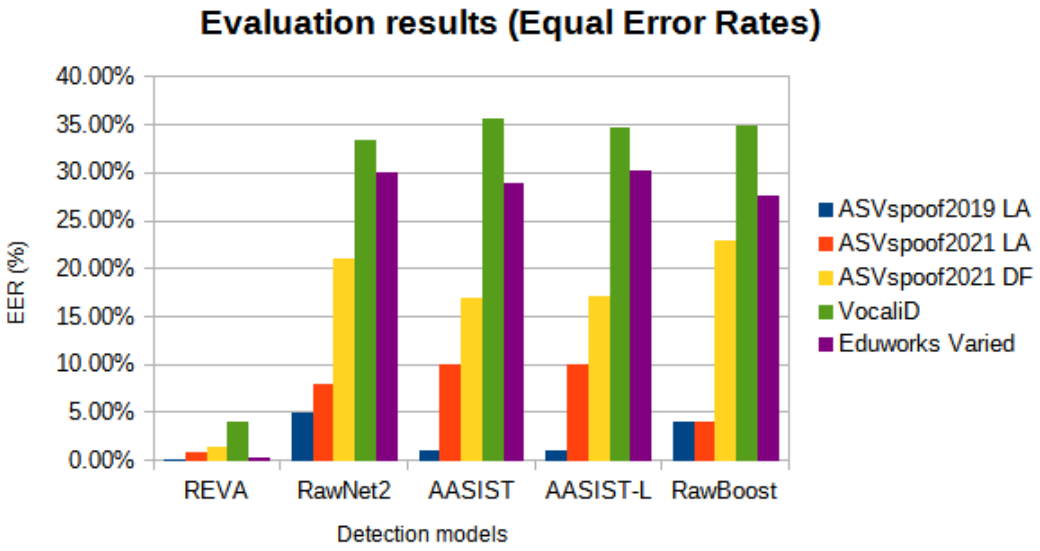

Figure 3. Evaluation results of deepfake detection models on three publicly available datasets (ASVspoof) and the VocaliD and Eduworks Varied datasets in terms of Equal Error Rates**.**

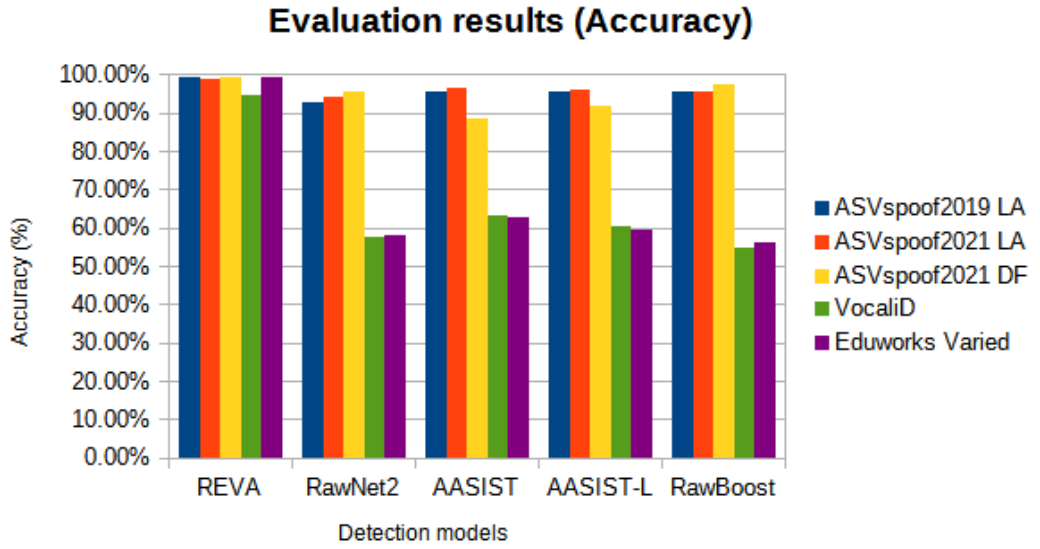

Figure 4. Evaluation results of deepfake detection models on three publicly available datasets (ASVspoof) and the VocaliD and Eduworks Varied datasets in terms of Accuracies**.**

Voice authentication

We evaluated the performance of REVA's voice authentication model using two datasets:

- The VoxCeleb1 dataset (Nagrani et al., 2017), which contains over 100,000 utterances of 1251 speakers (all pristine human speech)
- The VocaliD dataset, from which we create comparison pairs of trials of person-specific synthetic vs synthetic comparison and pristine vs. synthetic of the same source speaker comparisons.

We first evaluated the performance of REVA's voice authentication in a verification scenario on VoxCeleb1 when no deepfakes are present and on the VocaliD dataset to demonstrate that voice authentication, by itself, is susceptible to deepfakes. Results are shown graphically as EER rates in Figure 5 where shorter bars indicate better performance. As shown in Figure 5, voice authentication is very reliable in the absence of deefakes (note the low error rate in the left column). With deepfake (high quality synthesized voice samples), voice authentication techniques are far less effective (note the high error rate in the middle column). However, when combined with REVA's deepfake detection, typical voice authentication performance is nearly restored (right column).

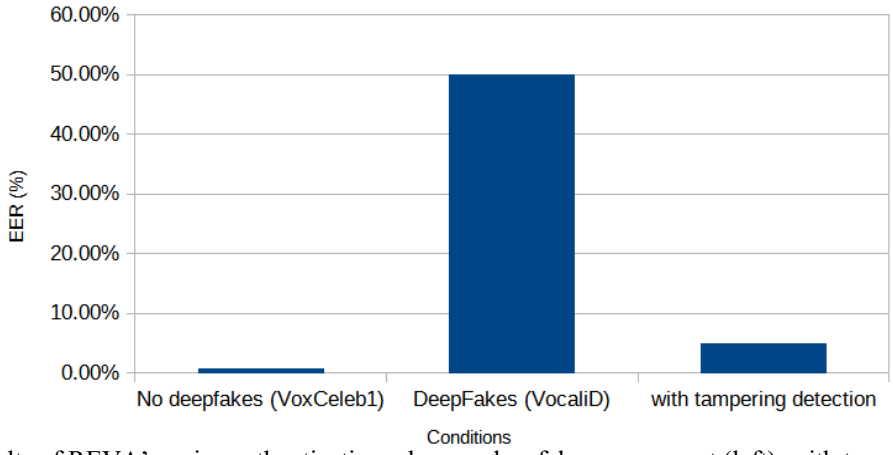

Figure 5. Results of REVA's voice authentication when no deepfakes are present (left), with targeted person-of-interest (POI) deepfakes (center) and when combined with REVA's deepfake detection (right).

We then compared the performance of the voice authentication model used in an identification scenario to that of multiple other recently proposed models: (1) ECAPA-TDNN (Desplanques et al., 2020), (2) SpeakerNet (Koluguri et al., 2020), and (3) WavLM (Chen et al., 2022). Each model was used to obtain voice profiles from both enrollment samples and test samples in the VoxCeleb1 dataset (identification split). The enrollment samples were then entered in a vector storage database using a flat index with an inverted file with exact post-verification (IVFFlat) to allow for fast approximate nearest neighbor search. Results presented in Table 2 follow the general convention for evaluating ML models, showing accuracy, precision, recall, and F-1 scores. As the results show, REVA's voice authentication model archives very competitive results, outperforming all other models. This demonstrates that REVA's voice authentication models can also be applied in identification scenarios.

Table 2. Results of voice identification performance on the VoxCeleb1 dataset.

	Top-1 Acc (%)	Top-5 Acc (%)	Precision	Recall	F1 Score
ECAPA-TDNN	0.946	0.949	0.952	0.950	0.944
SpeakerNet	0.916	0.938	0.922	0.921	0.910
WavLM-base-SV	0.917	0.955	0.920	0.919	0.909
REVA	**0.957**	**0.959**	**0.960**	**0.959**	**0.954**

REVA SERVICE ARCHITECTURE - USE CASES AND APPLICATIONS

REVA's algorithms and modeling approach are integrated into applications using REVA's API for use in various use cases and applications. At the heart of this API is the concept of streaming audio data processing enabled by REVA's WebSocket APIs. It enables asynchronous communication between the REVA Service and client applications implementing specific use cases. REVA's API can be backed directly by a worker analyzing the audio concurrently as well as by a work queue to which the REVA API can offload processing in off-line use cases.

In this section we summarize two current use cases:

- Real-time voice authentication for detecting deepfake telephone calls on a mobile phone.
- Forensic investigative analysis, search, and triage.

Real-time voice authentication on mobile phones

The real-time voice authentication use case involves live monitoring of audio streams of a telephone call on a mobile phone while it is being received. It is aimed at detecting and thwarting voice phishing ("vishing") attacks, a method used by social engineers to obtain information or influence actions of unsuspecting victims. It is an increasingly common and serious threat. In 2020 alone, the FTC's Consumer Sentinel Network recorded nearly 500,000 reports of imposter phone scams in which people lost $1.2 billion (FTC, 2020). The Communications Fraud Control Association estimates that global losses exceed $28B annually (CFCA, 2020). Voice phishing is possible because telephone systems lack a means for verifying the authenticity of communications or the identity of callers.

REVA's algorithms aim to detect techniques used in voice phishing by analyzing the incoming audio stream as it is being transmitted and checking for various cues for fraud, such as tampering detection and detecting impersonation via voice authentication. Users receive low-latency feedback in real time while on the call to warn them of fraud attempts. The client application pictured in Figure 6 was implemented as an Android application that receives phone calls via the REVA service.

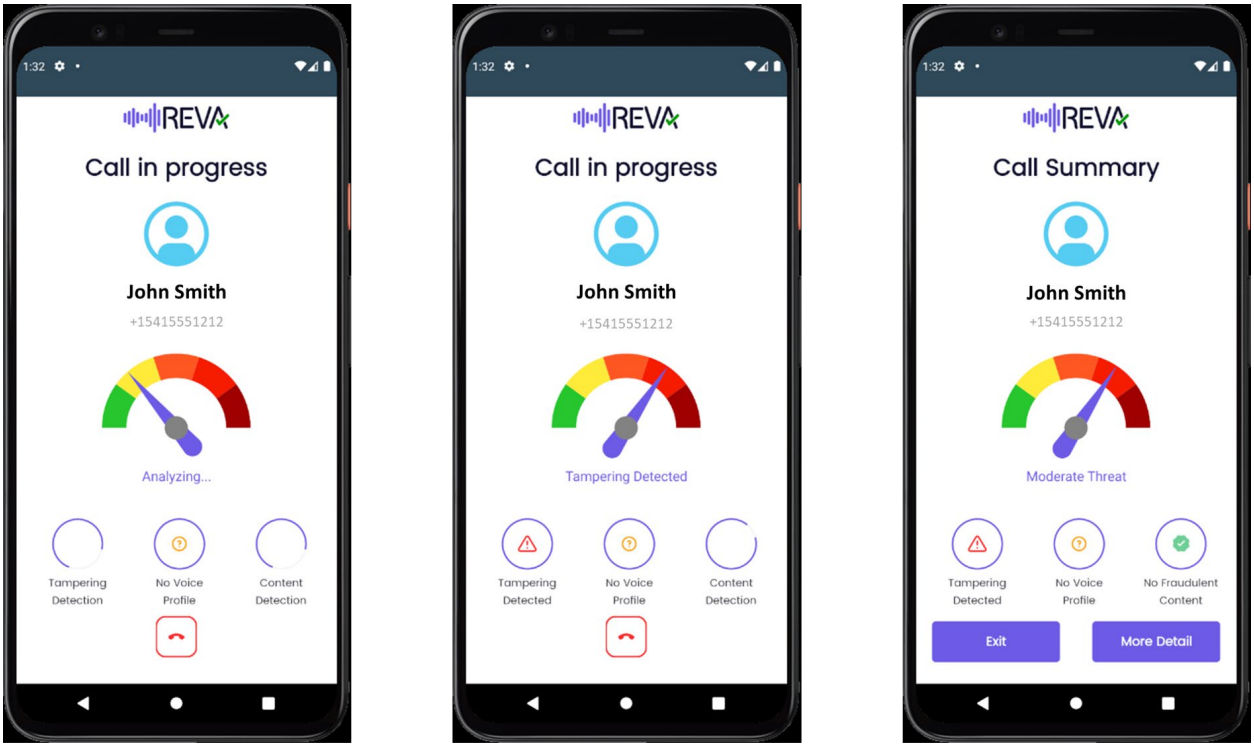

Figure 6: Screenshots of the real-time voice authentication use case for detecting deepfake telephone calls on a mobile phone. Left: During an ongoing phone call. Center: Tampering was detected. Right: Summary after the call

Forensic investigative analysis, search, and triage

The forensic investigative analysis, search, and triage use case involves forensic analysis of large volumes of recorded conversations. Analysis can support investigations into criminal activity or monitoring for threats, to name a few applications. Currently, analysis of recorded conversations or their transcripts is mostly performed manually, taking an organization large amounts of time and relying on human judgements on speaker identity and intent. These conversations (whether recorded phone calls or collected through other means) are thus typically underutilized, which hampers effectiveness.

REVA's algorithms and API can help analysts cope with high volumes of recorded voice conversations (from phone calls or other sources). REVA's detection capabilities can rapidly flag conversations where one interlocutor is using a digitally manipulated voice, for example using text to speech (TTS), voice conversion (VC), or Generative Adversarial Networks (GAN), and support keyword search (KWS) and recording condition detection. REVA's attribution capability enables speaker diarization as well as rapid interactive audio-based search for sections within or across recordings where a particular speaker is speaking (speaker search and linking). REVA's algorithms and API can help analysts triage their work by focusing attention on conversations that REVA recommends for further analysis. Figure 7 shows a mockup of REVA's instantiation of the investigative analysis, search, and triage use case.

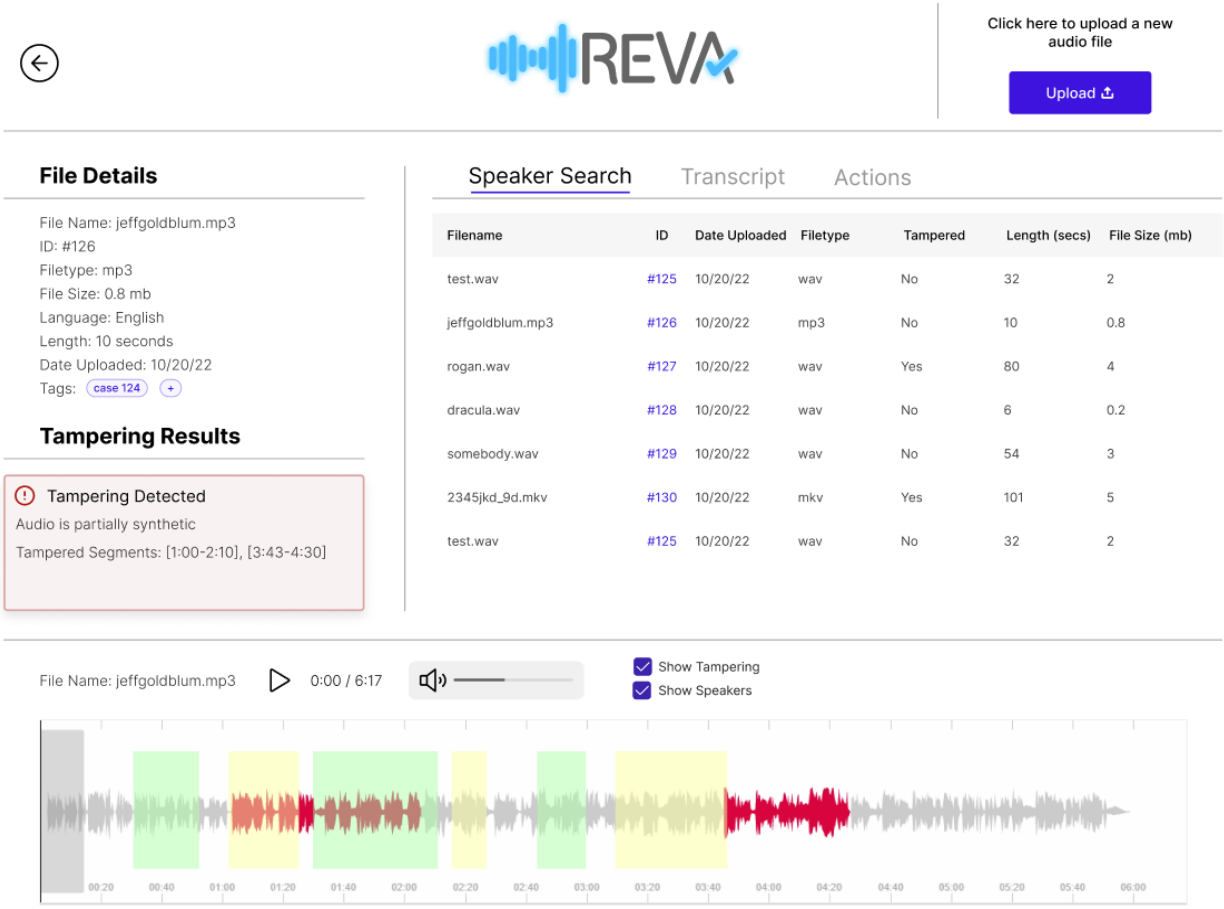

Figure 7: Mockup of the investigative analysis, search, and triage use case.

Generalizing REVA to adjacent domains: Monitoring of maritime VHF radio for hoax calls

REVA's capabilities in ML-powered audio forensics can be adapted for applications beyond deepfake detection. To explore this space, we performed a proof-of-concept for DHS designed to assist U.S. Coast Guard (USCG) personnel detect hoax distress calls. USCG watchstanders monitor transmissions of vessels at sea for aid calls via maritime VHF radio channel 16. Watchstanders continuously listen for distress or "mayday" calls to coordinate search and rescue missions. Fraudsters send hoax distress calls which cause damage in terms of the cost of ships and helicopters sent out for rescue as well as the risk to USCG personnel involved in these missions.

Using REVA's advanced modeling approach, we implemented a prototype system (shown schematically in Figure 8) that continuously monitors maritime radio channels for speech activity, listens for specific keywords, performs speaker identification to recognize known hoax callers, and tries to detect other speaker characteristics, such as intoxication. This proof-of-concept validated our approach and supports the generalization of REVA to other applications.

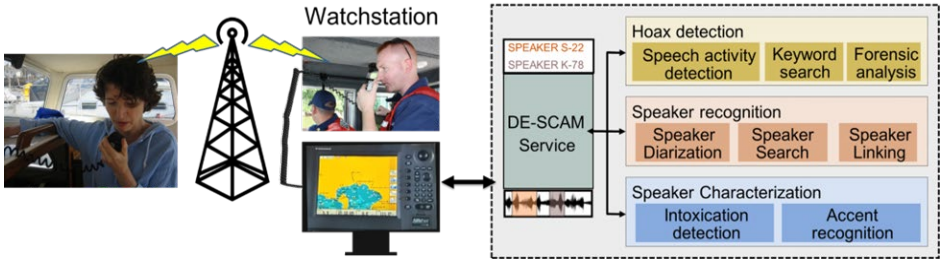

Figure 8: Conceptual diagram of the maritime VHF radio monitoring use case.

CONCLUSIONS

In this paper we gave an overview of REVA's technical approach, in particular deepfake detection and voice authentication. We presented results of evaluations of these capabilities on large varied datasets, demonstrating REVA's ability to generalize to unseen conditions. Through its modular architecture and API, REVA's approach can be generalized for application in adjacent domains, as showcased by the two use cases including real-time voice phishing monitoring of incoming calls on a mobile phone, and forensic investigative analysis, search and triage. We also presented a proof-of-concept for the USCG demonstrating REVA generalizing into adjacent applications such as detecting hoax distress calls.

REFERENCES

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint* arXiv:2111.09296.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. Wu, J., Zhou, L., Ren, S., Qian, Y.,Qian, Y., Wu, J., Zeng, M. & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1505-1518.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690-4699).

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *arXiv preprint* arXiv:2005.07143.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547.

Jung, J. W., Heo, H. S., Tak, H., Shim, H. J., Chung, J. S., Lee, B. J., Yu, H.-J., & Evans, N. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6367-6371). IEEE.

Koluguri, N. R., Li, J., Lavrukhin, V., & Ginsburg, B. (2020). SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification. *arXiv preprint* arXiv:2010.12653.

Koluguri, N. R., Park, T., & Ginsburg, B. (2022). TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8102-8106). IEEE.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, (pp. 2616-2620).

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2021). End-to-end anti-spoofing with rawnet2. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6369-6373). IEEE.

Tak, H., Kamble, M., Patino, J., Todisco, M., & Evans, N. (2022). Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6382-6386). IEEE.

Tak, H., Todisco, M., Wang, X., Jung, J. W., Yamagishi, J., & Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proc. Odyssey 2022 - The Speaker and Language Recognition Workshop*.

Wang, X., & Yamagishi, J. (2021). Investigating self-supervised front ends for speech spoofing countermeasures. In *Proc. Odyssey 2022 - The Speaker and Language Recognition Workshop.*