VOICESHIELD: AUDIO DEEPFAKE & DURESS DETECTION FOR BECU

STUDENTS: Kun Yang, Puyuan Du, Silu Zheng, Wenhan Zhang, Yuechen Guo, Yuxuan Wang , Zeyuan Zhao

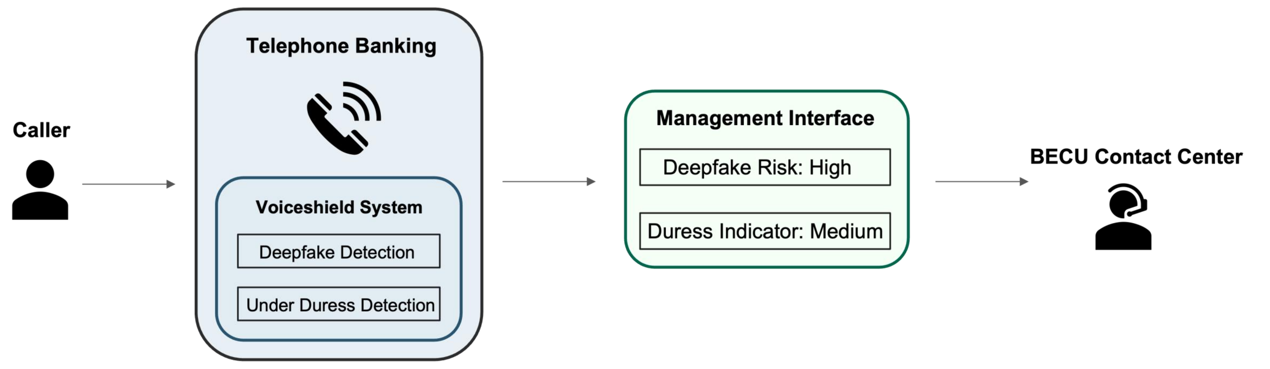
B|E|C|U

Objective

- BECU Contact Center needs a solution that performs voice signature match analysis of a member's voice (on the call) to identify if it's an AI deepfake, smart assistant (e.g., Google Smart Assistant), a caller under duress (e.g., pressured to withdraw money against their will) or the legitimate member.

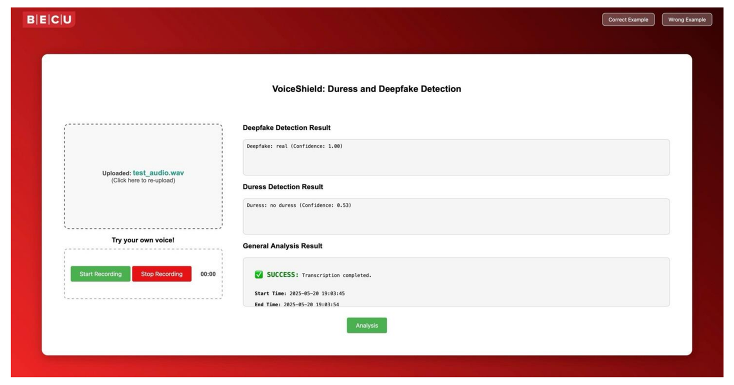
Requirements

- The solution should provide an analysis of the voice signature match, along with a confidence level. This information would be displayed to the Contact Center Representative. The solution would prompt the Representative to ask further questions to authenticate the user when the confidence level is low.

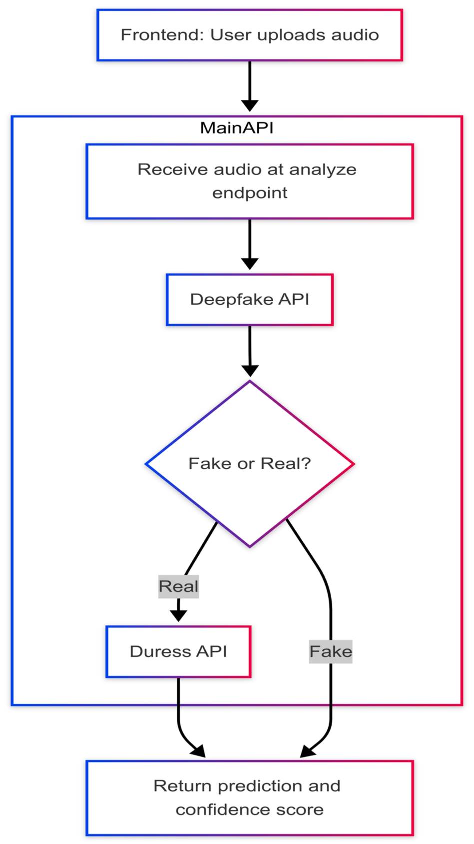


Front-end & Back-end Design

Front-end:
- Users can drag and drop or manually upload an audio file into the upload area.



Back-end:
- The API supports a two-stage workflow. It first calls the deepfake detection service and if the prediction is real, the audio is then forwarded to the duress detection service.
- The system is built with FastAPI and deployed on AWS EC2. It uses asynchronous I/O and a thread pool to handle concurrent requests efficiently and safely.



Deepfake Detection Technical Design

- We evaluated two recent top-performing speech deepfake detection models [2] [3]. AASIST2 showed better accuracy and accent robustness. Hence, it was selected as the base model for our work.

Models	CVoiceFake+ DeepVoice Datasets	Self-Collected Dataset	2s Audio Performance	4s Audio Performance	8s Audio Performance	Accent Variation Effect
RawNet	0.55	0.84	**0.67**	0.68	0.68	Yes
AASIST2(base)	**0.74**	**0.94**	0.58	**0.81**	**0.88**	**No**

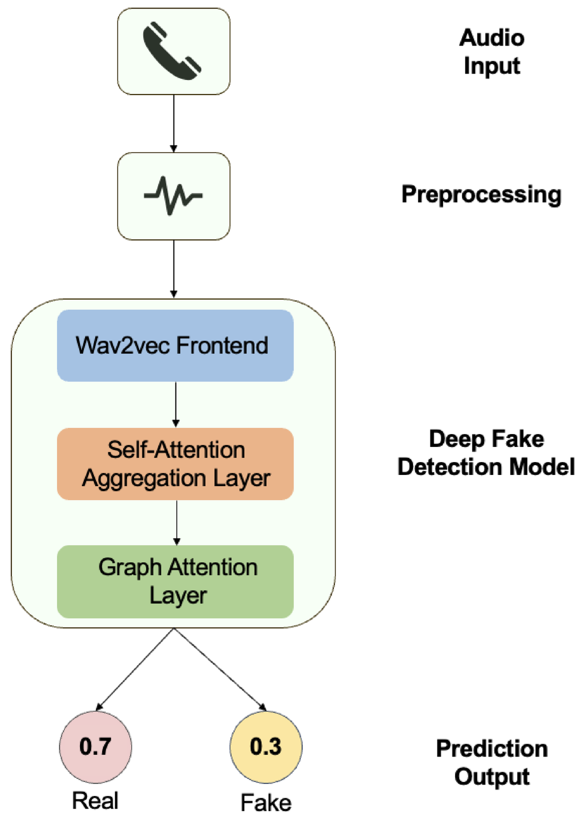
Metric: AUC

- During our experiments, we found that different deepfake speech generation techniques have distinct characteristics, and strong performance on one does not ensure good results on others. To improve robustness, we expanded the dataset with diverse techniques and retrained the model.

Models	CVoiceFake+ DeepVoice Datasets	Self-Collected Dataset
AASIST2(base)	0.74	**0.94**
AASIST2(finetune)	0.88	0.81
AASIST2(retrain)	**0.99**	0.84

Metric: AUC

Datasets	Deepfake Technology
CVoiceFake+ DeepVoice	Retrieval-based Voice Conversion, Griffin-Lim, WORLD, Parallel WaveGAN, DiffWave, MelGAN
BECU	Yourtts, xtts, VALL-E X
Self-Collected	DupDub AI Voice Generator, ByteDance's Doubao, Google Text-to-Speech AI



- Workflow Overview: User telephone audio is preprocessed (segmentation and resampling) and then passed to the AASIST2 model to predict the probability of being real or spoofed.

Under Duress Detection Technical Design

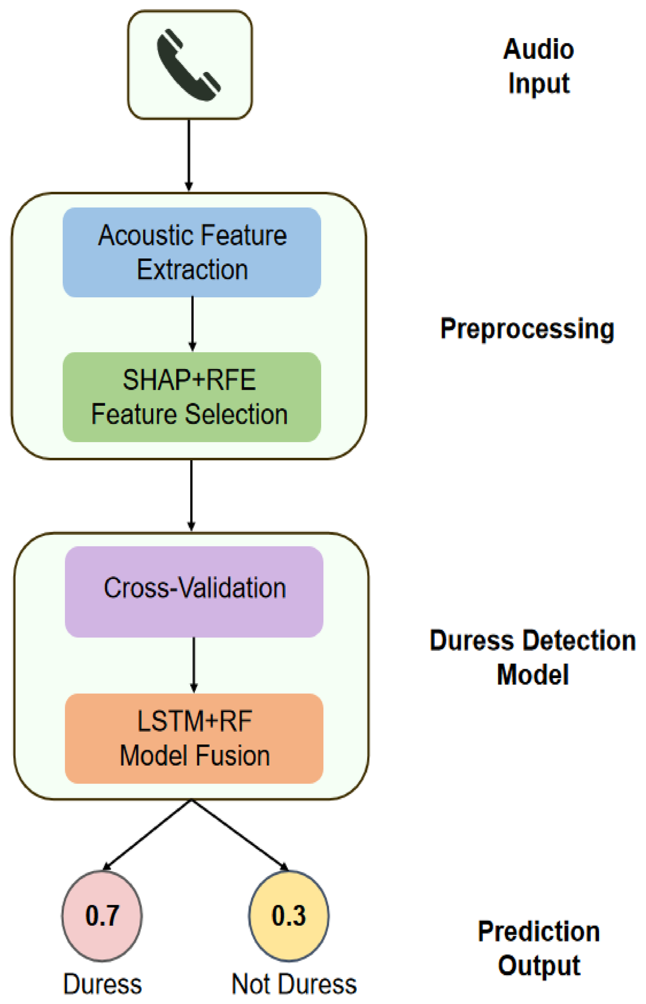
- The selection of the current duress detection system was informed by an evaluation of multiple technical paths:

Method (Key Tools)	Key Notes / Focus
✔ **Acoustic Feature Analysis + Machine Learning Classification (OpenSMILE, ML)**	Focuses on comprehensive sound characteristics; uses ML to identify duress patterns.
✘ Voice Features to Natural Language + LLM Analysis (Custom Extractor, LLM)	LLM analyzes text-converted voice features; challenges with specific emotion detection & false positives.
✘ Automatic Speech Recognition (ASR) + Sensitive Word Detection (Qwen2_Audio, ASR)	Relies on real-time transcription and predefined sensitive words/security codes for monitoring.
✘ Keyword Detection + BERT Semantic Analysis (Qwen2_Audio, spaCy, BERT)	Identifies keywords and uses BERT for contextual understanding; still reliant on keyword presence.

- Emerging from data analysis, critical insights into vocal duress markers directly informed the feature engineering and structural development of the subsequent model.

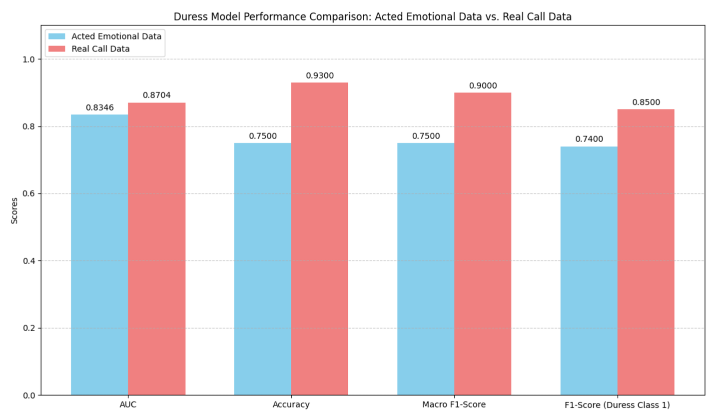
Vocal Cue	Duress Indication
Extreme Low Pitch Drops	Signals emotional distress
Distinct Pitch Patterns	Differentiates duress types
Steeper Loudness Slopes	Indicates panic/fear
Variable Loudness Increase	Links to deception/anxiety
Reduced Speech Variation	Shows physical tension

- This duress detection model processes raw audio to provide binary duress classificatio and an associated confidence score for age guidance.

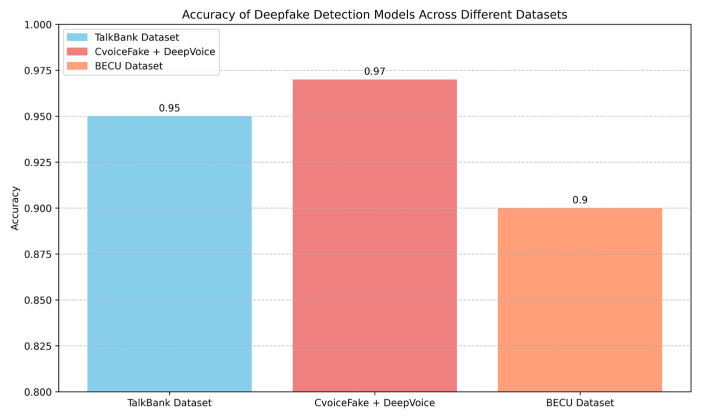


Results

- The under duress model and deepfake model results:



Under Duress model



Deepfake model

Future Work & References

- Improve the model to deal with more challenging situations such as a more advanced AI model.
- Explore the dynamic interplay of the features and their applicability across diverse real-world scenarios and populations.
- Reduce the latency of the system.
- Integrate with the company's real call scenarios.

[1] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10.2010.
[2] Tak, Hemlata, et al. "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation." arXiv preprint arXiv:2202.12233 (2022).
[3] C. Sun, S. Jia, S. Hou and S. Lyu, "AI-Synthesized Voice Detection Using Neural Vocoder Artifacts," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, BC, Canada, 2023, pp. 904-912, doi: 10.1109/CVPRW59228.2023.00097.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

FACULTY MENTOR: Jai Jaisimha
INDUSTRY MENTOR: Alan Wilson & Dan Gibbons
SPONSOR: BECU