A Novel Dataset for Testing Anti-spoofing Models in a Telephony Environment

Zachary N. Houghton[1,2], Dan Pluth[2], Jordan Hosier[2], Vijay K. Gurbani[2]

[1]University of California, Davis

[2]Vail Systems

UC DAVIS
DEPARTMENT OF LINGUISTICS

VAIL
https://www.freeclimb.com/research

RAILS
Research on AI in Linguistics and Systems

Background

- Recent rise in the quality of synthetic voices
- Increased security risk[1]
  - Scammers are already taking advantage of this to scam individuals and companies
  - Spoofing voices of individuals to get private information
  - Spoofing voices of loved ones or company execs to scam individuals out of money

[1]https://www.scmp.com/print/news/world/article/3025772/ai-first-voice-mimicking-software-used-major-heist

Background

- Research has begun to address these security risks using liveness detection models [1, 2, 3, 4, 5]
  - Liveness detection models focus on learning whether a voice is authentic or spoofed (i.e., whether a voice is “live” or not)
- Several datasets and models have been proposed for liveness detection [1, 2, 6]
  - Mostly focus on clean, relatively noise-free contexts.

Liveness Detection in Telephony Domain

- Few studies have examined liveness detection in noisier domains
  - [6] created a dataset called *phonespoof*. However, their dataset no longer reflects the ecosystem of contemporary spoofing methods.
  - [7], while not explicitly examining telephony speech, examined the effect of channel conditions on liveness detection. However, they found that their models struggled substantially with out-of-domain data.

Roadmap

The present talk builds upon the previous work to create a novel liveness dataset.

- Introduce our novel dataset
- Introduce our cellularization methodology
- Present our model
    - Demonstrate the viability of a lesser-known finetuning approach
- Present the results of our model on the novel dataset
- Discuss future directions

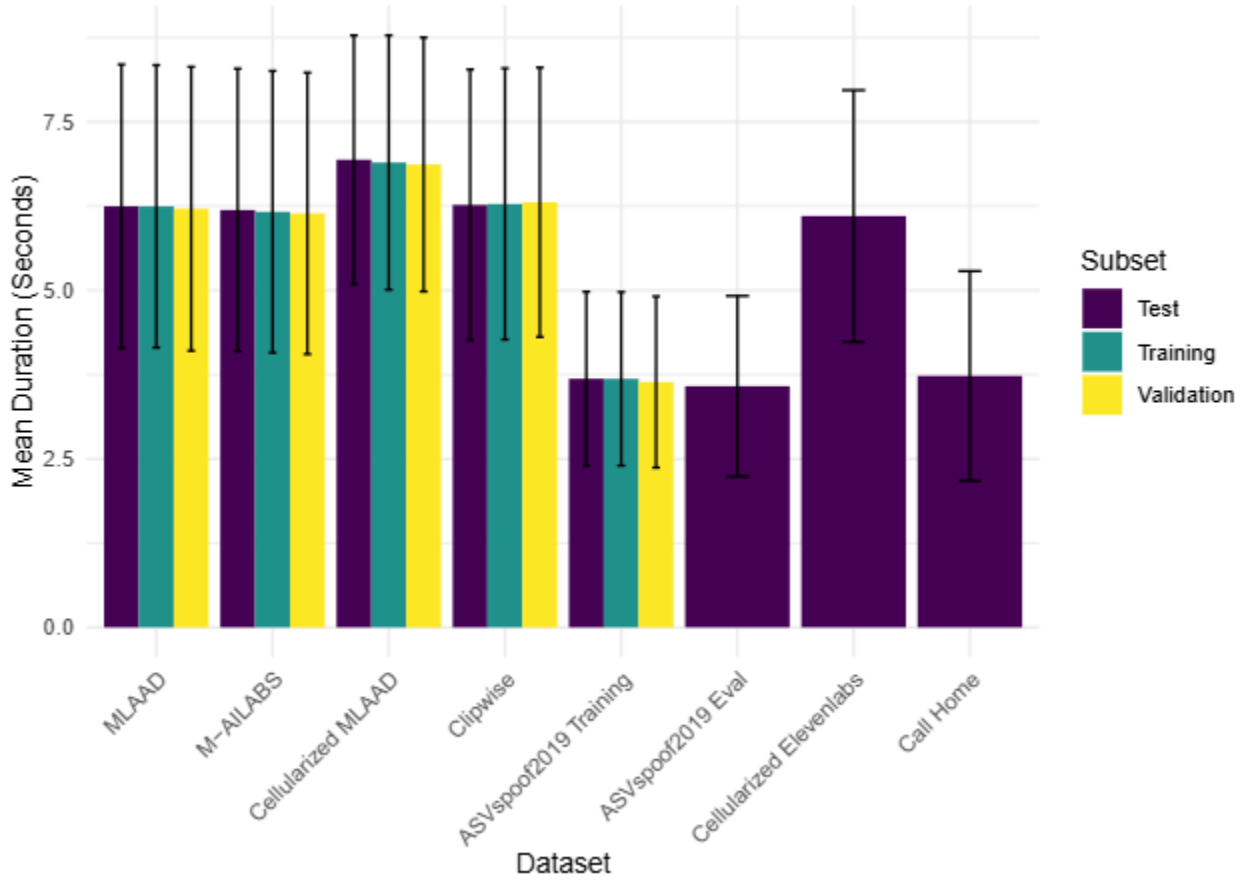
Dataset Breakdown[2]

Dataset	Description
M-AILABS	~1,000 hrs audiobooks, clean, multiple languages
Call Home	120 unscripted 30-min telephone conversations
MLAAD	175 hrs synthetic audio, 59 TTS models (based on M-AILABS)
Cellularized MLAAD	MLAAD passed through telephony pipeline
Cellularized ElevenLabs	175 voices used to clone LibriSpeech, then cellularized
ASVspoof2019	Training: 20 speakers, 6 synthesizersEvaluation: 13 novel synthesizers → tests generalization
Clipwise	Real financial institution calls

Dataset Breakdown

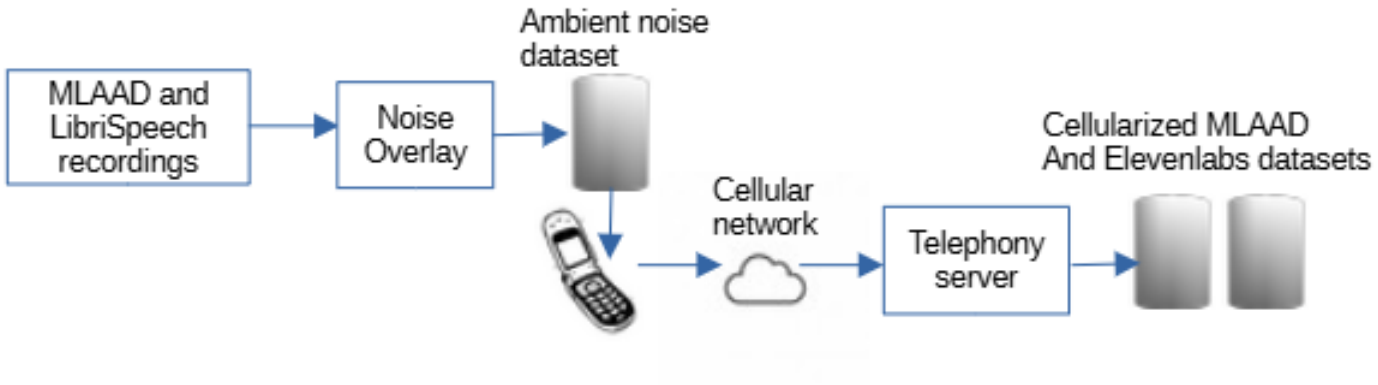
Dataset	Type	Train	Val	Test
MLAAD	Synthetic	36,000	4,500	4,500
M-AILABS	Human	24,000	3,000	3,000
Cellularized MLAAD	Synthetic	16,000	2,000	2,000
Clipwise	Human	40,000	5,000	5,000
ASVspoof2019 (Train)	Mix	16,000	2,000	2,000
ASVspoof2019 (Eval)	Mix	–	–	54,540
Cellularized ElevenLabs	Synthetic	–	–	1,788
Call Home	Human	–	–	11,549

Dataset Breakdown



Cellularization Method

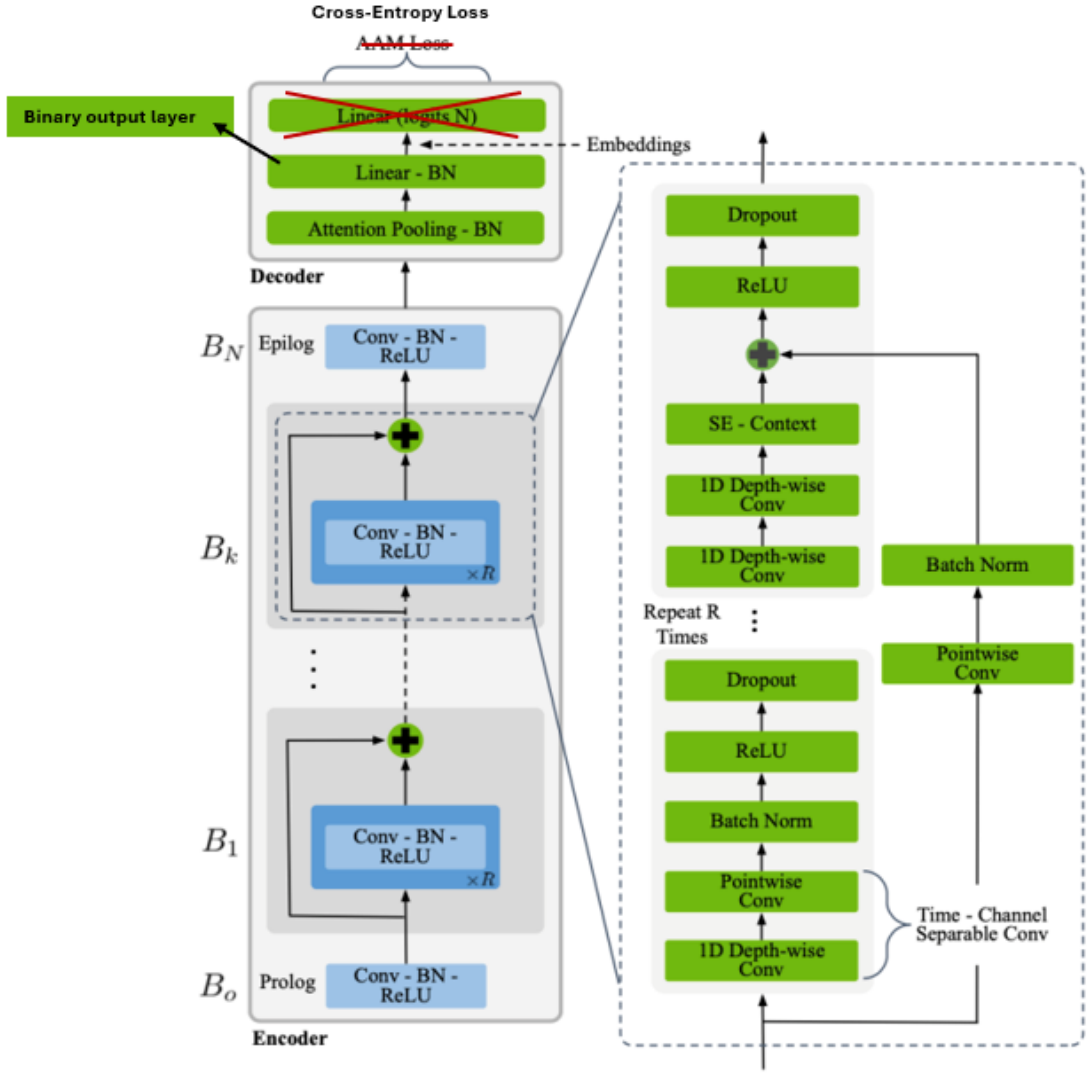
- The goal of the cellularization process is to embed telephony-related artifacts into the captured media.
  - Telephony-specific codecs like G.711, G.723, and G.729 and radio resource contention leading to jitter and delay



Model

- TitaNet pre-trained speaker recognition model
  - Output layer replaced with 2-dimension softmax layer
  - The motivation for this approach was that the speaker recognition model may have learned characteristics of speech that could be helpful for liveness detection
- Cross-entropy loss instead of additive angular margin (AAM didn’t perform well)

Model


Cross-Entropy Loss
AAM Loss
Binary output layer
Linear (logits N)
Linear - BN
Attention Pooling - BN
Decoder
Embeddings
B_N
Epilog
Conv - BN - ReLU
B_k
Conv - BN - ReLU
$\times R$
B_1
Conv - BN - ReLU
$\times R$
B_o
Prolog
Conv - BN - ReLU
Encoder
Dropout
ReLU
SE - Context
1D Depth-wise Conv
1D Depth-wise Conv
Repeat R Times
Dropout
ReLU
Batch Norm
Pointwise Conv
1D Depth-wise Conv
Time - Channel Separable Conv
Batch Norm
Pointwise Conv

Finetuning Method

- Output layers become specialized for the task
- Training all the layers with a new output layer can result in catastrophic forgetting

Our approach to deal with this:

- Freeze all layers and finetune output layer
- Unfreeze all layers and finetune again

Results – Confusion Matrix

		Actual	
		Synthetic	Human
Predicted	Synthetic	55,176	1,091
	Human	2,336	25,774

Results – Precision/Recall/Acc/EER

Metric	Value
Precision	0.981
Recall	0.959
Accuracy	0.959
EER	0.041

Results – In vs out-of-domain

Domain	Dataset	Mean Accuracy
In-domain	asvspoof2019_training	0.998
	cellularized_mlaad	1.000
	mlaad	1.000
	mailabs	0.972
	clipwise	0.993
Out-of-domain	asvspoof2019	0.952
	call_home	0.942
	cellularized_elevenlabs	1.000

Results – Breakdown by Synthesizer

Synthesizer	Type	Method	Mean Accuracy	# Obs.
A07	TTS	vocoder + GAN	1.000	2,942
A08	TTS	neural waveform	0.991	3,164
A09	TTS	vocoder	1.000	3,329
A10	TTS	neural waveform	0.991	2,797
A11	TTS	Griffin-Lim	1.000	2,852
A12	TTS	neural waveform	1.000	3,554
A13	TTS-VC	concat. + filtering	**0.886**	3,179
A14	TTS-VC	vocoder	1.000	3,895
A15	TTS-VC	neural waveform	1.000	3,895
A16	TTS	waveform concatenation	1.000	3,709
A17	VC	waveform filtering	**0.882**	4,717
A18	VC	vocoder	**0.713**	4,718
A19	VC	spectral filtering	0.998	4,702

Conclusion

- A model trained on our novel dataset performs well in a zero-day context
  - Though, unsurprisingly there is a slight decrease in performance on out-of-domain data
- Model trained our dataset performs well on telephony data
- However, seems sensitive to synthesizer type
- Demonstrates the need to account for channel-conditions of real-world use cases when researching methods
  - In this instance, telephony conditions

Future Directions

- Testing a wide variety of different models on our dataset
- Expanding the dataset to include more synthesizers
- Examining the effects of audio duration on model performance

Questions/Comments?

References

1. Kawa, P., Plata, M., & Syga, P. (2022). Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. *arXiv preprint arXiv:2206.13979*.
2. Müller, N. M., Kawa, P., Choong, W. H., Casanova, E., Gölge, E., Müller, T., ... & Böttinger, K. (2024, June). Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.
3. Lavrentyeva, G., Novoselov, S., Volkova, M., Matveev, Y., & De Marsico, M. (2019, May). Phonespoof: A new dataset for spoofing attack detection in telephone channel. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2572-2576). IEEE.
4. Kinnunen, T., Wu, Z. Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012, March). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4401-4404). IEEE.
5. Li, M., Ahmadiadli, Y., & Zhang, X. P. (2024). Audio anti-spoofing detection: A survey. *arXiv preprint arXiv:2404.13914*.
6. Lavrentyeva, G., Novoselov, S., Volkova, M., Matveev, Y., & De Marsico, M. (2019, May). Phonespoof: A new dataset for spoofing attack detection in telephone channel. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2572-2576). IEEE.
7. Pluth, D., Hosier, J., Zhou, Y., & Gurbani, V. K. (2025, March). Echoes Unveiled: Identifying Synthetic Voices. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 582-587). IEEE Computer Society.