Vocal Tract & Voice Source Features for Synthetic Speech Detection

Tadipatri Uday Kiran Reddy[1], Sahukari Chaitanya Varun[1], Pranav Kumar Kota[1],
Muhammed Fayis[2], Sankala Sreekanth[1], K. Sri Rama Murty[1]

1 Electrical Engineering, IIT Hyderabad , 2 Engineering Science, IIT Hyderabad

Problem Statement

- Identify the synthesis algorithm used to generate the given audio
  - Extract a fixed-length "synthesizer signature" from varying-length audio and classify into an algorithm.

Dataset

- Clean data
  - Five different synthesis algorithms + Unknown class
  - 1000 utterances synthesized using each class are given.
- Noisy data: Noise augmented(AWGN, Reverberation, MP3 compression) version of clean
- Natural samples from VCTK and CMU datasets

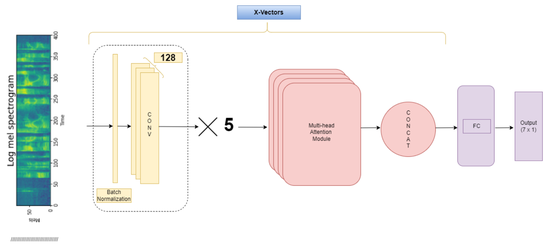
Speech Production

- Speech signal is an outcome of time-varying vocal-tract system driven by time-varying excitation
- Synthetic speech algorithms seek to mimic this natural system
  - May fail to incorporate subtle features of speech production
  - At system-level: coarticulation of phoneme sequences
  - At source-level: prosodic variations, jitter, shimmer etc.,

Vocal-Tract System Features

- VTS has concatenation of acoustic tubes
- Magnitude spectral envelope of speech represents resonances of VTS
- Mel-wrapped magnitude spectrum is used as features

Extracting & Classifying Synthesizer Signatures

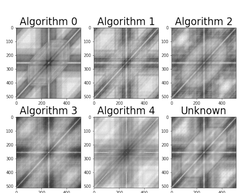


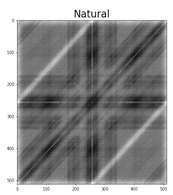
- X-Vector framework is adapted for synthesizer identification
  - Convolutional filter-bank captures long-term temporal dependencies
  - Statistical pooling extracts fixed-dimensional signature
  - Multi-head attention offers differential temporal weightage

Magnitude Spectral Features Results

- Training: 98%, Evaluation: 98%
- Major confusion between algorithm-1 and algorithm-2.

Bicoherence

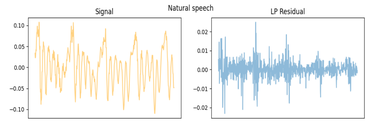


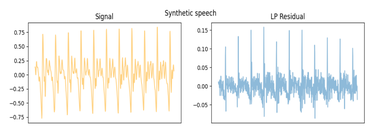


- Bispectrum capture phase spectral information
- Contains discriminative evidence - from voice source

Voice Source Features

- Speech signal is inverse filtered through estimated VTS response $V(z)$ (Linear prediction filter)
- The output of the inverse filter is referred to as LP residual
- Natural signal: Higher uncertainty around the glottal closure instant
- Lesser predictability across the instants - higher long-term residual





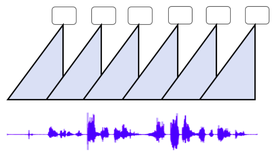
Feature Extractor

- Strided convolutional filter-bank for feature extraction
  - Filter sizes and strides decide the effective frame size and shift
  - Empirical studies were conducted with different filter sizes and strides
- Filter weights are estimated to minimize the classification loss

Results

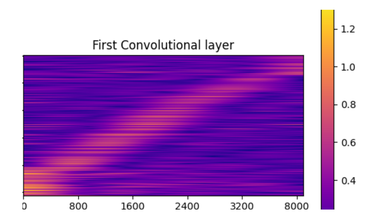
	Clean		Noisy	
Model (Parameters)	**Train%**	**Eval%**	**Train%**	**Eval%**
VGGish($\sim 72M$)	**99**	89	-	70
YAMNet($\sim 3.7M$)	**99**	92	-	76
X-vector($\sim 0.3M$)	98	**98**	**98**	93.3
LP Residual X-Vectors($\sim 1.3M$)	98	95.4	**98**	**96.6**

Effective Receptive Field



Receptive Field	Accuracy
35ms	94.7%
131ms	**96.67**%
310ms	92.9%

Frequency Filter Maps



- Frequency response of the first layer of convolutional filter bank.
- Sorted according to peak in the magnitude response.
- This follows Mel-structure: around 50 filters below 800 Hz.

Future Course

- Analyze the attention weights to locate unnatural regions
- Combine evidence from source and system level signatures

References

- Snyder et al, *X-Vectors: Robust DNN Embeddings for Speaker Recognition*, ICASSP-2018
- Albadawy, Ehab et. al., *Detecting AI-Synthesized Speech Using Bispectral Analysis*