AASIST: AUDIO ANTI-SPOOFING USING INTEGRATED SPECTRO-TEMPORAL GRAPH ATTENTION NETWORKS

Poster 2372

Jee-weon Jung[1], Hee-Soo Heo[1], Hemlata Tak[2], Hye-jin Shim[3], Joon Son Chung[4], Bong-Jin Lee[1], Ha-Jin Yu[3], Nicholas Evans[2]
[1]Naver Corporation, South Korea, [2]EURECOM, Sophia Antipolis, France
[3]School of Computer Science, University of Seoul, [4]Korea Advanced Institute of Science and Technology, South Korea

Overview

❖ Objective: Develop an **efficient, single system** that can detect a broad range of different spoofing attacks spanning in both spectral and temporal domains

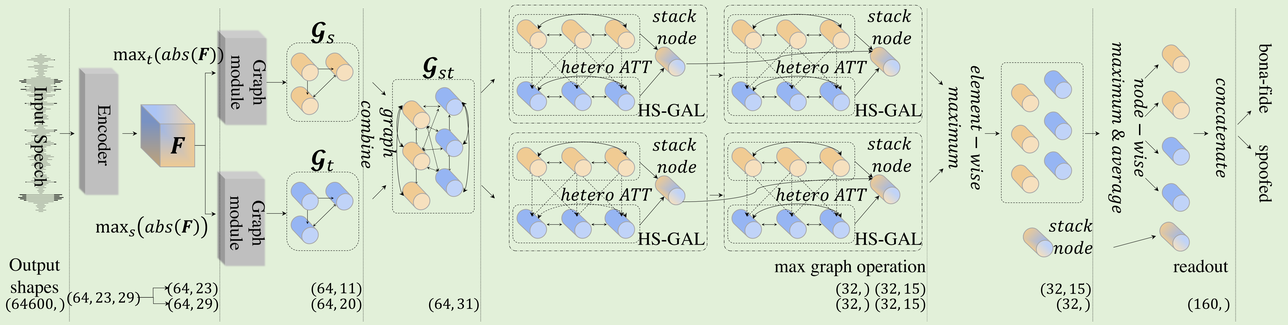
❖ Proposed model: **AASIST**

- Builds upon previous state-of-the-art system that extracts two (spectral and temporal) views (graphs) from raw waveform
- Models **heterogeneous graph** using proposed mechanism concurrently

❖ EER 0.83% / min t-DCF 0.0275 on the ASVspoof2019 LA dataset

- Includes 19 different voice conversion and text-to-speech attacks

❖ Code available in https://github.com/clovaai/aasist



EER: equal error rate; DCF: detection cost function; LA: logical access; HS-GAL: heterogeneous stacking graph attention layer; MGO: max graph operation

Proposed architecture & techniques

❖ Spoofing artefacts can lie in specific sub-bands or frames

- Depends on the attack algorithm

❖ **Strategy**: extract *spectral* & *temporal* representations → combine

❖ **Architecture**

- RawNet2-encoder: extracts 3-dimensional feature map from raw waveforms
  - (channel, spectral bins, temporal frames)
- Element-wise maximum on either spectral or temporal dimension → Two graph representations
- Graph module: graph attention layer + graph pooling layer
- Graph combination: add edges to all possible node pairs
- **HS-GAL** jointly models two heterogeneous graphs
  - *Heterogeneous attention*: utilise different parameters for attention
  - *Stack node*: receives information from all other nodes
- **MGO** exploits two same branches
  - Different parameters, each branch includes two HS-GALs
- Readout: concatenate node-wise maximum, average, and stack node

Dataset & Configurations

❖ Dataset: ASVspoof2019 LA

	# bona fide utterance	# spoofed utterance
Train	2,580	22,800
Development	2,548	22,296
Evaluation	7,355	63,882

❖ Input: raw waveform (4 seconds)

❖ RawNet2-encoder: 6 residual blocks

❖ Graph pooling: reduce 50% nodes

❖ Optimiser: Adam w/ learning rate of 0.0001

Experiment results

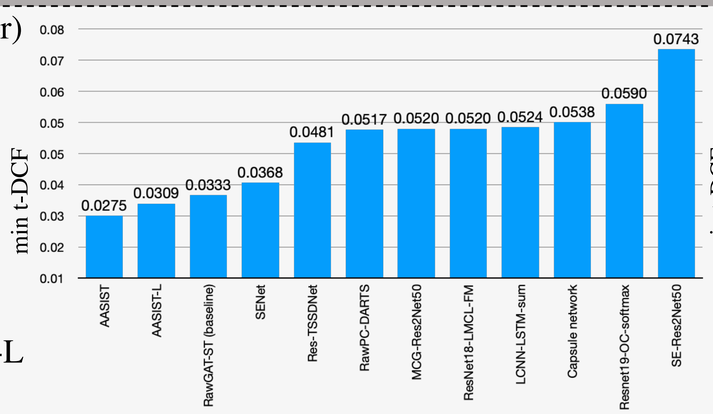
❖ Metrics (lower is better)

- EER(%)
- min t-DCF

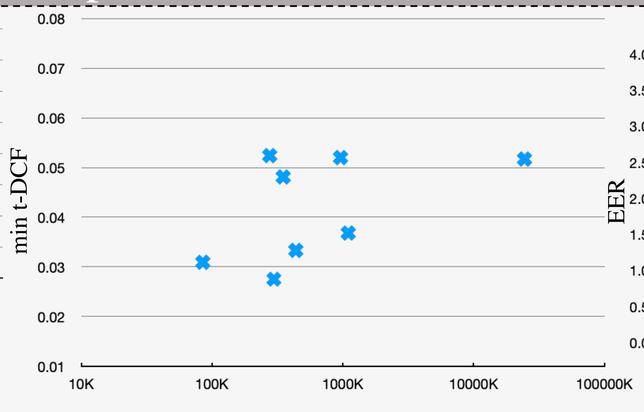
❖ Two model sizes

- AASIST: 297k
- AASIST-L: 85k

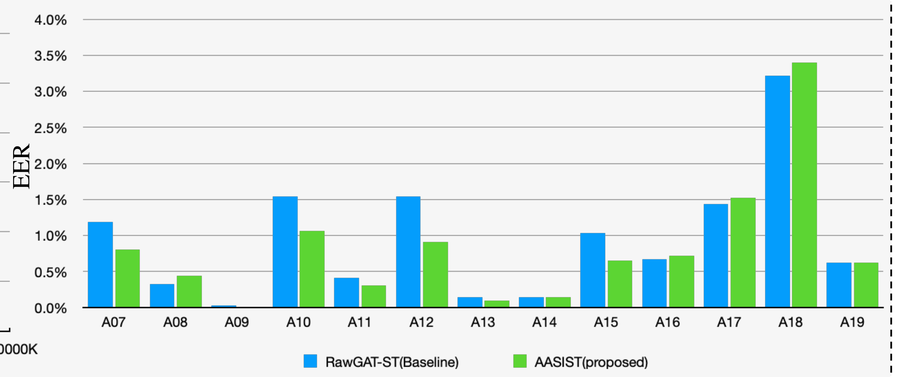
❖ AASIST and AASIST-L show state-of-the-art performance



Recent systems comparison



parameter / min t-DCF plot



Attack-wise comparison