Audio Replay Spoof Attack Detection Using Segment-based Hybrid Feature And Densenet-LSTM Network

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Chi Man Pun

Outline

1. Introduction
2. Segment-based Hybrid Feature Extraction
3. The Proposed DenseNet-LSTM Classifier
4. Experimental Results
5. Conclusion and Future Work

Introduction

- Automatic speaker verification (ASV) system have experienced explosive growth.
- The highest risk is that spoofed speech may gain unauthorized access.
- The genuine and spoofing discriminative ability is one of the key issues in multimedia information security.

Introduction (Cont.)

- It is well known that ASV systems can be vulnerable to spoofing

- There are four main types of audio spoof attacks.

Speech synthesis (SS)

Voice conversion (VC)

Impersonation

Replay

Introduction (Cont.)

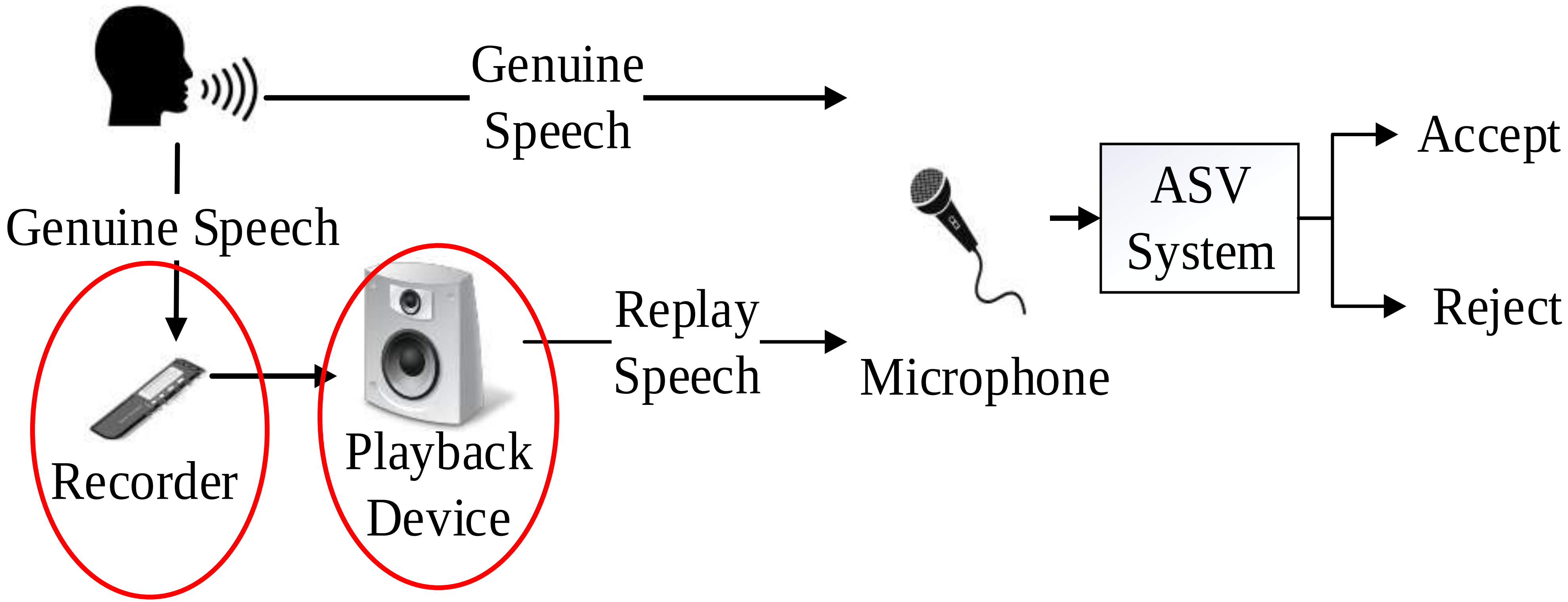
- 2015 'Automatic Speaker Verification Spoofing and Countermeasures Challenge' (ASVspoof 2015)
  - SS, VC, or other unknown spoof attacks
  - MFCC and CFCCIF features achieved EER of 1.211%
    - Mel-frequency cepstral coefficient (MFCC)
    - Cochlear filter cepstral coefficients and change in instantaneous frequency(CFCCIF)
  - CQCC based features with average EER of 0.255%
    - Constant-Q cepstral coefficient(CQCC)
  - CNN,RNN

Introduction (Cont.)

- 2017 'Automatic Speaker Verification Spoofing and Countermeasures Challenge' (ASVspoof 2017)
  - Focus on replay spoof attack
  - Baseline system: CQCC+GMM (EER 24.77%)
    - Gaussian mixture model (GMM)
  - Fusion CQCC GMM+MFCC ResNet+CQCC ResNet (EER 13.30%)
  - CQCC+DNN, CQCC+ResNet,MFCC+ResNet

Introduction (Cont.)

➢ The implementation process of the replay attack



Introduction (Cont.)

- The audio spoof detection methods are divided into two steps:

  a) Extract the features of speech fragments

  CQT, MFCC, CFCCIF, CQCC or spectrum (by STFT)

  b) Classify all input speech based on their features

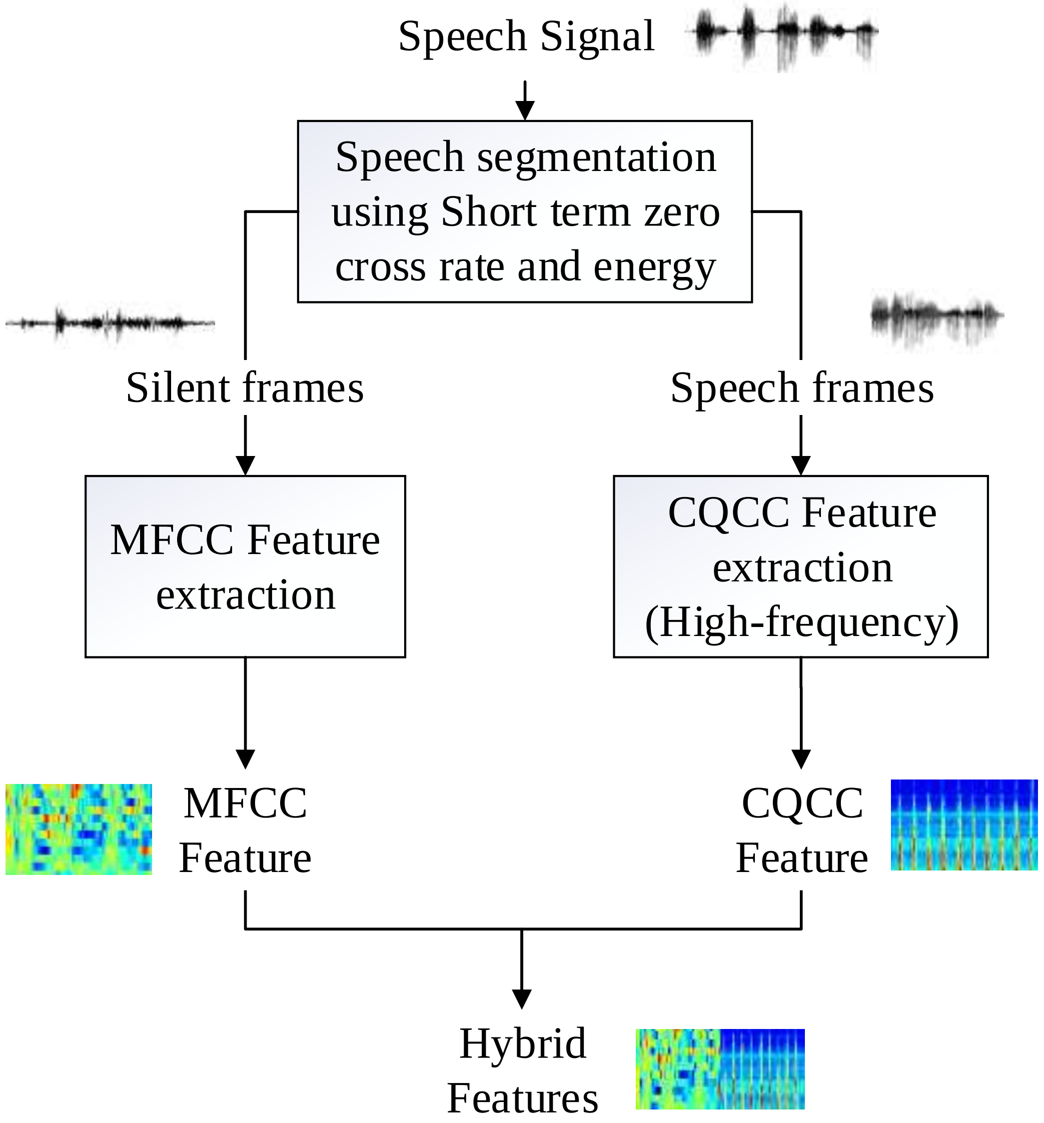
  GMM, DNN, LSTM

Introduction (Cont.)

- Contributions
  a) A segment-based hybrid feature extraction method is proposed
  b) Compared to traditional features, the hybrid feature is much better for distinguishing between genuine and replay spoof speech.
  c) A novel DenseNet-LSTM architecture is proposed as back-end classifier.

Segment-based Hybrid Feature Extraction

- ➢ Speech segmentation
- ➢ Feature extraction (Respectively)
- ➢ Concatenate to Hybrid Features



Segment-based Hybrid Feature Extraction(Cont.)

- Waveforms of the genuine speech and the replayed speech

Segment-based Hybrid Feature Extraction(Cont.)

Part 2

- Speech Segmentation Using Short-term Zero Cross Rate and Energy

a) Short-Term Zero Crossing Rate (ZCR)

$$st_{zcr} = \frac{1}{T-1}\sum_{t=1}^{T-1}\pi\{\underbrace{S_t S_{t-1} < 0}\}$$

The value of the sample point

T is frame length

$\pi\{A\}$ is 1 when A is true, otherwise 0

a) Short-Term Energy

$$E_n = \sum_{m=-\infty}^{+\infty}[S(m)W(n-m)]^2$$

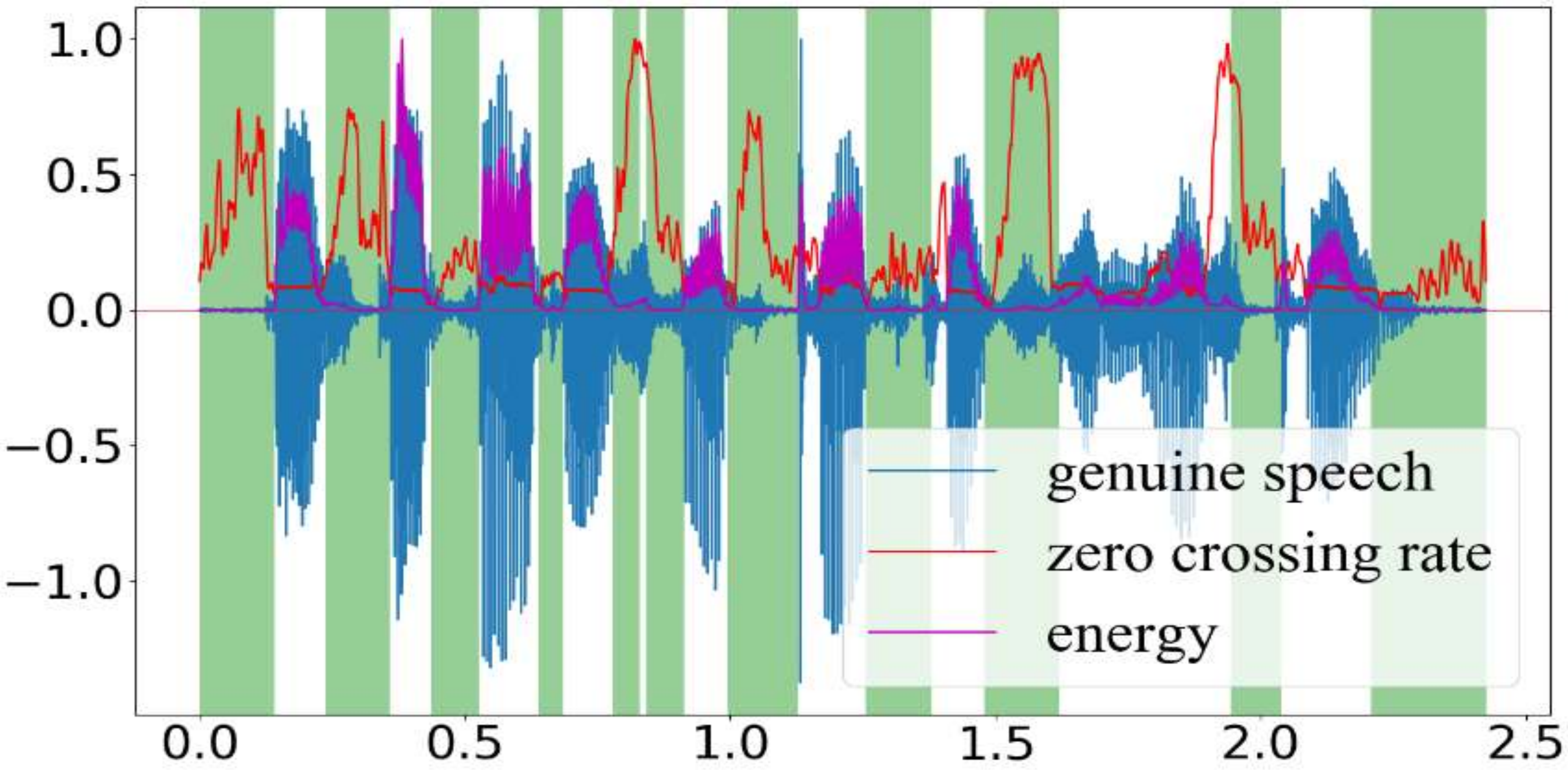
$$= \sum_{m=n-(N-1)}^{n}[S(m)W(n-m)]^2$$

N is the window length

W is a Hamming window

Segment-based Hybrid Feature Extraction(Cont.)

- Short-term zero-crossing rate and energy segmentation



Segment-based Hybrid Feature Extraction(Cont.)

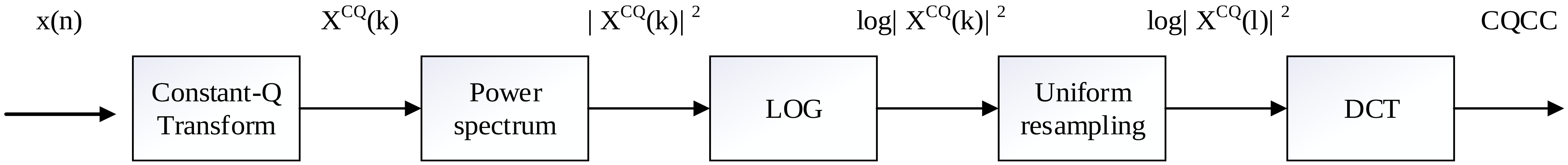
- MFCC Features Extraction

a) Mel Frequency Cepstral Coefficients:

- pre-emphasis (sub-frame and adding window)
- Fast Fourier Transform
- absolute value and square operation
- Mel-scaled triangle filters
- a Logarithmic operation
- Discrete cosine transform

b) MFCC feature is extracted in the approximate silent segment

Segment-based Hybrid Feature Extraction(Cont.)



- CQCC Features Extraction
- It provides a higher frequency resolution for low frequencies and a higher time resolution for high frequencies

The Proposed DenseNet-LSTM Classifier

➢ CNN, DNN and RNN Architectures

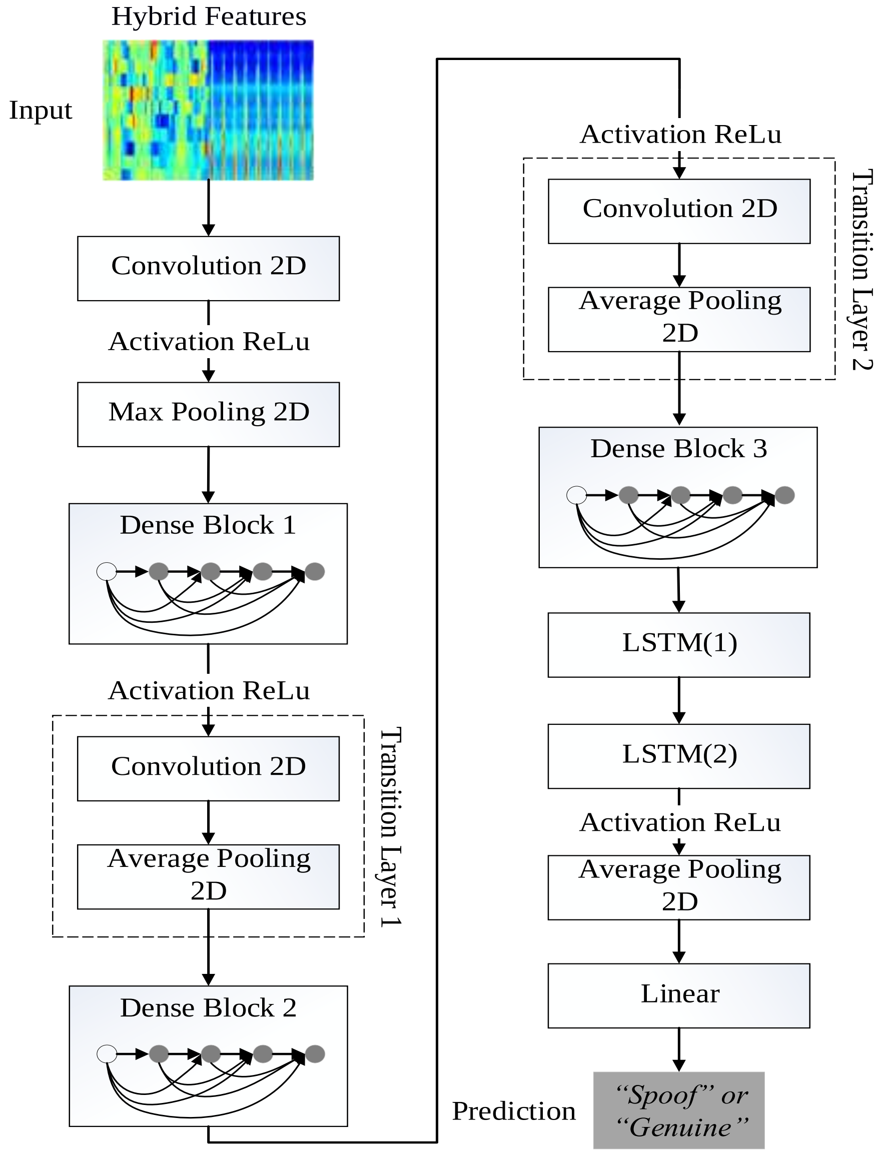
- CNN and DNN particularly dependent on the availability of large quantities of training data.
- The replay spoof audio dataset is smaller than other datasets
- LSTM can extract more useful information
- The use of CNN, DNN or RNN architecture directly in the replay spoof attack detection does not yield convincing results

The Proposed DenseNet-LSTM Classifier(Cont.)

- Dense Convolutional Network (DenseNet)
    - Create short paths from early layers to later layers (ResNet)
    - Each layer takes all preceding feature-maps as input (Feature Reuse)
    - Reduce Vanishing-gradient
    - Good resistance to overfitting without enough samples

The Proposed DenseNet-LSTM Classifier(Cont.)

- DenseNet-LSTM Classifier
- Modify the last layer of DenseNet and add two more LSTM layers before the linear layer



Experimental Results

- Dataset: BTAS2016 & ASVspoof 2017
- All audio signals have a resolution of 16 bits and a sampling rate of 16 kHz
  - BTAS2016 dataset (selected the spoof type of replay)

Biometrics Theory Applications and Systems 2016

Subsets	#utterances	
	#genuine	#replay
Train	4973	2800
Dev	4995	2800
Eval	5576	4800
Total	15544	10400

Experimental Results (Cont.)

- The ASVspoof 2017 dataset
  - The dataset is only for replay spoof attack
  - The replay speech is re-recorded using different recording devices in different acoustic environments.

2017 ‘Automatic Speaker Verification Spoofing and Countermeasures Challenge’

Subsets	#speakers	#utterances	
		#genuine	#replays
Training	10	1508	1508
Devel.	8	760	950
Eval.	24	1298	12008
Total	42	3566	14466

Experimental Results (Cont.)

- Baseline system
  - Based on the CQCC feature and the Gaussian mixture model (GMM)
  - It focuses on estimating a likelihood ratio

$$\omega(u) = \frac{P(u|H_g)}{P(u|H_s)}$$

For a speech utterance u, it will decide whether u belongs to the genuine speech H_g or to the spoof speech H_s

Experimental Results (Cont.)

- Metrics
  - Evaluating the system by Equal Error Rate (EER)
  - EER is the error rate when a certain threshold θ is taken and FRR (false rejection rate) == FAR (false acceptance rate), FAR and FRR can be calculated using θ as follows:

$$FAR(\theta) = \frac{\#\{replay\ trials\ with\ score > \theta\}}{\#\{Total\ replay\ trials\}}$$

$$FRR(\theta) = \frac{\#\{non\!-\!replay\ trials\ with\ score \leq \theta\}}{\#\{Total\ non\!-\!replay\ trials\}}$$

EER corresponds to the threshold θ_{EER} at which EER == FAR(θ_{EER}) == FRR(θ_{EER})
(To be determined in development dataset)

Part 4

Experimental Results (Cont.)

- Details of DenseNet-LSTM architecture

Layers	Output Size	Layer config
Convolution	30×63	7×7 conv, stride 2
Pooling	15×32	3×3 max pool, stride 2
Dense Block 1	15×32	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
Transition Layer 1	15×32	1×1 conv， stride 1
	7×16	2×2 average pool, stride 2
Dense Block 2	7×16	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
Transition Layer 2	7×16	1×1 conv， stride 1
	3×8	2×2 average pool, stride 2
Dense Block 3	3×8	$\begin{bmatrix} 1\times 1\ conv \\ 3\times 3\ conv \end{bmatrix}\times 6$
LSTM	1×48	2 Layers
Classification Layer	1×1	1×48 global average pool
	（None,1）	Linear

Experimental Results (Cont.)

- Evaluation results in ASVspoof 2017 Dataset

Individual System	EER(%)		
	Dev set	Eval set(T)	Eval set(T+D)
Baseline (CQCC+GMM)	10.83	30.60	24.77
CQCC+DenseNet[1]	7.65	17.73	15.27
MFCC+DenseNet[1]	6.77	15.86	13.45
CQCC+DenseNet-LSTM[1]	6.87	12.64	11.67
CQCC+DNN[2]	5.18	19.41	-
CQCC+ResNet[2]	5.05	18.79	-
MFCC+ResNet[2]	10.95	16.26	-
CQCC GMM+MFCC ResNet+CQCC ResNet[2]	**2.58**	13.30	-
Hybrid Feature+GMM[3]	8.67	25.63	18.11
Hybrid Feature+DenseNet[3]	5.62	12.39	11.08
Hybrid Feature+LSTM[3]	9.45	15.64	14.78
Hybrid Feature+DenseNet-LSTM (Proposed)[3]	3.32	**9.56**	**8.84**

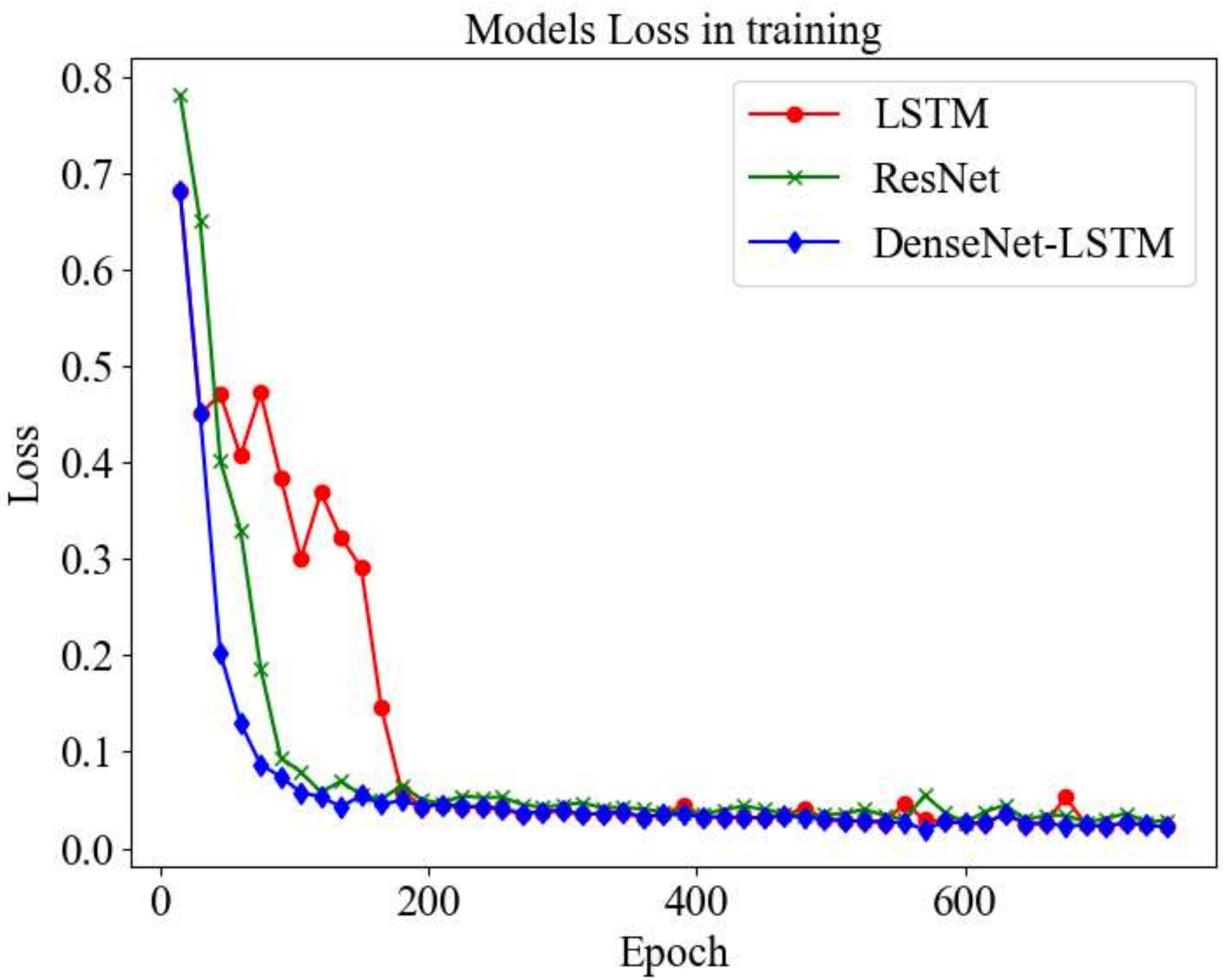
Experimental Results (Cont.)

- Evaluation results in BTAS2016 Dataset

Individual System	EER(%)		
	Dev set	Eval set(T)	Eval set(T+D)
Baseline (CQCC+GMM)	2.36	8.42	6.37
CQCC + DenseNet[1]	1.32	1.67	1.24
MFCC + DenseNet[1]	0.26	1.53	1.20
CQCC + DenseNet-LSTM[1]	0.42	1.35	1.12
CQCC+DNN[2]	1.25	2.08	-
CQCC+ResNet[2]	1.18	1.87	-
MFCC+ResNet[2]	1.12	1.97	-
CQCC GMM+MFCC ResNet + CQCC ResNet[2]	0.89	1.27	-
Hybrid Feature + GMM[3]	1.87	7.92	5.43
Hybrid Feature + DenseNet[3]	**0.04**	1.42	1.34
Hybrid Feature + LSTM[3]	0.28	2.25	1.25
Hybrid Feature+DenseNet-LSTM (Proposed)[3]	0.31	**0.96**	**0.89**

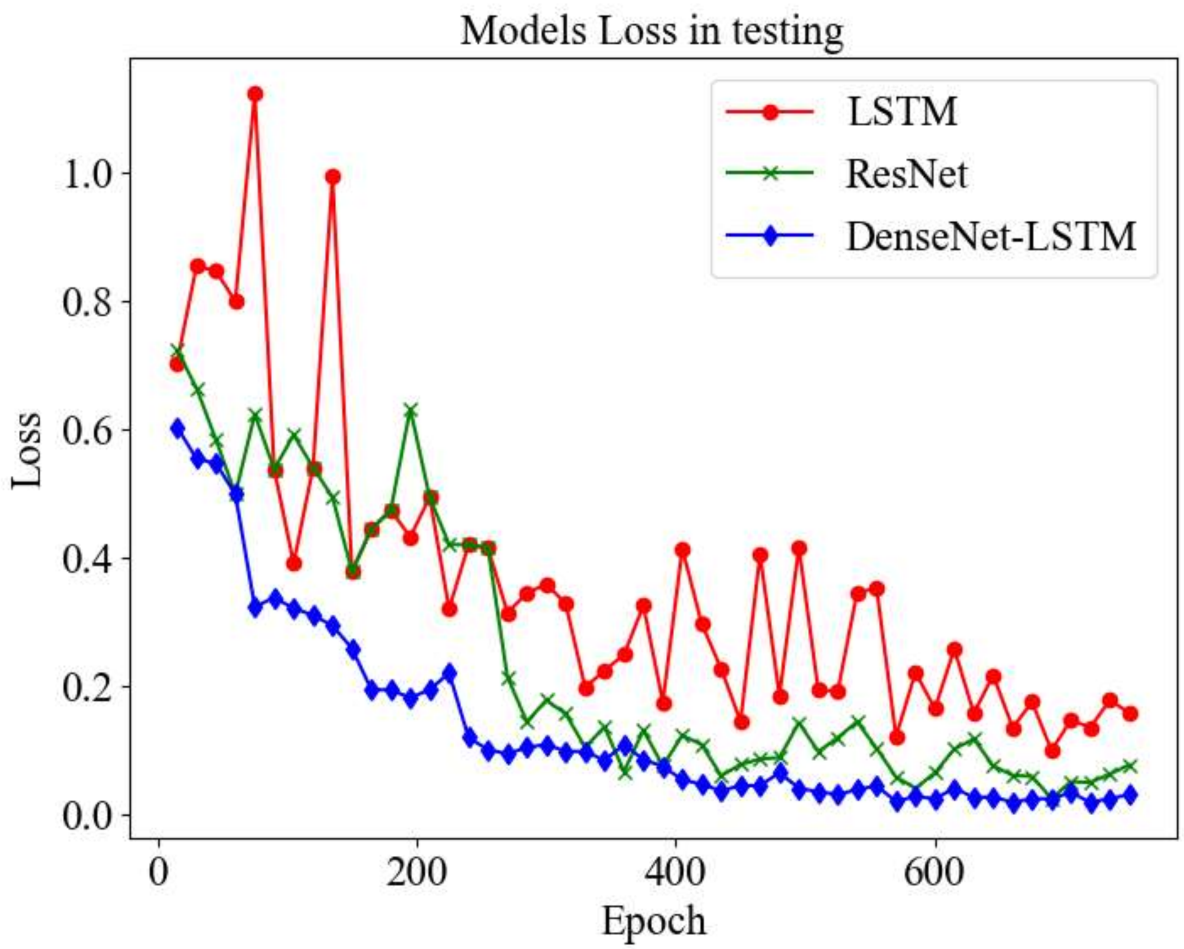
Experimental Results (Cont.)

- Models loss in training



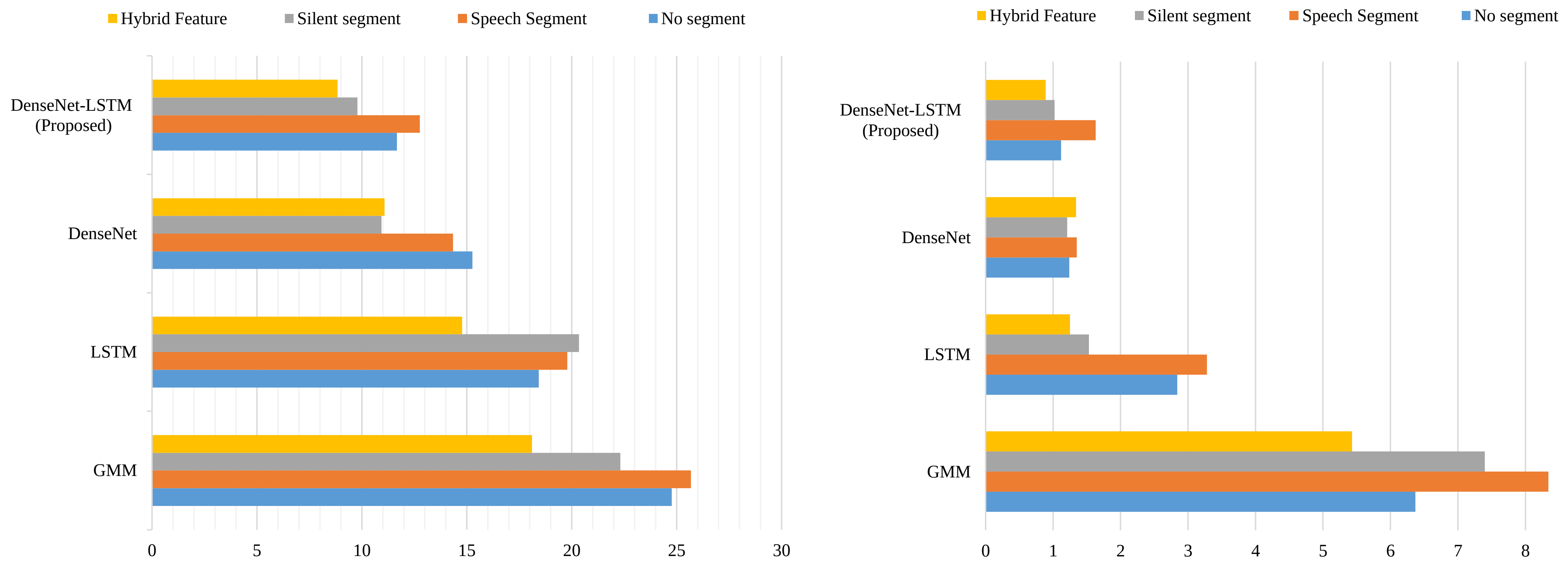
Experimental Results (Cont.)

➢ Models loss in testing



Experimental Results (Cont.)

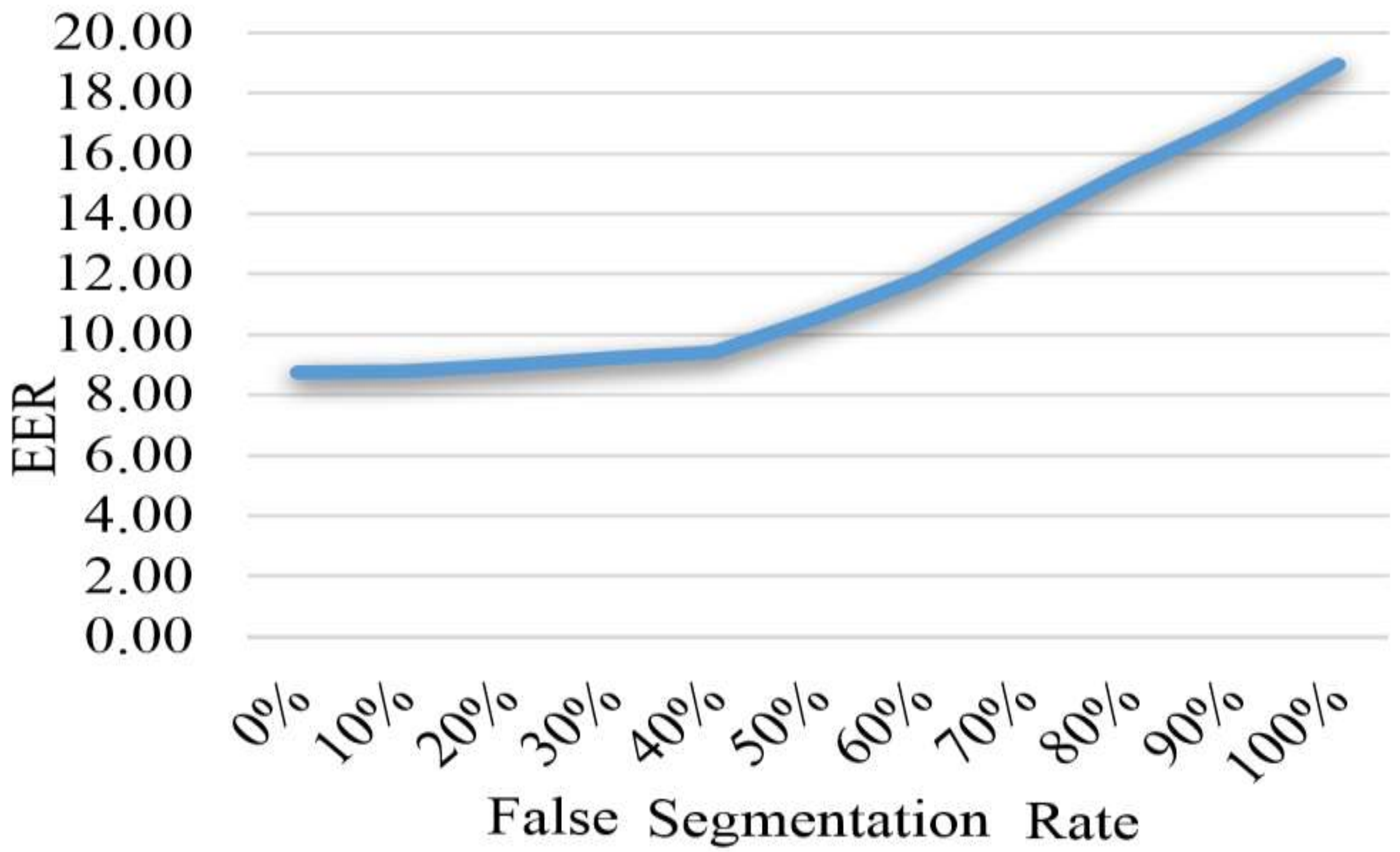
➢ Segmentation's effect on EER



The EER based on CQCC with different models. (left) Evaluation results in ASVspoof 2017 Dataset. (right) Evaluation results in BTAS2016 Dataset.

Experimental Results (Cont.)

- The impact of false segmentation rate on the EER



Conclusion and Future Work

Conclusion:

- Proposed a novel feature extraction method: Segment-based Hybrid Feature Extraction.
- The hybrid feature extraction method emphasizes the background noise characteristics and the results show that it can improve the performance in detecting replay spoofed speech.
- The proposed DenseNet-LSTM classifier enhance the classification accuracy.
- The DenseNet-LSTM classifier can reduce the overfitting problem.

Conclusion and Future Work (Cont.)

Future work:

- With high-quality hardware in replay spoofing, we may need to explore in the future work.
- Further research with raw-wave and end-to-end approach may simplify the detection process.
- It may need to explore a detection method with better generalization capability.

Publication

- Lian Huang and Chi-Man Pun, “Audio replay spoof attack detection using segment-based hybrid feature and densenet-LSTM network,” *in IEEE ICASSP* 2019, pp. 2567–2571.
- L. Huang and C.-M. Pun, “Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28(6), pp. 1813 – 1825, 2020.

Thank You!

Q&A