AN EXPERIMENTAL STUDY ON AUDIO REPLAY ATTACK DETECTION USING DEEP NEURAL NETWORKS

Bekir Bakar, Cemal Hanilçi
Bursa Technical University
Department of Electrical and Electronics Engineering
Bursa, Turkey

Replay Attack Detection

- Given a speech signal S, spoofing detection is a hypothesis test:
  - ➢ H_0 **: S is genuine speech signal**
  - ➢ H_1 **: S is re-played/spoof speech signal**
- Using the feature vectors extracted from speech signal S, log-likelihood ratio (***LLR***) test can be applied to decide between two hypothesis:
  - ➢ $LLR(X) = logp(X|\lambda_{H_0}) - logp(X|\lambda_{H_1})$
- $X = \{x_1, x_2, \dots x_T\}$ are the feature vectors.
- λ_{H_0} ***and*** λ_{H_1} are the acoustic models representing natural/genuine and re-played/spoof speech classes.

Features

- **CQCC**
  - ➢ *18 dimensional* $(c_1 - c_{19})$ static ***constant Q cepstral coefficients*** [1] *are* computed using ***constant-Q transform.***
  - ➢ Greater time resolution for the higher frequencies and greater frequency resolution for the lower frequencies.
- **MFCC**
  - ➢ ***Mel-frequency cepstral coefficients*** are extracted from the **pre-emphasized** speech signal [2].
  - ➢ 20 ms frames in 10 ms overlap, Hamming window.
  - ➢ ***27-channel mel-filterbank***.
- **LTAS**
  - ➢ The **514** dimensionel **long-term** average spectrum [3].
  - ➢ Represents an utterance in a **long-term** rather than **short-term**.
  - ➢ 20 ms frames in 10 ms overlap.
  - ➢ 512 point discrete Fourier transform
  - ➢ Concatenation of the mean and standard deviation statistics of logarithmic magnitude spectrum.

Classifiers

- **GMM**
  - ➢ 512 component Gaussian mixture model.
  - ➢ Expectation maximization (EM) algorithm.
  - ➢ Genuine and replay clases (λ_{replay}, λ_{genine}) replay attack detection score is computed as:

$$LLR_{GMM}(X) = logp(X|\lambda_{genuine}) - logp(X|\lambda_{replay})$$

- **DNN**
  - ➢ Fully-connected feed-forward neural network.
  - ➢ Batch training, batch normalization and dropout.
  - ➢ The softmax function.
  - ➢ The posteriors obtained at the output layer (two units) of deep neural network is transformed into LLR score as:

$$LLR_{GMM}(X) = logp(X|\lambda_{genuine}) - logp(X|\lambda_{replay})$$

	CQCC	MFCC	LTAS
Input Size	18	57	514
Hidden Layers	3	3	5
Unit Size	256	256	1024
Dropout	0.2	0.2	0.5

Optimized empirically.

These values are highly depend on data (features).

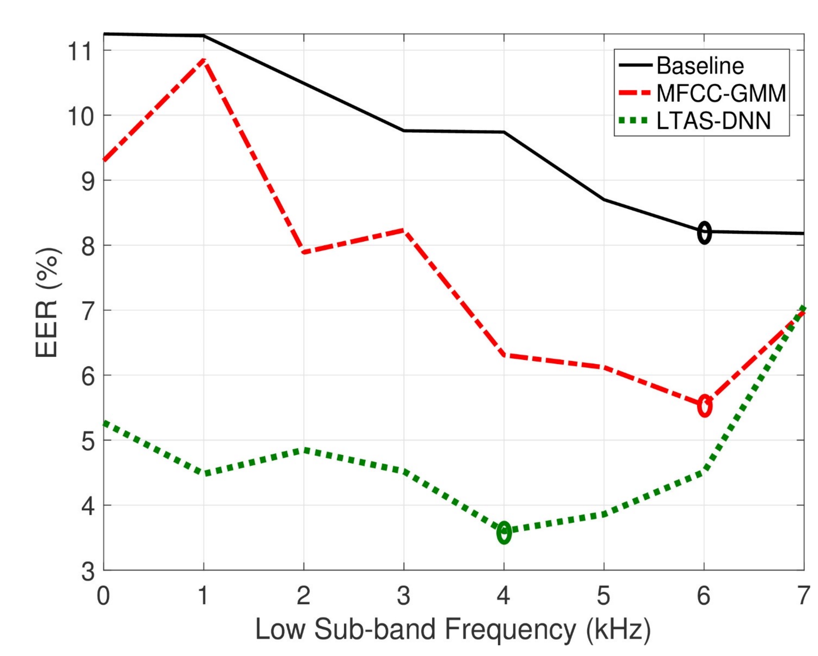
Results

- **Development Set**

System	EER (%)
CQCC-GMM	8.18
MFCC-GMM	5.54
CQCC-DNN	10.05
MFCC-DNN	6.44
LTAS-DNN	**4.10**

System	EER (%)
$CQCC_{CMVN}$	15.15
$MFCC_{CMVN}$	13.40
$CQCC_{CMVN}$	17.18
$MFCC_{CMVN}$	12.51
$LTAS_{CMVN}$	**6.05**

Short-term are inferior to long-term features independent of the classifier.



Cepstral mean and variance normalization reduces performance

High-frequency components convey more discriminative information.

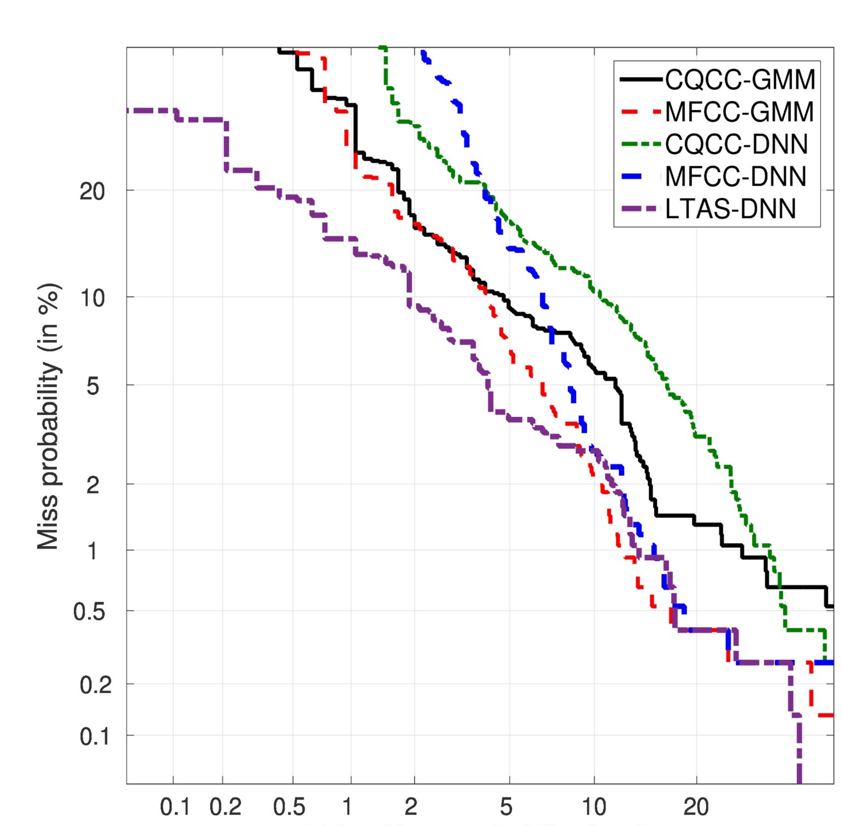
- **Evaluation Set**

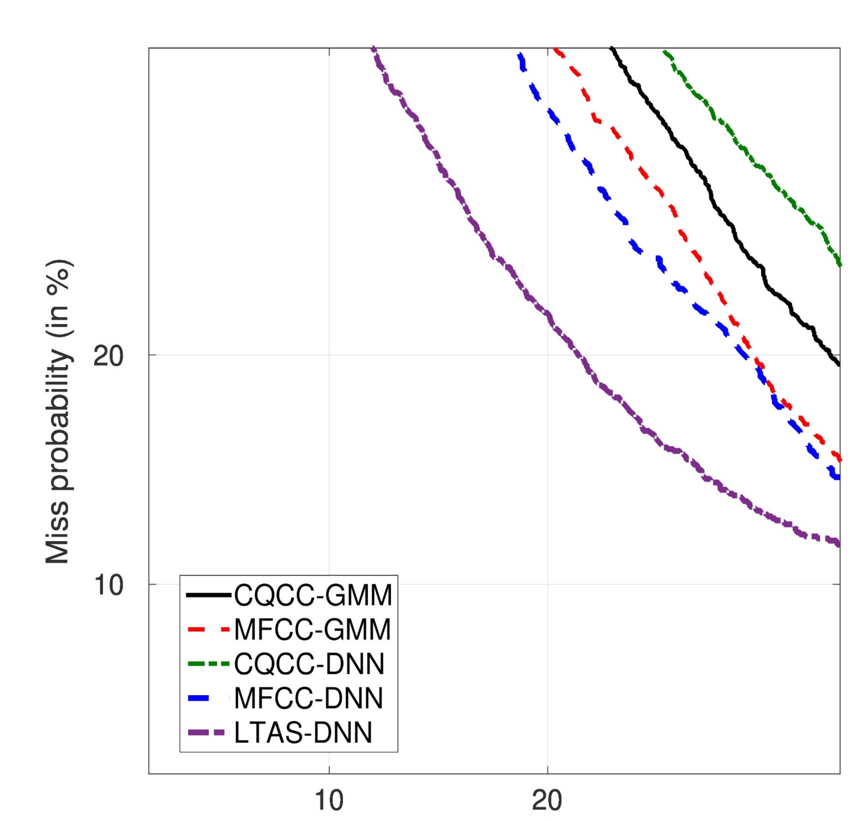
System	EER (%)
$CQCC_{CMVN}$	15.15
$MFCC_{CMVN}$	13.40
$CQCC_{CMVN}$	17.18
$MFCC_{CMVN}$	12.51
$LTAS_{CMVN}$	**6.05**

DNN slightly improves the performans in comparison to GMM classifier for MFCC features.

GMM classifier is superior to DNN back-end for CQCC features.

- **Det Curves**





References

[1] Massimiliano Todisco, Hector Delgado, and Nicholas Evans, “A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients,” in Proc. Speaker Odyssey Workshop, 2016, pp. 249–252.

[2] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 357–366, Aug. 1980.

[3] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “Presentation attack detection using long-term spectral statistics for trustworthy speaker verification,” in Proc. BIOSIG, 2016, pp. 1–6.

[4] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, ”Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.*