A new approach for robust replay spoof detection in ASV systems

Shaik Mohammad Rafi B, Sri Rama Murty K, Shekhar Nayak

Department of Electrical Engineering,
Indian Institute of Technology, Hyderabad, India.

Introduction to Spoofing attacks

Manipulation of a biometric system by a fraudster impersonating another enrolled person through voice conversion, replaying, speech synthesis, mimicry etc.

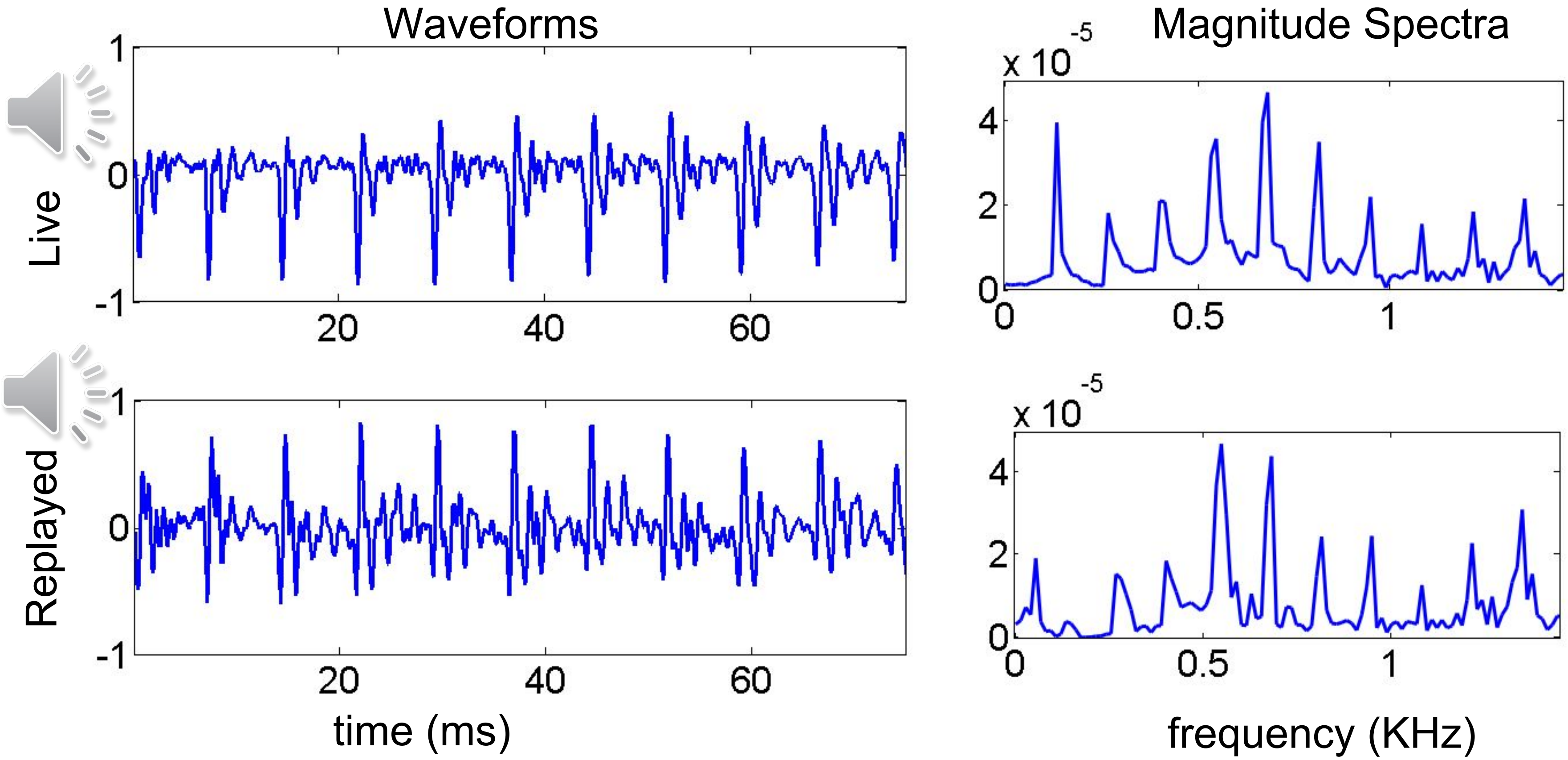
Focus on Replay attacks:

Playing pre-recorded voice of the target speaker to gain unauthorized access to the secured information.

- With the widespread use of mobile devices, it is easy to record the target speaker's voice, without his/her knowledge making replay spoof attacks easier to implement.
- It also does not require any signal processing knowledge.

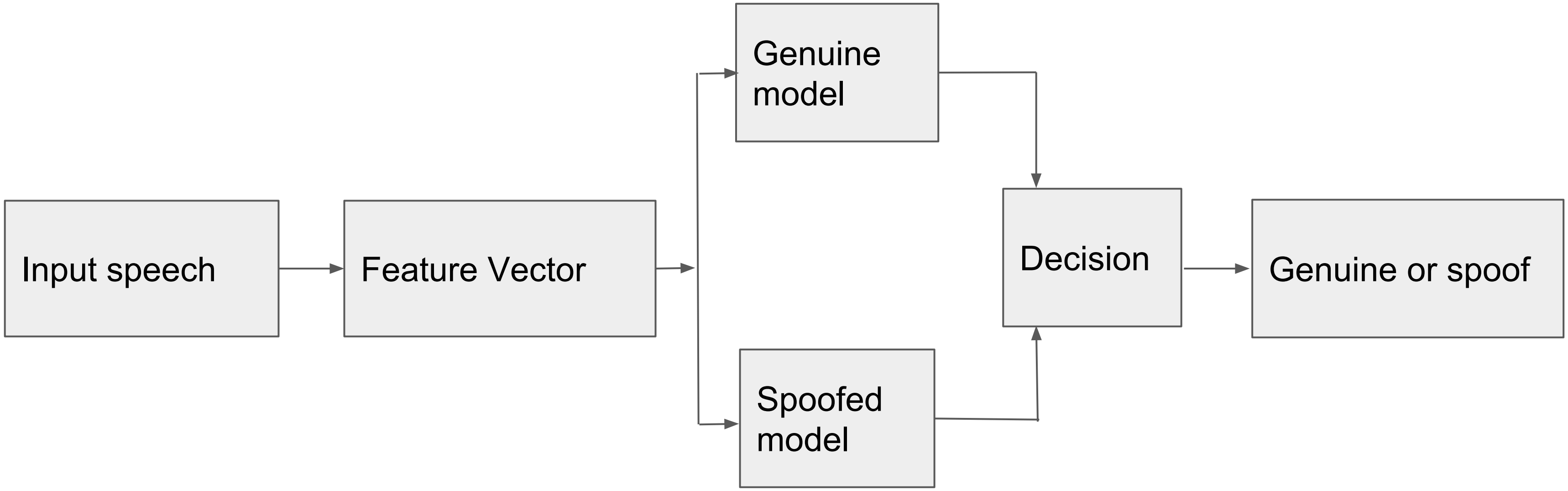
Motivation

The distortion is inherent in the replayed signal and is reflected as lower damping effect around glottal closure regions.



Spoofing Counter Measures

- Aim to distinguish between natural and spoofed speech.
- Artefact detection encompassing relatively standard **feature extraction and statistical pattern recognition** techniques.
- Capturing the **obvious signs** of manipulation.
- Design of spoofing countermeasures should better **focus on feature engineering**.

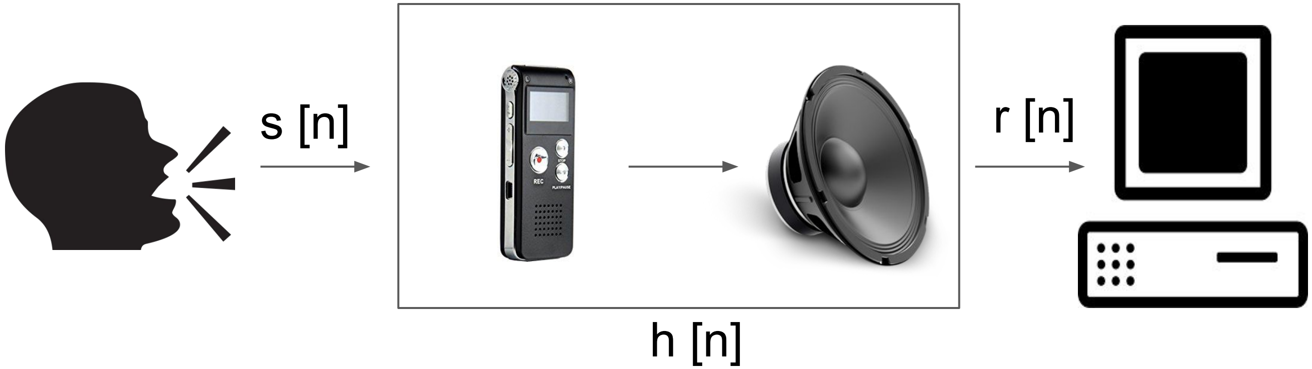


Spoof Detection System

Proposed method

Replay spoof detection is the task identifying whether a given speech utterance is recorded from a live speaker or an **intermediate (recording+playback) device**.

Assuming intermediate device is linear and time invariant, $r[n] = s[n] * h[n]$

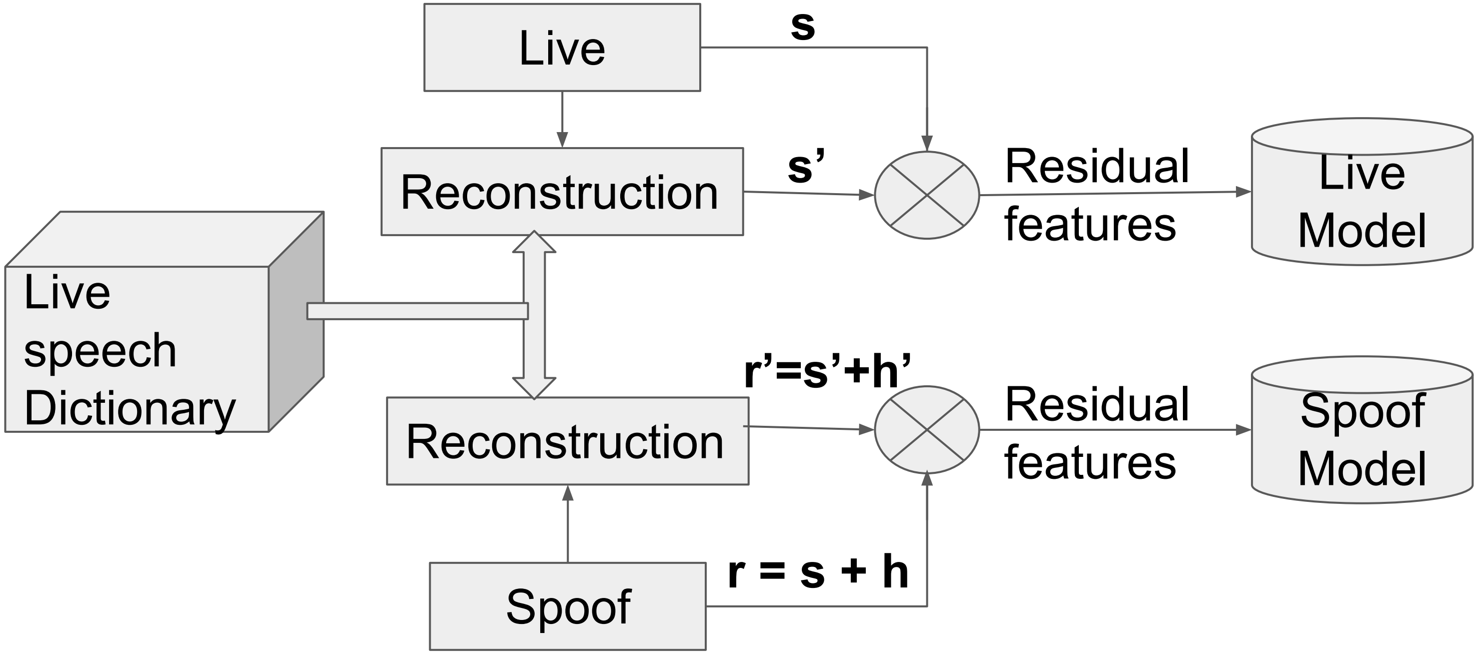


The intermediate device introduces **convolution distortion** and hence features representing the characteristics of the intermediate device should be extracted in order to discriminate live speech signal from replayed speech signal.

Proposed method

The convolutive relationship between live speech and intermediate device characteristics, transforms to an additive relationship in the cepstral domain

Convolution in time	Multiplication in frequency	Addition in Cepstral
r[n] = s[n] * h[n]	R = S H	**r = s + h**



Proposed method

Given the set of live speech cepstral coefficients **S**, the dictionary of atoms **A** can be estimated by minimizing the following objective function

$$\min_{\mathbf{A},\mathbf{Y}} \|\mathbf{S} - \mathbf{A}\mathbf{Y}\|_F^2 \text{ , s.t } \|\mathbf{y}_i\|_0 \leq \tau.$$

Where $\|.\|_F$ denotes Frobenius norm, $\|.\|_0$ denotes number of nonzero elements and τ denotes sparsity constraint parameter.

The K-singular value decomposition (K-SVD) algorithm is commonly used for joint estimation of dictionary **A**, and sparse weights **Y**.

The dictionary **A** approximates the cepstral coefficients from live speech better than the replay speech. Hence the residual (deviation from Live) vector

$$\mathbf{e} = \mathbf{x} - \mathbf{A}\mathbf{y} \quad \text{subject to} \quad ||\mathbf{y}||_0 = \tau$$

can be used as a feature for detecting replay speech

Experiments

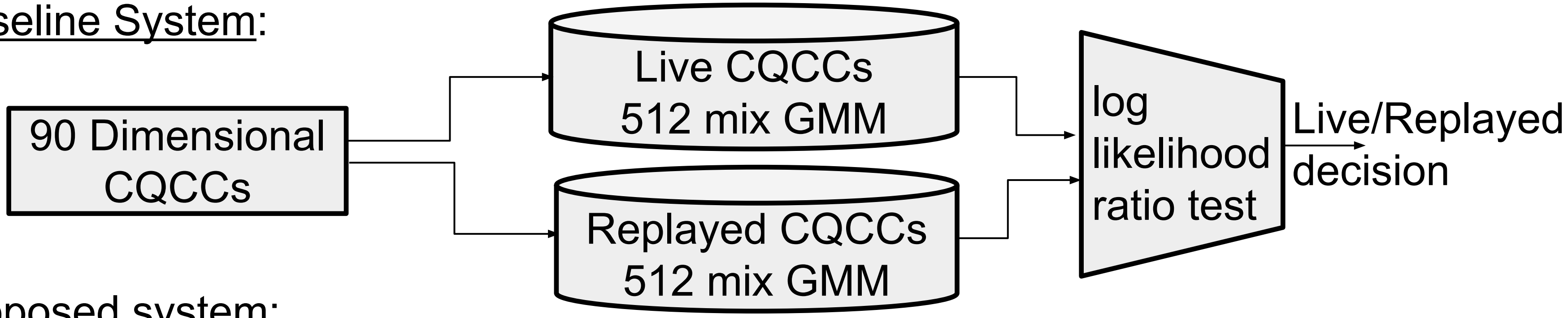
The performance of the proposed approach for replay spoof detection is evaluated on ASVspoof 2017 corpus.

The replay speech data is collected by playing back the RedDots utterances and re-recording with heterogeneous devices in and diverse acoustic environments.

Subset	# Speakers	# Live Utterances	# Replay Utterances
Train	10	1508	1508
Dev	8	760	940
Eval	24	1298	12922

Experiments

Baseline System:



90 Dimensional CQCCs

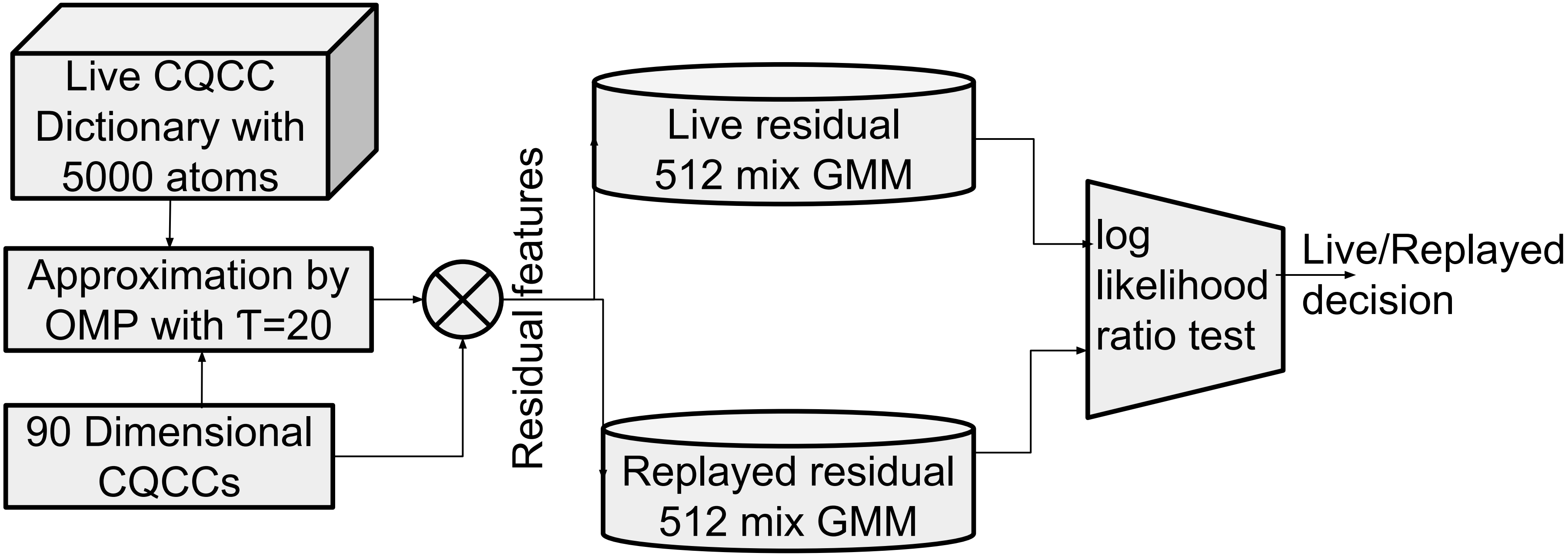
Live CQCCs 512 mix GMM

Replayed CQCCs 512 mix GMM

log likelihood ratio test

Live/Replayed decision

Proposed system:



Live CQCC Dictionary with 5000 atoms

Approximation by OMP with T=20

90 Dimensional CQCCs

Residual features

Live residual 512 mix GMM

Replayed residual 512 mix GMM

log likelihood ratio test

Live/Replayed decision

Results and Discussions

Two different systems are trained, one with only the data from train (T) subset, the the other with combined data from train and dev (T+D) subsets.

Feature & Modelling	**EER (T)**	**EER (T+D)**
CQCC & GMM (baseline system)	30.60	24.77
CQCC-KSVD & GMM	24.14	22.45
Score Fusion	23.39	19.77

Results and Discussions

- The proposed method exploits convolutional distortion introduced by the intermediate device.

- This approach only requires live data and does not require replay data for learning dictionary and hence its independent of intermediate devices.

- Most of low quality hardware devices introduce significant convolutional distortion, and hence can be detected in the proposed approach.

- However, recordings made with high quality hardware in benign acoustic environment may be indistinguishable from the live audio recordings.

- In such cases, we cannot rely only on the convolutional distortion introduced in the magnitude spectral domain alone and the distortion introduced in the phase domain also needs to be explored.

THANK YOU