Importance of Analytic Phase of the Speech Signal for Detecting Replay Attacks in Automatic Speaker Verification Systems

Shaik Mohammad Rafi B, K. Sri Rama Murty
ee17resch01003@iith.ac.in, ksrm@iith.ac.in

Speech Information Processing (SIP) lab, Department of Electrical Engineering, IIT Hyderabad, India.

Objectives

- Verify whether the given speech utterance is collected from a live human or playback device
- Playback device characteristics can be exploited to detect spoof attacks

Instantaneous Frequency Feature Extraction

- The analytic signal of a continuous time signal $s(t)$ is

$$s_a(t) = s(t) + js_h(t)$$

where $s_h(t) = \frac{1}{\pi t} * s(t)$.

$$s_a(t) = |s_a(t)|exp(j\phi(t))$$

- Instantaneous frequency (IF) is the time-derivative of the unwrapped instantaneous phase of $s_a(t)$.
- IF can be computed from the Fourier transform relations as

$$\phi'(t) = \frac{d\phi(t)}{dt} = Im\left\{\frac{s_a'(t)}{s_a(t)}\right\}$$

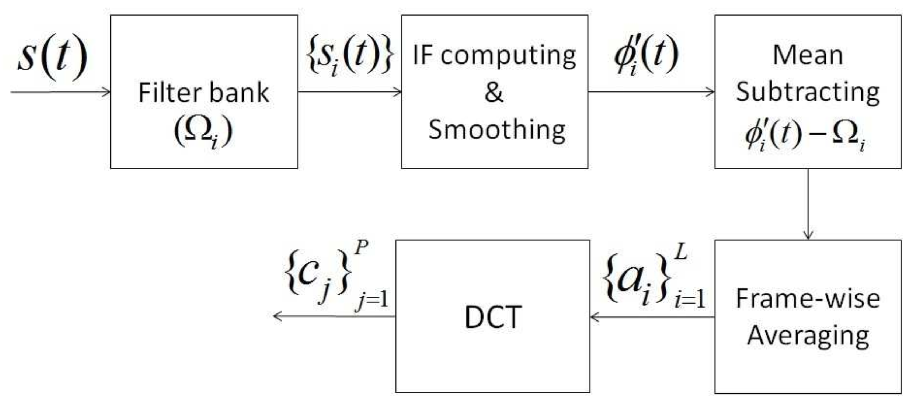
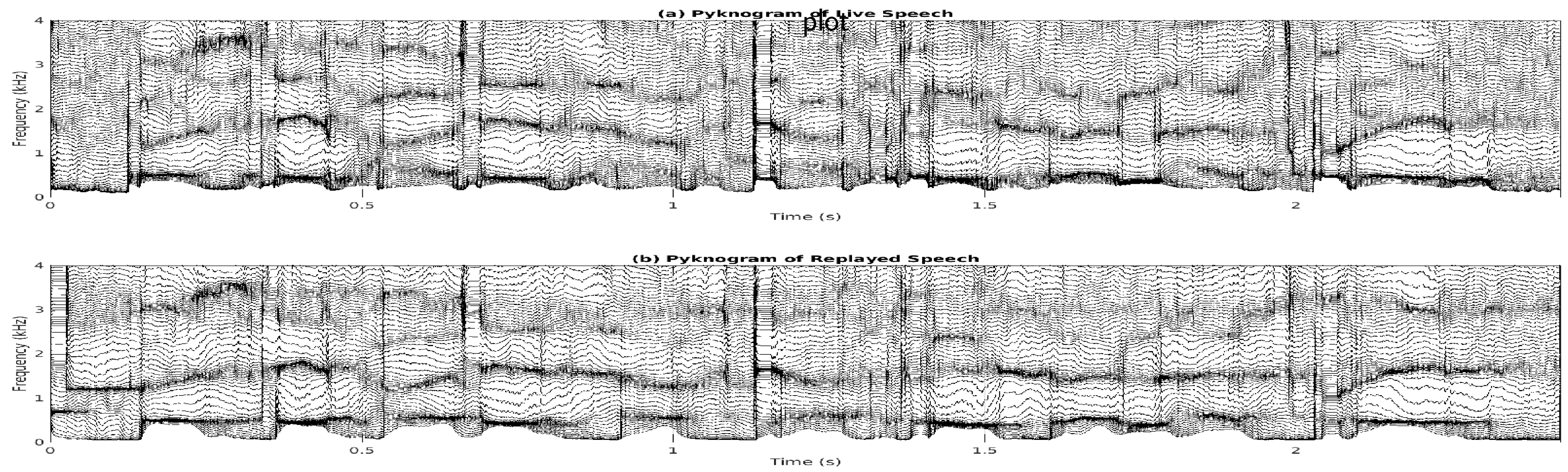


Figure 1: Instantaneous Frequency Cosine Coefficients (IFCC) features extraction

Device characteristics extraction

- Playback device introduces convolutional distortion to replayed speech
- It is manifested as additive distortion in the phase domain $\mathbf{r} = \mathbf{s} + \mathbf{h}$

$\mathbf{r}$, $\mathbf{s}$ and $\mathbf{h}$ denote the phase features of replayed speech, live speech and playback device, respectively.



- An overcomplete dictionary $\mathbf{A}$ is trained on live speech s so that it approximates live better than replayed speech.
- K-SVD dictionary learning algorithm involves two steps
  - Sparse Coding: For the given features $\mathbf{y}$
    - Initialize the dictionary $\mathbf{A}$ randomly
    - Find the best k-sparse vector $\mathbf{x}$ such that

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}.$$

    using the orthogonal matching pursuit (OMP) algorithm.
  - Update the atoms of $\mathbf{A}$ by optimizing

$$\min_{A,x} \|y - Ax\|_F^2, \text{s.t} \|x\|_0 \leq k.$$

- The residual error vector in this approximation is $\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}$
- The dictionary approximates of live speech better than the replayed speech, hence the residual error can be used as a feature for spoof detection.

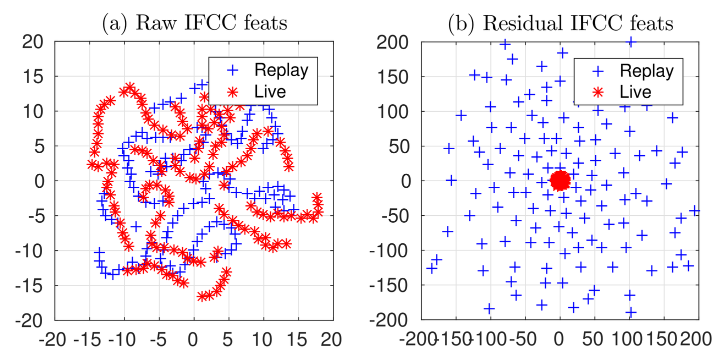


Figure 2: Highlighting device-specific characteristics by t-SNE plot

Experimental Evaluations

- The residual live and replayed features are modelled with GMMs.
- The experiments are evaluated on ASVspoof2017 challenge dataset.
- Baseline system: Constant-Q Cepstral Coefficients(CQCCs) of Live and replayed speech are modelled with GMMs.

Feature	Raw Features	Residual Features
CQCC	24.65*	22.45
MFCC	30.48	21.4
MGDC	30.00	34.5
IFCC	**23.44**	**15.00**
MFCC + IFCC	-	**13.99**

Conclusions

- IFCCs capture acoustic variations in live and replayed speech.
- The dictionary learns the contribution of live speech which helps in discriminating from replayed speech.
- IFCC features perform better than magnitude based features (MFCCs & CQCCs) and also other phase based features (MGDCs).

References

[1] Karthika Vijayan, Vinay Kumar, and K Sri Rama Murty. Feature extraction from analytic phase of speech signals for speaker verification. In *INTERSPEECH*, pages 1658–1662, 2014.

[2] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[3] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. *Proc. Interspeech 2017*, pages 2–6, 2017.