**REPLAY AND SYNTHETIC SPEECH DETECTION WITH RES2NET ARCHITECTURE**

1. **Key Technical Innovations:**

* This technique utilizes Res2Net for synthetic speech detection
* Res2Net focuses on the revision of the ResNet block to enable multiple feature scales
* It splits the feature maps within one block into multiple channel groups and designs a residual-like connection across different channel groups.
* Such residual-like connection increases the possible receptive fields, resulting in multiple feature scales.
* This additional multi-scale reception improves the system’s capacity and helps the system perform better when generalized to unseen spoofing attacks.
* It also decreases the model size compared to traditional ResNet-based models.
* Integration with the squeeze-and-excitation (SE) block can further enhance performance.
* For feature engineering, the generalizability of the Res2Net model that incorporates different acoustic features, and observe that the constant-Q transform (CQT) achieves the most promising results in both PA and LA subsets of ASVspoof 2019.

1. **Reported performance metrics:**

* Upon comparing ResNet34 and ResNet50, we observe that ResNet34 outperforms ResNet50 in all conditions, which indicates that simply increasing the depth is not efficient for generalizability enhancement.
* Res2Net50 significantly outperforms both ResNet34 and ResNet50.
* Res2Net50 respectively outperforms ResNet34 and ResNet50 by a relative EER reduction of 37.7% and 42.8% on the PA evaluation set, and 20.9% and 29.3% on the LA evaluation set. Similar gains are also observed under the t-DCF metric,
* We also observe that when compared to ResNet34 and ResNet50, the model size of Res2Net50 is relatively reduced by 33.8% and 16.2%, respectively, which verifies the efficiency of the Res2Net architecture in detecting spoofing attacks.
* Furthermore, integration with the SE block further improves the performance for all model architectures.
* For Res2Net50, the SE block achieves a relative EER reduction of 18.7% and 36.9% for PA and LA evaluation set, respectively.
* For horizontal comparisons, SE Res2Net50 outperforms SE-ResNet34 and SE-ResNet50 by a relative EER reduction of 43.9% and 46.0% respectively on the PA evaluation set, and 38.8% and 43.3% respectively on the LA evaluation set. Similar gains under the t-DCF metric.

1. **Promising aspects of this approach:**

* Achieves promising performance in both Physical Access(PA) and Logical Access(LA) scenarios.
* Decreases the model size compared to ResNet-based models.
* Integration with the squeeze-and-excitation(SE) block further enhances performance.
* Works well for unseen spoofing attacks.
* Fewer parameters make it computationally efficient for faster inference times which is crucial for real-time or near real-time detection.

1. **Potential limitations or Challenges:**

* The effectiveness of Res2Net is highly dependent on the quality of features extracted (eg- CQT, LFCC, spectrograms).
* The scale dimensions s, block expansion, and other parameters needs careful tuning to achieve optimal performance.
* The evaluation is primarily conducted on controlled datasets like ASVspoof 2019 (PA and LA partitions). Real-world audio data with environmental noise and diverse attack methods may yield different results.