"Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World

Rahavee Prabakaran

The research paper addresses speech synthesis attacks that cause damage to both computer systems and humans. Due to advancements in Deep Learning , several tools are developed that produces synthetic speech spoken in a voice of a target speaker. The paper aims at highlighting the need to raise awareness and describe results of an analysis on this speech synthesis attacks. The DNN-speech synthesis system is described, which are either text-to-speech or voice conversion which produces a synthetic version of the target's voice. The setting for the analysis of DNN-based Synthesis systems is a "Zero-shot system" (with target's voice sample which is <5minutes) with focus on peer-reviewed, published paper with public code implementations and pre-trained models. Based on an experimenting with several synthesis systems, it was observed that such systems did not perform well to speakers who were not on the training dataset. So SV2TTs and AutoVC was chosen because they performed best in such situations. The SR systems used for measurement study are Resemblyzer, Microsoft Azure, WeChat and Amazon Alexa. The speaker recognition datasets used in the paper are VCTK, LibriSpeech, SpeechAccent and a Custom dataset. Later, the paper describes an experiment on finding out the vulnerability of the SR-Systems to such attacks. First the efficiency of non-DNN synthesis attacks against today's SR system is evaluated as a reference, then it is tested against Resemblyzer. Some of the crucial results of this experiment were that Azure is easily attacked by DNN-synthesized speech, the attacks against female speakers were uniform while using SV2TTS, there is a difference in the attack success rate between native English speakers (high) and non-native English speaker. The paper moves on to evaluating the impact of DNN-synthesized speech on humans by conducting two user studies covering static survey and trusted interaction settings. The user study A (an online survey) is conducted to compare how well participants can identify synthesis speech for voice samples with mix of real, fake speech and different levels of familiarity. The study settings resulted in DNN-synthesized speech fails to have attacks on human's identification, since the users were able to distinguish between real and fake accents. The User Study B was in a zoom setting when the humans were fooled by the synthesized voice. The crucial findings of such studies were that context and demographics impact the creditability of synthesized speech for the users. The paper then goes on to make it a point that there is a need for defenses and there is a need for more comprehensive investigation that is required. There is also a need for human-centric defenses against such attacks by designing authentication methods or evaluating a stronger defense that combines Void and Attack-VC.

Three strong points:

1.Experimental setup for analyzing the effect of synthesized speech on machines is explained in detail for every speech recognition system.

2.The paper tends to attend to all type of readers by explaining the basics of SR systems, the workflow of synthesized based voice spoofing attacks.

3.The baselines models and the metric tables regarding the performance of the models are self-explanatory.

Three weak points:

1.The impact of English speakers on the attacks were described , but impact on attacks by other languages were not addressed.

2.Only Zero-shot systems were considered for experiments. The effect of the attack when larger speech sample is taken is not addressed.

3.There was no experimental analysis for evaluating synthesized attacks on humans which makes the information given on the paper less efficient because user case studies cannot be trusted entirely for edge cases.