Test Against Alexa	$\phi = 0°$		
$d_t =$	$4.2ft$	$7.2ft$	$10.2ft$
Random Music generated by the Karplus-Strong(KS) algorithm	0/10	0/10	0/10
Random Notes generated by the Karplus-Strong(KS) (single note) algorithm	1/10	0/10	0/10

Table 1: Times of the real Amazon Alexa being able to respond to the wake-word under the influence of baseline noises.

Thanks to all the reviewers for your time and detailed comments! For the most part, we fully agree with many of the statements that the reviewers make about this paper. As was recognized by all, this paper is about introducing a denial-of-service (DoS) attack against voice assistant systems, and as such its motivation and value lies largely in the evident fact that such attacks are possible in the real world against a commercial-grade product, rather than in the algorithmic components.

-Reviewer #1 Thank you for your suggestion on problem setup, and we will modify our problem statement accordingly to focus on DoS attack (false-negative). To answer your question about the baseline, we experimented with two new sample audio generated by the same (Karplus-Strong) algorithm and tested against Alexa. The result is shown in Table.1. The musical audio does not fool Alexa. (Please see another demo video here: `https://youtu.be/TRHGpYzv_Sk`, uploaded anonymously. Though we understand if the reviewers are not able to take it into account for review, as the response guidelines mentioned not to reference external links.) As for the timing offset, we have been conducting a more rigorous quantitative analysis. If we were given a chance to present in the camera-ready version in the main conference, we would definitely include a thorough comparison with the different length of timing offset from alignment. So far, from our rough empirical evaluation and as you can see in the original demo video, we observe that our adversarial music does not have to be aligned with the wake word perfectly, but the audio tends to work the best while being looped. If the wake word starts first, it would be recognized by Alexa. If the noise starts first, the wake word is nullified. We will also include a digital experiment in our camera-ready version with specific digital simulation evaluating the performance of the model as a function of SNR. We will also add more physical experiments to study the trend of SNR VS accuracy in the real world. Thank you again for your constructive feedback!

-Reviewer #3 The Karplus-Strong (KS) algorithm is used here since the formulation, and implementation of the algorithm is fully open-sourced and differentiable with current configuration (controlling the frequency and the intensity of notes), which could be plugged nicely into our adversarial training paradigm. We expect other MIDI synthesizers could be effective given the right configurations as well. If we were given a chance to present our work at the main conference, we would add our result of our on-going effort trying to generate our adversarial music using Google Magenta Synthesizers Nsynth. As for your question about the baseline, please see our answer to Reviewer #1 and the additional demo video.

Currently, we are also trying to activate the wake-word using our adversary. So far, we observe false positives are way more challenging to produce than false-negatives if we were using the same music format in the same adversarial training paradigm. However, a distortion of the recorded wake-word could easily trigger the detection of false-positives, which makes false positive less attractive in the scope of our discussion. Thank you again for your encouragements and constructive feedback!

-Reviewer #4 Thank you again for your constructive feedback! We agree with your suggestion calling our model "gray box" attack. Please see our answer to R#3 for the motivation behind the KS algorithm. We conjecture that the simplicity of the KS algorithm also constrains us since the parameters and complexity heavily constrain the variability of our adversaries that the KS algorithm defines. This could be the main reason for the transient-heavy and high-volume sound. Currently, we are working on other synthesizers and other instruments to make our adversary sound more pleasing to humans. As of now, we observe guitar works the best. A more complex synthesizer (e.g., a neural-network-based synthesizer Nsynth) might be able to provide us more attacking budgets due to more degree of freedom with a lot more parameters in constructing the adversary. (even though itself is a black box) If given the opportunity for the camera-ready submission, we will try our best to provide more low-volume low-attack adversaries and explore the SNR threshold for our adversary to be effective as you suggested. Although from what we have observed, the universal VA adversary could be very difficult if not impossible to achieve in the real world, since the adversary itself needs to be robust against other environmental noise.