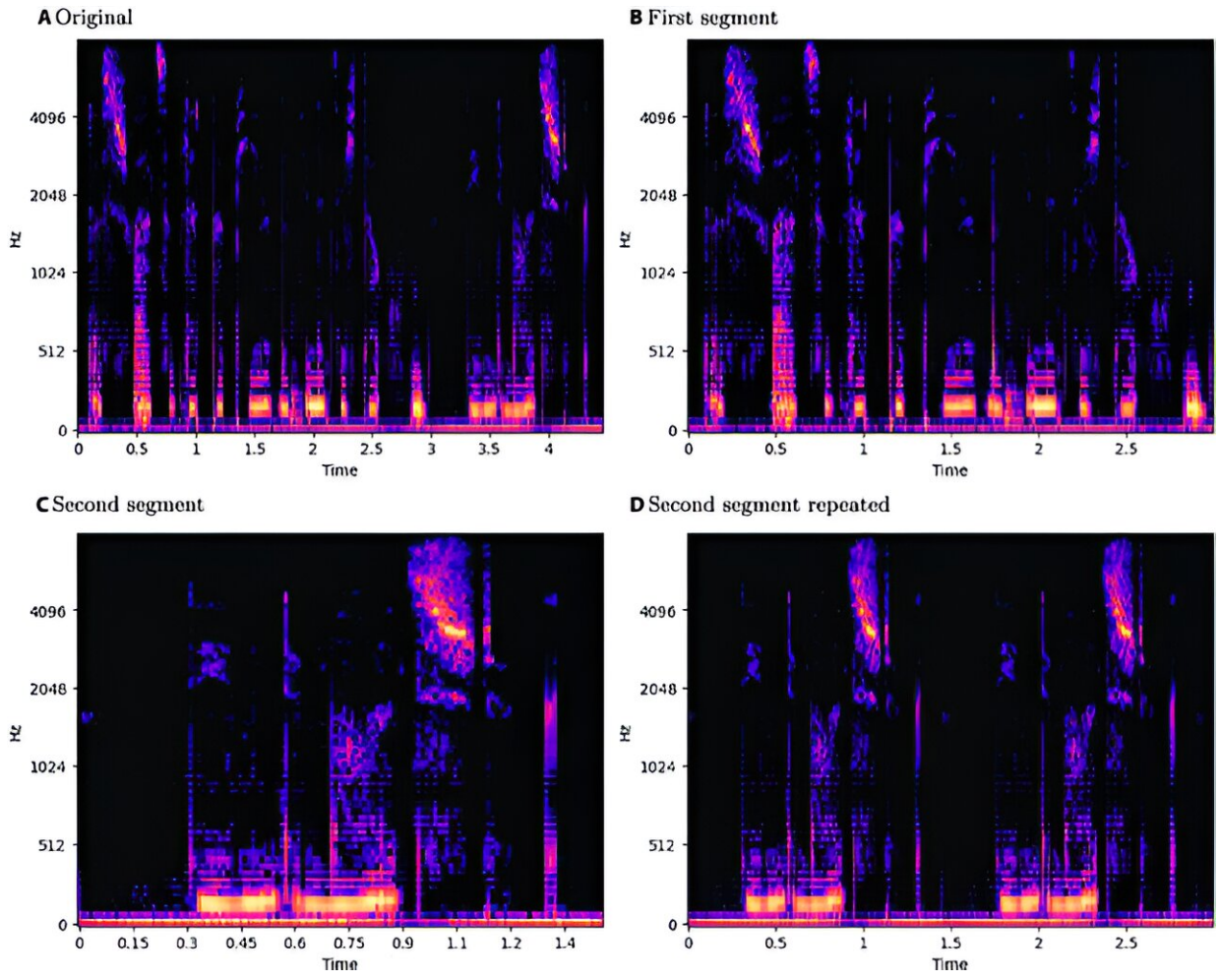
Researchers expose vulnerability of speech emotion recognition models to adversarial attacks

August 9 2024



Example of split and repeat process on log-Mel spectrograms. Original log-Mel spectrogram (A), the sliced segments (B and C), and segment (C) repeated to 3 s (D). Credit: *Intelligent Computing* (2024). DOI: 10.34133/icomputing.0088

Recent advancements in speech emotion recognition have highlighted the significant potential of deep learning technologies across various applications. However, these deep learning models are susceptible to adversarial attacks.

A team of researchers at the University of Milan systematically evaluated the impact of white-box and black-box attacks on different languages and genders within speech emotion recognition. The research was published May 27 in *Intelligent Computing*.

The research underscores the considerable vulnerability of convolutional neural network long short-term memory models to adversarial examples, which are carefully designed "perturbed" inputs that lead models to produce erroneous predictions. The findings indicate that all considered adversarial attacks can significantly reduce the performance of speech emotion recognition models. According to the authors, the susceptibility of these models to adversarial attacks "could trigger serious consequences."

The researchers proposed a methodology for audio data processing and feature extraction that is tailored to the convolutional neural network long short-term memory architecture. They examined three datasets, EmoDB for German, EMOVO for Italian and RAVDESS for English. They utilized the Fast Gradient Sign Method, the Basic Iterative Method, DeepFool, the Jacobian-based Saliency Map Attack and Carlini and Wagner for white-box attacks and the One-Pixel Attack and Boundary Attack for black-box scenarios.

The black-box attacks, especially the Boundary Attack, achieved impressive results despite their limited access to the internal workings of the models. Even though the white-box attacks had no such limitations,

the black-box attacks sometimes outperformed them; that is, they generated adversarial examples with superior performance and lower disruption.

The authors said, "These observations are alarming as they imply that attackers can potentially achieve remarkable results without any understanding of the model's internal operation, simply by scrutinizing its output."

The research incorporated a gender-based perspective to investigate the differential impacts of adversarial attacks on male and female speech as well as on speech in different languages. In evaluating the impacts of attacks across three languages, only minor performance differences were observed.

English appeared the most susceptible while Italian displayed the highest resistance. The detailed examination of male and female samples indicated a slight superiority in male samples, which exhibited marginally less accuracy and perturbation, particularly in white-box attack scenarios. However, the variations between male and female samples were negligible.

"We devised a pipeline to standardize samples across the 3 languages and extract log-Mel spectrograms. Our methodology involved augmenting datasets using pitch shifting and time stretching techniques while maintaining a maximum sample duration of 3 seconds," the authors explained. Additionally, to ensure methodological consistency, the team used the same convolutional neural network long short-term memory architecture for all experiments.

While the publication of research revealing vulnerabilities in speech emotion recognition models might seem like it could provide attackers with valuable information, not sharing these findings could potentially be

more detrimental. Transparency in research allows both attackers and defenders to understand the weaknesses in these systems.

By making these vulnerabilities known, researchers and practitioners can better prepare and fortify their systems against potential threats, ultimately contributing to a more secure technological landscape.

More information: Nicolas Facchinetti et al, A Systematic Evaluation of Adversarial Attacks against Speech Emotion Recognition Models, *Intelligent Computing* (2024). DOI: 10.34133/icomputing.0088

Provided by Intelligent Computing