In this video, we'll go over some linguistic tools you can use to spot fake audio when you encounter it in the wild. I'll also give an overview of our project and the multidisciplinary team of data scientists and sociolinguists who developed this strategy for detecting audio deepfakes.

To first give an introduction to deepfakes, for those that may not be familiar, deepfake is a combination of the words "deep learning" and "fake," and it means any content that is generated or manipulated using artificial intelligence and then employed for malicious purposes. Deepfakes can include audio, video, images, and text. The type of deepfake we're concerned with today is audio deepfakes. Deepfakes can mimic a real person's voice, making them a powerful tool for deception; for example, they can be used to trick voice authentication systems or to make fake content using a public figure's voice.

Because audio deepfakes are becoming harder and harder to detect, we want to give you, the listener, some strategies to help you spot deepfakes.

Our research team includes experts in data science, machine learning, and variationist sociolinguistics, a subdiscipline of linguistics that deals with how language varies with respect to an assortment of different social and individual factors. Team members with a background in variationist sociolinguistics listened to hundreds of audio files of fake and real English speech, gathered with the help of our colleagues in data science, and then identified easily discernible linguistic features that are helpful in differentiating between authentic speech and deepfakes. We came up with five of what we call expert-defined linguistic features, or EDLFs, that can help individuals get better at discerning fake speech from real speech.

The first EDLF is pitch, or the perceived relative high or low tone of a speech sample. Pause includes a break in speech production within a speech sample. Word-Initial and Word-Final Consonant Bursts include the bursts of air when the English stop consonants sounds are released. Next, Breath is any audible intake or outtake of breath within a speech sample, and, finally, Audio Quality encompasses an overall estimation of the quality of the audio in a particular speech sample.

These five EDLFs–pitch, pauses, word-initial and word-final consonant bursts, breath, and audio quality–include commonly occurring, variable, and distinguishing phonetic and phonological characteristics of spoken English.

For each real or fake speech sample in our dataset, the sociolinguist team members listened to them and then perceptually identified and labeled whether any of these features were present or absent, then annotated for any anomalies in their production. Those labels indicate linguistic characteristics that are potential indicators of real versus fake speech. Those expert-annotated datasets were then used to augment models that our machine learning colleagues developed to help detect deepfake audio, with promising results.

In our other videos on the EDLFs, we'll explain each of the five features in greater detail and show you how to listen for them when you encounter potentially fake audio.