This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Deepfake detection

In this caselet, you'll explore a benevolent use case for speech recognition and synthesis technology. Voice-AI technologies can cause unintended harm, but they also have exciting applications to benefit society. Read the background information carefully, and use it to answer the multiple choice questions that follow.

By the end of this caselet, you will learn some uses of voice-AI for benevolent purposes and brainstorm some possible future uses of voice-AI for societal benefit.

Background

In the last decade, synthetic speech – speech generated by a computer – has become easy to create, highly naturalistic, and ubiquitous throughout society. Because of its convincingly humanlike sound, malicious people have started using synthetic voices for scams, fraud, and to distribute misinformation. An increasingly common malicious use of synthetic speech is the creation of audio deepfakes, which are synthetic voices designed to sound virtually identical to a real person's voice. A convincing deepfake of a person's voice can be created from a recording of only three to seven seconds of that person's speech.

This poses serious financial and political risks, as well as individual risk of identity theft or fraud. Because of this, deepfake detection has already become a fast-growing area of research; but, most approaches to deepfake detection rely on artificial intelligence algorithms. While deep learning algorithms do a reasonably good job of detecting deepfaked audio, they are time-consuming to create, very expensive and energy-intensive to run, and often cannot be accessed by the public. Beyond these limitations, even the best deepfake detection models currently do not catch the most sophisticated deepfaked audio samples. Although deep learning and artificial intelligence are a promising approach to develop solutions for identifying deepfaked audio, other methods like linguistic analysis can be used as a complimentary method to both improve detection algorithms and make deepfake detection less resource-intensive.

Another benefit of using linguistic analysis to detect voice deepfakes is that ordinary people can learn these strategies and apply them in situations where they are not sure if the voice they are hearing is a real person, or a voice clone. In this caselet, you will develop strategies to identify audio based on linguistic cues.

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Problem context

You are a researcher who has been tasked with finding new ways to identify deepfake and synthetic audio without relying on machine learning algorithms. To do this, you need to apply knowledge about how real, human speakers of language typically speak, and identify reliable and generalizable patterns in how human voices differ from fake voices. Ideally, you also want to find linguistic features that apply across as many languages and dialects as possible– just like deep learning algorithms, we want our linguistic features to be generalizable.

One way to systematically and reliably identify fake audio is by applying knowledge of *expert-defined linguistic features*, or EDLFs. EDLFs were developed by language scientists who specialize in understanding how people vary their speech, which is a specialty known as sociolinguistics. To identify these linguistic features, expert linguists listened to hours of real and fake English audio recordings and identified common linguistic features that point to whether the audio is real or fake.

First, let's discuss four EDLFs already identified by data scientists and linguists at the University of Maryland, Baltimore County. Then, you will come up with your own linguistic features that can be used to identify fake audio.

They identified these four features as useful and recognizable speech cues:

Pitch – Each person differs in how high or low-pitched their voice sounds, and people vary their pitch across words, phrases and sentences in predictable ways.

Pause – A break in the middle of speech production. People pause during speech for various different linguistic functions, like to emphasize a particular phrase in a sentence or to check for comprehension by their conversation partner.

Intake/Outtake of Breath – After speaking for some time, people must breathe. Real humans choose breath pauses at contextually-appropriate breaks in speech, like the end of a phrase or before a stylistic pause.

Audio Quality – Overall audio quality of a sample. This includes any disturbances or distortion in the sound. It also includes audio that sounds tinny or muffled. A recording that has poor quality does not always mean it is fake, but it can be a possible red flag.

A major benefit of these four ELDFs is that they can generalize across different dialects and speaking styles of English. They can also somewhat generalize across other languages. For example, every human needs to take a breath while talking, so we can expect that to be the case regardless of what language the person is speaking. The same is true for pauses; across languages, people tend to not pause in the middle of words but tend to pause at the end of

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

phrases or sentences. Audio quality is another generalizable feature because it does not have to do with language itself, but rather how the speech was recorded (or manipulated!).

Q1: Which of these features would generalize across languages?

	Intake/outtake of breath	Pitch variation	Length of and context of pauses within a sentence
True	☐	☐	☐
False	☐	☐	☐

The features we discussed in the last section have already been tested by research as possibly reliable ways to judge if audio is fake. In the next section, we will talk about two new linguistic features that could work for detection methods.

Stress In all words with more than one syllable, one syllable will "stand out" as the most emphasized part of the word. This is something that speakers intuitively learn when they learn a language, and it also varies depending on the language or dialect being spoken. For example, in American English most people agree that the stressed syllable in *computer* is the second one: com-PU-ter. Words can also vary in syllable stress based on the word meaning. For example, the word *object* has different syllable stress depending on whether it is being used as a verb or a noun. "OB-ject" is the noun form, like in the sentence "There was an OB-ject on the table." The verb form stresses "ob-JECT," like in the sentence "I ob-JECT to your proposal." Fluent speakers of English are familiar with the correct syllable stress for the word they want to use, and what stress is appropriate based on the context. However, sometimes text-to-speech fake audio chooses the incorrect syllable stress for a word based on the context—for example, producing "ob-JECT" when the meaning of the sentence calls for "OB-ject." Incorrect stress patterns can be a telltale sign that audio is fake.

Speech rate Speech rate refers to how fast a speaker is talking. In general, speakers will use a relatively consistent speech rate across an entire sentence. When speakers speed up or slow down their speech, it is usually to emphasize a particular portion of a sentence or phrase. Audio recordings that have sped up or slower portions that don't make sense in the context of the sentence are a red flag for fake audio.

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Q2: Which of these features do you think would be easiest to integrate into a deepfake detection algorithm?

- ☐ Speech rate, because it can be easily measured
- ☐ Stress, because dictionary knowledge exists on where the correct stress in a word should be

Finding new "tells" for deepfaked audio

Consider what linguistic features you might tackle next to identify fake audio. On the one hand, we want features that can generalize across languages and speakers, so that they can reliably be used to identify audio as real or fake. On the other hand, however, some features are more generalizable across languages than others. And for every language, including English, there is a wide variation in how it is spoken in different locations and contexts. One important principle to remember about language is that it varies and is always changing. For this reason, linguistic characteristics and features are not always interchangeable within or across languages.

Q3: What can serve as a source of inspiration for what linguistic features you can look for next?

- ☐ Listening to recordings of fake audio and comparing them to real audio
- ☐ Reading about common linguistic patterns in people's speech

Q4: When brainstorming new linguistic features, which of these factors is the most important?

- ☐ Identifying features that might exist across languages, not just English alone
- ☐ Identifying features that exists across dialects and individual differences in how people speak
- ☐ Zooming in on features that are most likely to exist in one particular variety of English

Staying ahead of realistic deepfakes

Voice deepfakes get dramatically more realistic each year, and so developing reliable ways to detect them is a moving target.

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Follow this link to listen to listen to real and deepfaked audio samples.

This website contains a very large dataset of real and deepfaked audio. You can listen to as many as you want, but try listening to at least ten. Keep going until you get at least one answer wrong.

Q5: Based on what you heard, list three different features that stood out to you as signs of deepfaked audio. When you heard these features, did you notice it in every deepfaked audio clip you heard, or just some of them?

_____

_____

_____

_____

_____

Q6: Based on what you heard, did you notice any features that you feel confident could not be a deepfake? Think carefully about this question– deepfakes are surprisingly sophisticated.

_____

_____

_____

_____

_____

This caselet was developed by student Chloe Evered as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and supervised by Dr. Vandana P. Janeja, Dr. Christine Mallinson and Dr. Karen Chen at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Q7: Do you think there are any features of speech that a computer could not replicate? If so, list what features you think those could be. Either way, explain your reasoning.

_____

_____

_____

_____

_____

Q8: What three linguistic features might you try next? Write down your answer below. Remember that these features don't have to work perfectly 100% of the time. In fact, they usually don't– the best deepfake detection strategies use a variety of different methods. Instead, think of linguistic features that might be difficult to fake, or a malicious actor might not consider when creating a deepfake.

_____

_____

_____

_____

_____