УДК 004.934.2

МЕТОДИКА ВЫЯВЛЕНИЯ ГОЛОСОВЫХ ДИПФЕЙКОВ

И.С. Дейкало, В.Д. Вольфович

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», г. Минск, Республика Беларусь

Аннотация. С развитием информационных технологий и искусственного интеллекта появились инструменты создания голосовых дипфейков, представляющие серьезную угрозу для информационной безопасности и доверия к аудиоконтенту. Данная технология может быть использована для манипуляции общественным мнением и введения в заблуждение. Голосовые подделки активно применяются в мошеннических схемах, что создает значительные риски для финансовых организаций, корпоративных структур и персональных данных пользователей. В данной работе рассматриваются ключевые технологии синтеза речи, включая Text-to-Speech и Voice Conversion. Описан процесс создания

голосового дипфейка с использованием сервиса iMyFone VoxBox. Проведен анализ методов выявления подделок, включая онлайн-сервис Deepfake-o-Meter и спектральный анализ аудиозаписей. Работа направлена на демонстрацию процесса создания голосового дипфейка и использования доступных инструментов для его последующего анализа и возможного применения в контексте защиты данных.
Ключевые слова: голосовой дипфейк, обнаружение голосового дипфейка, Text-to-Speech, iMyFone VoxBox, анализ спектрограммы.

METHODOLOGY OF VOICE-DEEPFAKE DETECTION

I.S. Deikalo, V.D. Volfovich
Educational Institution "Belarusian State University of Informatics and Radioelectronics", Minsk, Belarus

Abstract. With the development of information technology and artificial intelligence, tools for creating voice deepfakes have emerged, posing a serious threat to information security and trust in audio content. This technology can be used to manipulate public opinion and mislead. Voice forgeries are actively used in fraudulent schemes, which creates significant risks for financial organizations, corporate structures and personal data of users. This paper discusses key speech synthesis technologies, including Text-to-Speech and Voice Conversion. The process of creating a voice deepfake using the iMyFone VoxBox service is described. The analysis of methods for detecting fakes, including the online Deepfake-o-Meter service and spectral analysis of audio recordings, was carried out. The work is aimed at demonstrating the process of creating a voice deepfake and using available tools for its subsequent analysis and possible application in the context of data protection.
Keywords: voice deepfake, voice deepfake detection, Text-to-Speech, iMyFone VoxBox, spectrogram analysis.

Введение

Стремительное развитие информационных технологий, в частности искусственного интеллекта, открыло обширный набор инструментов для создания дезинформации – дипфейк или же синтез подходящего для манипулятора материала в формате видео, аудио или изображения для распространения вымысла в пагубных целях. В последнее время данную технологию используют для введения людей в заблуждение, подрывая репутацию некого объекта. Результативность дипфейк-технологий и дальнейшие перспективы развития заставляют остерегаться и стимулируют выявление способов противодействия. В данной статье будет рассмотрен именно голосовой дипфейк, так как является наиболее распространенным в мошеннических махинациях.

Основная часть

Для лучшего и более глубокого понимания темы, стоит ознакомиться с основными методами реализации данной технологии.

Преобразование текста в речь (Text-to-Speech). Данный метод основан на технологиях искусственного интеллекта, где текстовая информация преобразуется в синтезированную речь, имитирующую голос реального человека. Нейронные сети и модели глубокого обучения позволяют создавать аудиозаписи с высокой степенью реалистичности, что затрудняет их различение от подлинных. На данный момент существуют такие модели Text-to-speech, как VoCo, MelGAN, AdaSpeech, Tacotron 2, DeepVoice 3, MelNet, GlowTTS [1].

Преобразование голоса (Voice Conversion), включая имитацию. Преобразование голоса предполагает изменение характеристик исходного голоса говорящего таким образом, чтобы он звучал как голос другого человека. Этот процесс также реализуется с помощью алгоритмов глубокого обучения и нейронных сетей, которые анализируют и воспроизводят уникальные особенности голоса, такие как тембр, интонация и

акценты. Существуют такие модели, как Voice Conversion: Cotatron, Assem, Mellotron VC, StarGAN VC, PPG-VC [1].

Для целей исследования авторами был создан экземпляр голосовой подделки с помощью сервиса iMyFone. Процесс включал несколько последовательных этапов, начиная с записи образцов реального голоса и заканчивая получением голосового дипфейка.

После записи образцов голоса было решено дополнительно улучшить качество записи путем шумоочистки. Этот процесс позволил снизить уровень фоновых помех и повысить четкость звука, что было критически важно для последующего использования записи в качестве исходных данных для синтеза голосового дипфейка.

Полученная аудиозапись была загружена в сервис iMyFone VoxBox, который использует методы машинного обучения для синтеза речи. Было введено текстовое сообщение «*Мне очень тяжело, попал в беду и нуждаюсь в помощи. Каждая ваша поддержка важна для меня. Перевести средства можно на карту или через кошельки*», необходимый для реализации эксперимента, а также аудиофайл с записью голоса. На основе предоставленных данных, сервис использовал алгоритмы, которые анализируют особенности произношения, тембр, интонации и другие параметры голоса.

Была проведена проверка существующих сервисов обнаружения подделок. Первым был рассмотрен онлайн сервис Deepfake-o-Meter, который предоставляет возможность проверки аудиофайла на составляющие дипфейка с помощью определенной модели. По результатам модели «RawNet3(2023)» была высчитана процентная вероятность в 83,3 %, что экземпляр представляет синтезированную подделку (рис. 1).

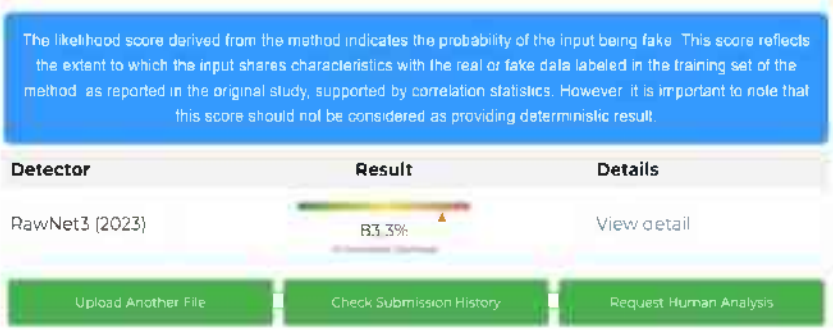

Рис. 1. Результаты проверки сервисом Deepfake-o-Meter
Fig. 1. Results of the Deepfake-o-Meter service check

Вторым методом является спектральный анализ. Для проверки была записана речь такого-же содержания настоящим человеком, голос которого был взят за основу при генерации с помощью сервиса iMyFone VoxBox. С помощью другого сервиса были извлечены спектрограммы человеческого и синтезированного происхождения.

На рис. 2 и рис. 3 представлены спектрограммы, на которых можно наблюдать некоторые схожести на начальных и конечных сегментах. Однако, стоит заметить, что в большинстве своем синтезированный экземпляр является обрезанной и неполной версией человеческой речи, хоть на выходе и звучит более чем правдоподобно. Самым главным наблюдением является аномалия в центральном сегменте, которая является главным доказательством работоспособности спектрального анализа для выявления подделок. Дело в том, что спектрограмма отображает изменение непрерывного аудио сигнала во времени [2]. А наличие аномальных пустот гласит о том, что присутствуют

незаметные для человеческого организма прерывания в сигнале, что противоречит природе человеческой речи.

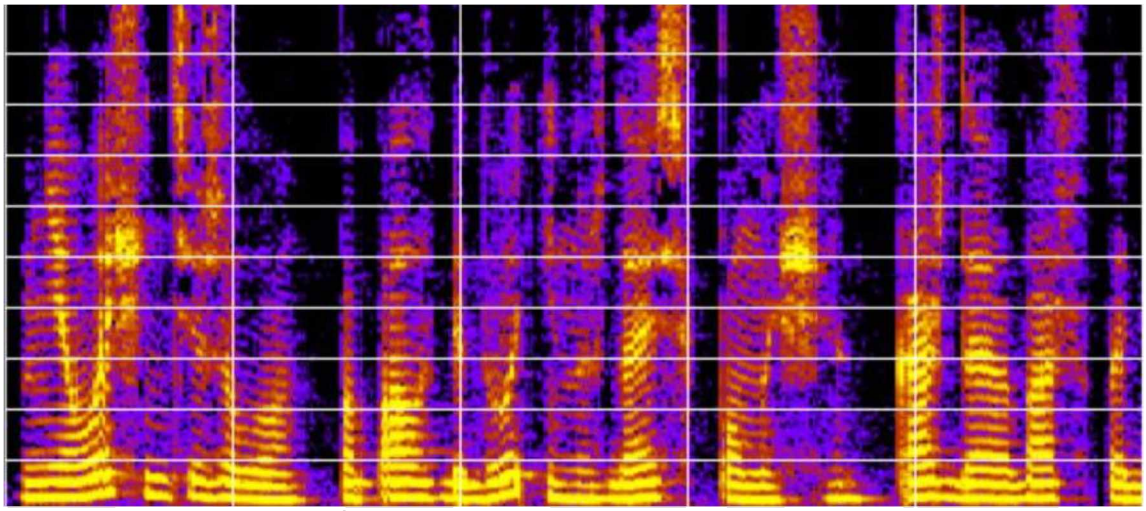

Рис. 2. Спектрограмма реального голосового сообщения
Fig. 2. Spectrogram of a real voice message

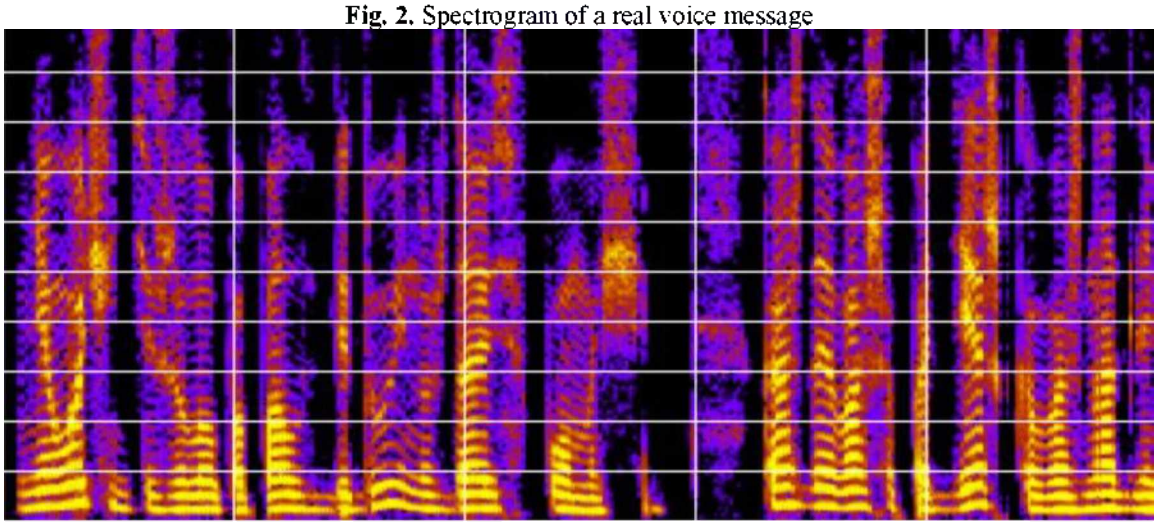

Рис. 3. Спектрограмма синтезированного голосового сообщения
Fig. 3. Spectrogram of a synthesized voice message

Заключение

В рамках исследования был синтезирован экземпляр голосового дипфейка и проведен его спектральный анализ. Анализ показал наличие выраженных различий в исходном и синтезированном речевом фрагменте, что можно использовать в дальнейших исследованиях методов выявления голосовых дипфейков.

Список использованных источников / References

1. Khanjani Z., Watson G., Janeja V.P., Audio deepfakes: A survey. Front. Big Data. 5:1001063. – 24 P. DOI: 10.3389/fdata.2022.1001063

2. Blue L., Warren K., Abdullah H., Gibson C., Vargas L., O'Dell J., Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. University of Florida. 978-1-939133-31-1.

Сведения об авторах

Дейкало И.С., студент факультета информационной безопасности, учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», ila.dejkalo.rabota@gmail.com.
Вольфович В.Д., студент факультета информационной безопасности, учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», volfovich-v@inbox.ru.

Information about the authors

Deikalo I.S., student of the Faculty of Information Security, Educational Institution "Belarusian State University of Informatics and Radioelectronics", ila.dejkalo.rabota@gmail.com.
Volfovich V.D., student of the Faculty of Information Security, Educational Institution "Belarusian State University of Informatics and Radioelectronics", volfovich-v@inbox.ru.