ЦРТ | ГРУППА КОМПАНИЙ

# Дипфейки в КЦ: миф или реальность

# Растущие объемы данных

ЦРТ | ГРУППА КОМПАНИЙ

**3,7+ млн**

видеороликов загружается
на YouTube каждый день

**270+ тыс.**

часов видеоконтента ежедневно

**328,77 млн**

терабайт данных создается ежедневно

**~120** зеттабайт данных

будет сгенерировано в этом году

**~181** зеттабайт данных

будет сгенерировано в 2025 году

50% интернет-трафика — видеоконтент

# Растет доступность технологий

ЦРТ | ГРУППА КОМПАНИЙ

## 11 января 2023

### Microsoft создала инструмент для подделки голоса любого человека, включая эмоции и тон

Microsoft разработала систему на основе искусственного интеллекта, которая может преобразовать текст в речь, произнесенную голосом любого человека, с передачей эмоций и тона говорящего. Для этого лишь понадобится трехсекундный образец его речи. По заявлению исследователей, испытавших инструмент под названием Vall-E, он значительно превосходит существующие системы синтеза речи.

**VALL-E – технология Microsoft синтеза речи:**

- на базе алгоритма EnCodec
- обучена на 60 000 часах англоязычной речи от более чем 7000 носителей
- 3 сек достаточно для синтезирования целевого голоса
- синтезированные записи только на английском языке

Выложены небольшое кол-во записей оригинальных и синтезированных голосов

## 20 февраля 2023

https://www.twitch.tv/atheneaiheroes


**На Twitch создали канал, где проходят круглосуточные стримы с звездами, сгенерированными ИИ**

- Генерация фейк видео известных персон
- Генерация синтезированной речи известных персон
- Генерация контента

ЦРТ | ГРУППА КОМПАНИЙ

**Эксперты предсказывают, что до 90% онлайн-контента может быть создано синтетически в течение нескольких лет**

Доклад Europol о борьбе с дипфейками

# Мировые тенденции

- С 2021 года фрод в контакт-центрах вырос **на 40%,** особенно – для получения доступа к счетам

- **64%** корпоративных организаций обеспокоены мошенничеством из контакт-центров

- **Более 80%** контакт-центров планируют внедрить методы аутентификации перед ответом, чтобы предотвратить риск атак социальной инженерии на операторов

ЦРТ | ГРУППА КОМПАНИЙ

Источник: TRUSTID State of Call Center Authentication 2021


# В России в 2022 году число кибератак на контакт-центры увеличилось в 10 раз

Intelligent Enterprise

# Атаки с применением ИИ

ЦРТ | ГРУППА КОМПАНИЙ

| Каналы | Риск |
| --- | --- |
| Контакт-центры | Атака на сервисы с целью получения каких-то услуг или данных 3-их лиц |
| ВКС и внутренние коммуникации | Подмена голоса/лица на видеозвонке, с последующими устными указаниями от имени подмененного сотрудника либо публичными заявлениями, телефонные звонки, почта |
| Публичные видео-сервисы | Размещение медиаконтента с публичными лицами компании, делающими дискредитирующие заявления, либо находящимися в дискредитирующих ситуациях |
| СМИ, соц. сети, мессенджеры | Распространение непроверенных материалов в СМИ |

## Цели

- Операторы КЦ, поддержки и иных цифровых / удаленных клиентских сервисов
- VIP-клиенты
- Массовые клиенты
- Внутренние сотрудники компании и контрагенты

# Примеры атак и сложность

ЦРТ | ГРУППА КОМПАНИЙ

Сложность атаки зависит от

- Стоимости создания
- Времени подготовки
- Сложности создания
- Длительности речи жертвы

## Replay spoofing

Запись голоса жертвы с помощью различного оборудования,
сложно запись голос жертвы с определенным текстом

## Text-to-speech

1. Старые технологии синтеза – это в основном UnitSelection технология, параметрический синтез с использования различных вокодеров (примеры в базе открытой базе ASVspoof2015, некоторые записи из ASVspoof2019, и старый синтез - ЦРТ, Amazon, Google и других)
2. Открытые решения из различных репозиториев github и voice constructors, в которые загружаются записи голоса,
   а результат получается автоматически (WaveNet, Tacotron, WaveGlow)
3. Hi-End технологии синтеза – включают в себя синтез ЦРТ, современный синтез Google, Amazon, IBM и других, которые предлагают готовые голоса из коробки. Создать целевой голос с помощью такого подхода дорого и долго

## Voice conversion spoofing

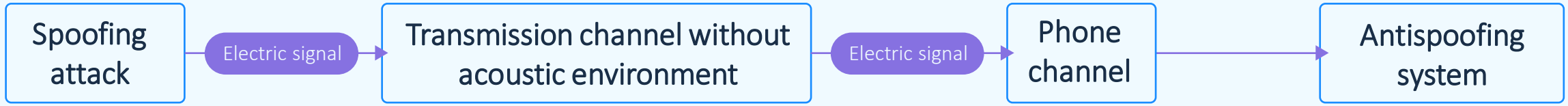
1. Открытые сервисы Voice Conversion
2. Research team create a target voice

# Разделение по типу каналов

ЦРТ | ГРУППА КОМПАНИЙ

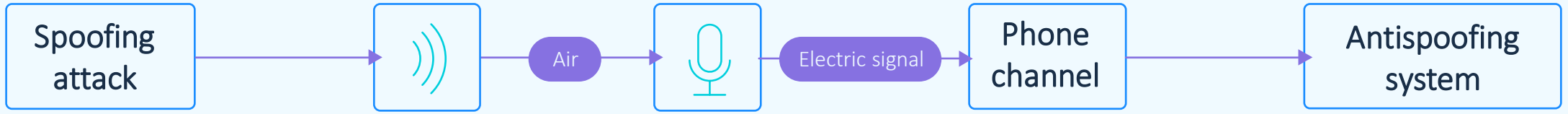
**Logical access**

Записи голоса человека и атаки, генерируемые с помощью TTS и VC, подаются непосредственно в телефонный канал и VoIP канал

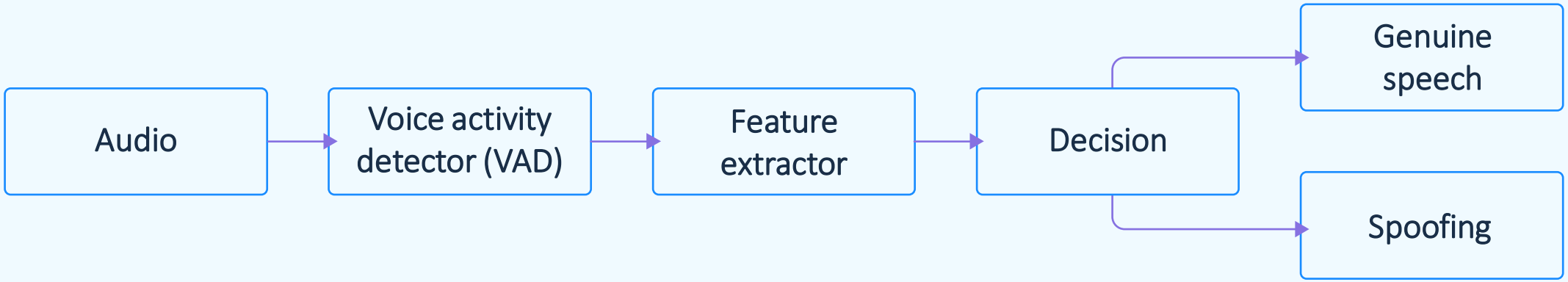



**Physical access**

Записи голоса человека и атаки (replay, TTS, VC) проигрываются в телефонную трубку, при этом в сигнал добавляются различные искажения – шумы, дополнительное эхо, также влияют устройства воспроизведения



# Как работает Antispoofing




| | | | |
|---|---|---|---|
| Запись голоса | VAD (детектор речи) отмечает сегменты с речью | Из выделенных сегментов извлекаются признаки | Генерируется общий вектор признаков(Embedding, vector) и принимается решение – запись живого человека или атака |

# Только ли дипфейки?

ЦРТ | ГРУППА КОМПАНИЙ

**ОКОЛО 30%**
исходящих коммуникаций компаний выходят на голосовых помощников и автоответчики, что провоцирует издержки

**БОЛЕЕ 10**
коммерческих голосовых помощников и различных вендоров синтеза на рынке РФ

# Кейс


## Проблема

При исходящем обзвоне звонок может принимать голосовой помощник или автоответчик. Операторы не заканчивают звонок, что ведет к увеличению расходов на оплату исходящего трафика, времени работы оператора

## Решение

Автоматически детектировать голосовых помощников и завершать звонок с помощью антиспуфинга

# Атаки на КЦ с использованием дипфейков

- 11% экспертов пока не видят угрозы в дипфейках, синтезированных при помощи речевых технологий, еще 22% не видят в них особой опасности. Остальные 2/3 респондентов считают это достаточно серьезной проблемой, если не сейчас, то в будущем.


- Атаки с помощью дипфейков могут стать серьезной проблемой в перспективе нескольких лет
- Атаки фиксируются, но не представляют особой опасности
- Пока атаки не зафиксированы, но в ближайшее время это станет серьезным вызовом для КЦ
- Атаки уже фиксируются и представляют серьезную опасность
- Не вижу особого риска существенного влияния таких атак на деятельность КЦ

Вопрос: Как оцениваете серьезность рисков атак на КЦ с помощью дипфейков?

**Образцы клонированных голосов, которые появляются в социальных сетях, мессенджерах и источниках Dark Web, часто используют голоса общественных деятелей, знаменитостей, политиков и интернет-личностей и предназначены для создания комедийного или вредоносного контента. Этот контент способствует распространению дезинформации, поскольку пользователи социальных сетей иногда обманываются высоким качеством образцов.**

**Kandinsky 2.2 – Злой пират из Dark Web*

# Драйверы и прогнозы

Люди (клиенты и сотрудники, в том числе операторы КЦ) подвержены эмоциональным триггерам и манипуляции


Генеративный ИИ становится проще использовать, технологии становятся быстрее и качественнее. Снижается стоимость атаки.

Отдельные законы и регуляция против использования фейк контента в РФ преимущественно слабы или не проработаны

## Риск:

На горизонте 2-5 лет развитие технологии синтеза и генеративного ИИ может привести к:

- Резкому удешевлению создания синтезированного клона голоса и лица
- Ускорению темпов роста качества синтезированных голосов по минимальным отрезкам речи для тренировки ML
- Рост атак на коммерческие и гос. сервисы по видео, голосовым и текстовым удаленным каналам взаимодействия с применением voice/face deepfakes и генерированного контента

# Атаки на КЦ с использованием дипфейков

- Несмотря на то, что треть респондентов не видят серьезной угрозы от дипфейков, всего 7% отказываются от внедрения защиты против этой угрозы. При этом всего четверть КЦ готовы устанавливать защитное ПО, 67% – раздумывают.

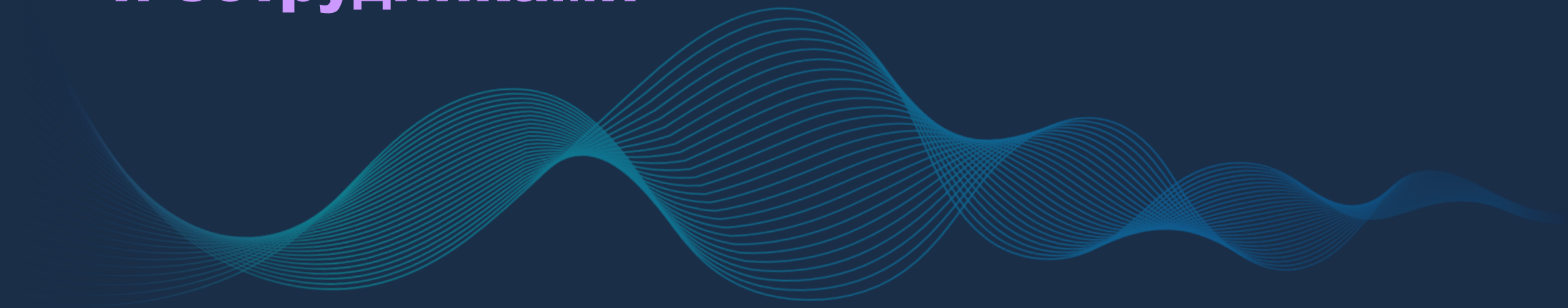

Вопрос: Готовы ли Вы внедрять ПО для защиты КЦ от атак с помощью дипфейков?

ЦРТ | ГРУППА КОМПАНИЙ

## Исследуйте коммуникации

# Извлекайте ценность из каждого диалога

# Постройте крепкие отношения с клиентами и сотрудниками



www.speechpro.ru