Bastam segundos para clonar uma voz com inteligência artificial. Vamos ter de aprender a desconfiar do que ouvimos

observador.pt/especiais/bastam-segundos-para-clonar-uma-voz-com-inteligencia-artificial-vamos-ter-de-aprender-a-desconfiar-do-que-ouvimos/

▲A OpenAI tem o Voice Engine, mas por agora considera que é demasiado arriscado fazer um lançamento ao público, temendo a desinformação

Cátia Rocha

A inteligência artificial está a conseguir imitar vozes de forma mais realista. O que, aos olhos dos especialistas, traz “muitos riscos pela frente”, como esquemas de extorsão ou a desinformação.

03 abr. 2024, 18:47 2

Imagine acordar a meio da noite com um telefonema. Do outro lado da linha, ouve a voz de um familiar, a chorar e a pedir ajuda. De repente, outra voz surge e exige um resgate em tom autoritário – tem é de fazer uma transferência rápida. Mas e se, na realidade, o familiar que chora ao telefone estiver, afinal, em segurança e aquela nem sequer for a sua voz?

Este é um dos exemplos de esquemas que estão a acontecer nos Estados Unidos, usando a inteligência artificial (IA) na clonagem de vozes. Desde o ano passado que estão a ser documentados vários casos, de acordo com a *CNN* internacional ou a revista *New Yorker.* A situação escalou ao ponto de já haver organizações a aconselhar as famílias a definirem palavras-passe ou expressões para verificarem a autenticidade nas comunicações.

“O problema não é a IA poder fazer coisas perigosas, é o mau uso que podemos fazer dela”, reconhece ao Observador a investigadora Isabel Trancoso, que lançou a unidade de processamento de discurso do INESC-ID e é professora no Instituto Superior Técnico. É que, se antes eram precisos muitos exemplos e minutos de som para conseguir imitar a voz de um humano, os avanços da IA fazem com que agora baste uma gravação com apenas alguns segundos para obter resultados convincentes. “E é complicado quando há muita gente a carregar a fala para a internet” em redes sociais e outras plataformas, completa.

“Há relativamente pouco tempo eram precisas grandes quantidades da voz de um orador para tentar criar um sistema” para ter uma imitação credível criada com IA, explica Luís Caldas Oliveira, diretor-adjunto do iStartLab Laboratório de Inovação e professor do Instituto Superior Técnico. Agora, “com poucos segundos de fala, consegue projetar-se como é que essa pessoa diria outras coisas que, na realidade, nunca disse”.

[Já saiu o sexto e último episódio de “Operação Papagaio” , o novo podcast plus do Observador com o plano mais louco para derrubar Salazar e que esteve escondido nos arquivos da PIDE 64 anos. Pode ouvir o primeiro episódio aqui, o segundo episódio aqui, o terceiro episódio aqui, o quarto episódio aqui e o quinto episódio aqui]

Torna-se cada vez mais difícil distinguir o que é uma voz humana de uma sintética, ou seja, desenvolvida por IA. Numa primeira fase, a ausência de expressividade e de nuances habituais no discurso – os *hum* ou as pequenas pausas – tornavam a distinção fácil entre o real e a máquina. Agora, até é possível escolher o tom do discurso e o idioma. Passou a ser “muito fácil copiar” uma voz e uma característica individual de determinada pessoa. “É isso o grande risco”, nota Luís Caldas Oliveira. “É muito fácil copiar e fazer de forma sintética essas características, enganando as pessoas.”

“O problema não é a IA poder fazer coisas perigosas, é o mau uso que podemos fazer dela”, reconhece ao Observador a investigadora Isabel Trancoso, que lançou a unidade de processamento de discurso do INESC-ID e é professora no Instituto Superior Técnico.

Há cada vez mais empresas a trabalhar nessa área, com destaque para a norte-americana OpenAI e a ElevenLabs, criada por uma dupla de polacos que pretendam evoluir nas dobragens de séries e filmes para uma voz mais natural. Ambas as empresas já reconheceram que as ferramentas que criaram **podem ser mal aplicadas e decidiram impor alguns limites**.

A OpenAI não disponibiliza ao público a tecnologia Voice Engine, por temer os riscos de imitações numa altura de campanha eleitoral nos EUA. Já a ElevenLabs garante que só permite criar imitações com alta qualidade a quem acede à versão paga e impôs medidas para prevenir a criação de clones de voz de “candidatos políticos envolvidos em eleições presidenciais ou legislativas”, a começar com os EUA e o Reino Unido.

“O público em geral já começa a ter literacia digital e começa a desconfiar do que lê”, reconhece Bruno Castro, especialista de cibersegurança da Visionware. **“Mas ainda não existe essa mentalidade, de desconfiança, do que ouvem e veem.”** Por isso considera que “a IA vai abrir a caixa de Pandora com os *deepfakes* [vídeos e imagens manipuladas]”. Uma tecnologia que, ainda por cima, “pode ser acedida a um custo muito baixo”.

Como é que funcionam os sistemas que imitam a voz? É difícil ter “deteção à prova de bala”

Os sistemas mais modernos para criar vozes sintéticas não necessitam de grandes quantidades de áudio porque detetam os padrões de fala, explica Luís Caldas Oliveira. Um padrão é a forma como cada pessoa diz uma certa frase e um conjunto de sons. “É por aí que se percebe como é que dizemos os ‘a’, ‘e’, a velocidade a que falamos, a entoação…”, enumera o especialista. “É preciso apenas um bocadinho desse exemplo para tentar encontrar uma voz que seja parecida” entre uma grande biblioteca de sons.

Antes não era assim. **“As pessoas tinham de dizer combinações de sons”** para que se conseguisse criar um sistema capaz de imitar alguém, , exemplifica Caldas Oliveira. “A forma como cada um de nós diz os sons é característica de como falamos.” Assim, “tomando apenas um bocadinho de uma frase”, os sistemas mais recentes conseguem “estimar como é que a pessoa diria outras frases”.

Depois, o sistema mapeia a configuração do trato vocal – “desde os pulmões até à boca” –, que é algo que, embora tenha particularidades individuais, na comparação com um universo alargado consegue apurar-se semelhanças. E, por isso, quanto maior número de vozes tiver o modelo, maior é a probabilidade de encontrar uma sonoridade parecida e uma correspondência com a voz original.

“Se formos buscar um bocadinho de cada pessoa, consegue-se criar uma espécie de pessoa virtual que fala como o indivíduo a quem se ‘apanhou’ apenas um pedaço da voz”, remata. Depois basta escrever o guião para essa “voz” dizer.

O investigador nota que, apesar de estarem cada vez mais evoluídos, há alguns pontos em que os sistemas cometem deslizes. “A palavra pinguim, em português europeu, é muito difícil” de reproduzir, explica Luís Calda Oliveira, já que o sistema vai ‘ler’ a palavra como se o “u” não existisse. Palavras homógrafas (escritas da mesma forma, mas com significado e pronúncia diferentes) também apresentam um desafio para estes sistemas.

A académica Isabel Trancoso reconhece que as vozes sintéticas “estão a melhorar cada vez mais” ao ponto de hoje **“não se conseguir fazer deteção à prova de bala”** do que é simulado e do que é real. O perigo “de desinformação é enorme” com a evolução desta tecnologia. “Temos de educar o público de que, a partir de agora, é muito mais difícil” fazer a distinção.

ElevenLabs continua a crescer. OpenAI evita lançamento geral da tecnologia devido aos riscos

Em apenas dois anos, a ElevenLabs conseguiu transformar-se num unicórnio, batendo a marca dos mil milhões de dólares de avaliação com uma ronda de 80 milhões, em janeiro. Tem sede em Londres, mas foi criada por Piotr Dabkowski, ex-engenheiro da Google, e Mati Staniszewski, ex-estratega da Palantir. Nasceram na Polónia e são amigos desde a infância. Desde miúdos que criticavam a qualidade das dobragens dos filmes e séries e decidiram criar uma empresa de clonagem de voz com IA para dar mais naturalidade ao discurso.

Entre Londres e os EUA, fizeram crescer a empresa, que já monetiza o serviço de clonagem de voz. Há diversas subscrições, mas por 11 dólares por mês tem-se acesso a uma ferramenta que pode ser usada para narrar áudiolivros ou vídeos e fazer dobragens de séries e filmes que se aproximem da voz original do ator noutro idioma.

Na versão gratuita, mais limitada, é possível escolher opções de uma biblioteca de vozes, selecionando a idade ou mesmo o sotaque disponível — há um Cristiano que consegue falar vários idiomas, mas com um ligeiro sotaque português. Noutra vertente, a empresa permite que os atores submetam a sua voz para que seja usada em vídeos de outras pessoas, recebendo uma compensação.

Depois de várias investigações revelarem que a tecnologia da ElevenLabs estava a ser usada para fins maliciosos, a empresa teve de agir a partir de janeiro de 2023. Começou por lançar uma ferramenta para analisar se um áudio é gerado com o seu *software*, possibilitando reportar usos inapropriados. Depois, introduziu limitações às contas grátis: o VoiceLab, que é a ferramenta para gerar voz, só está disponível nas versões pagas. Ainda assim reconhece que **"apresentar detalhes de pagamento não previne abusos"**, mas torna-os "menos anónimos e força-os a pensar duas vezes antes de partilharem conteúdo impróprio".

> Thank you everyone for your advice. We love what you're creating, but a set of actors use our tech for malicious purposes. We decided to take the following steps to address the issues:
>
> — ElevenLabs (@elevenlabsio) January 31, 2023

Já a OpenAI, a dona do ChatGPT, revelou na semana passada o Voice Engine, uma tecnologia que consegue recriar uma voz a partir de uma amostra de 15 segundos. Em teoria, basta carregar um áudio ou ler um pequeno parágrafo para ser gerada uma voz semelhante à do utilizador e consegue falar em vários idiomas. Por exemplo, alguém que só fala português pode ouvir a "sua" voz em mandarim.

▲ A OpenAI opta por não disponibilizar ao público em geral a ferramenta que permite clonar voz
Future Publishing via Getty Imag

A ferramenta foi desenvolvida pela OpenAI no fim de 2022, sendo, até aqui, usada para dar voz ao ChatGPT e para a funcionalidade de ler texto em voz alta. A companhia reconhece, no entanto, que a voz gerada artificialmente **traz perigos,** evitando, por isso, disponibilizá-la ao público em geral. "Estamos a adotar uma abordagem cautelosa e informada em relação a um lançamento alargado devido ao potencial mau uso de voz sintética", explica a companhia.

Considera que há usos benéficos para esta tecnologia, como ajudar alguém a ouvir um texto que não consegue ler ou ajudar quem não consegue comunicar. **Mas, por agora, teme os riscos do uso alargado desta tecnologia.** "Reconhecemos que gerar discurso que possa imitar a voz das pessoas tem riscos sérios, especialmente em momentos de

eleições.” Por isso, só vai permitir testes limitados com o Voice Engine para que seja possível perceber “a resiliência social contra os desafios gerados por **modelos generativos cada vez mais convincentes**”.

As empresas que vão testar esta tecnologia tiveram de aceitar as políticas de utilização, como a proibição da “imitação de um indivíduo ou organização sem consentimento ou direito igual” ou a obrigatoriedade de colocar uma indicação de que a voz foi gerada por IA. A OpenAI prevê ainda a criação de uma lista de vozes proibidas para evitar “vozes demasiado parecidas com as de figuras públicas”.

A OpenAI também diz que implementou medidas como a inclusão de uma marca de água nas vozes geradas pelo Voice Engine. A ideia é que seja possível perceber a origem de qualquer áudio, assim como “monitorizar de forma proativa o seu uso”.

Tanto Luís Caldas Oliveira como Isabel Trancoso destacam a existência de marcas de água como salvaguarda, **ainda que admitam que não são infalíveis.** “Com os sistemas digitais conseguimos colocar em cima da gravação de uma fala pequenas modificações no que está gravado, que são impercetíveis ao ouvido, mas que têm um certo código”, explica Caldas Oliveira. “Uma pessoa normal não consegue perceber, mas o computador consegue procurar a existência de código.”

“O problema é que é relativamente fácil retirar” as marcas de água, considera Isabel Trancoso – por exemplo, reproduzir o áudio e gravá-lo novamente. “As pessoas que agem de forma maliciosa encontram” outras formas de contornar as limitações, completa Luís Caldas Oliveira. “As leis não são perfeitas e há de haver alguém que vai conseguir fazer o sistema sem introduzir esse código.”

Sistemas de imitação de voz e imagem abrem “caixa de Pandora” dos ataques “a um custo muito baixo”

A IA é **“claramente uma área explosiva”** para os profissionais da área de segurança, reconhece Bruno Castro, especialista de cibersegurança da Visionware. “Até para nós, da área da segurança, tornou-se uma ameaça grande”

Nunca foi tão fácil aceder a ferramentas de IA, nota. E, mesmo nas opções pagas, é possível ter acesso a “um custo muito baixo”. Abre-se uma “caixa de Pandora” para ações maliciosas, acredita este profissional de segurança.

Com as ferramentas topo de gama da IA, é possível “construir um esquema fraudulento com base em IA” que põe em causa algo de que ainda não nos habituámos a desconfiar – a imagem e a voz de alguém, considera Bruno Castro, da Visionware.

Com as ferramentas topo de gama da IA é possível “construir um esquema fraudulento” que põe em causa algo que ainda não nos habituámos a desconfiar. “Fomos parametrizados para não desconfiar nem da voz nem da imagem. Se nos aparecer um

conteúdo multimédia de alguém a dizer algo, para nós, ser humano, é automaticamente credível.” E, acredita o especialista da Visionware, aprender a desconfiar desses dois conteúdos “vai passar muito pela questão da literacia” e “muito pouco pela tecnologia”.

Bruno Castro antecipa que vai **“haver muito mais ataques utilizando som e imagem”**. Para já, porque “qualquer pessoa com internet vai conseguir montar um esquema para uma fraude e fazê-lo de forma massiva”, com o potencial de taxas de sucesso elevadas. E, quanto mais se pagar para ter acesso a essa tecnologia, melhor serão os resultados e os números de vítimas, considera.

“A criação do vetor [de ataque], por si só, baseado em voz e imagem ou só voz, é altamente perigoso neste momento, altamente volátil e que facilmente fica fora de controlo”, avisa.