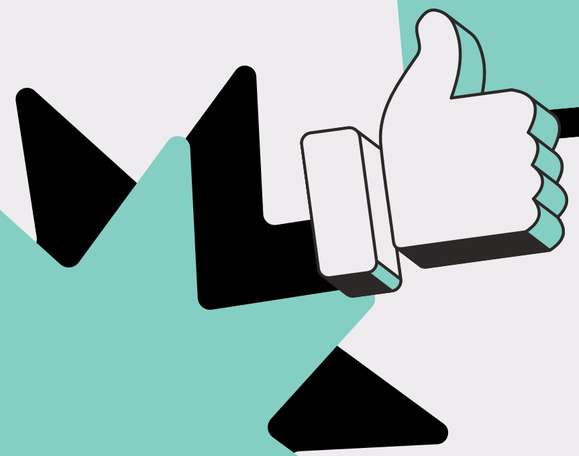
PANORAMA DAS FERRAMENTAS

GENERATIVAS PARA ÁUDIO



CONTEÚDO

Objetivos da aula:

- Compreender os fundamentos da geração de áudio e voz com IA
- Conhecer ferramentas clássicas e generativas
- Discutir aplicações práticas e responder dúvidas

O QUE É?

Áudio é uma representação das ondas sonoras que se propagam pelo ar e são percebidas pelo ouvido humano

Propriedades básicas

- são caracterizadas por propriedades como frequência (pitch), amplitude (volume) e timbre, que determinam aspectos como altura (grave ou agudo), intensidade (volume) e qualidade do som

O QUE É?

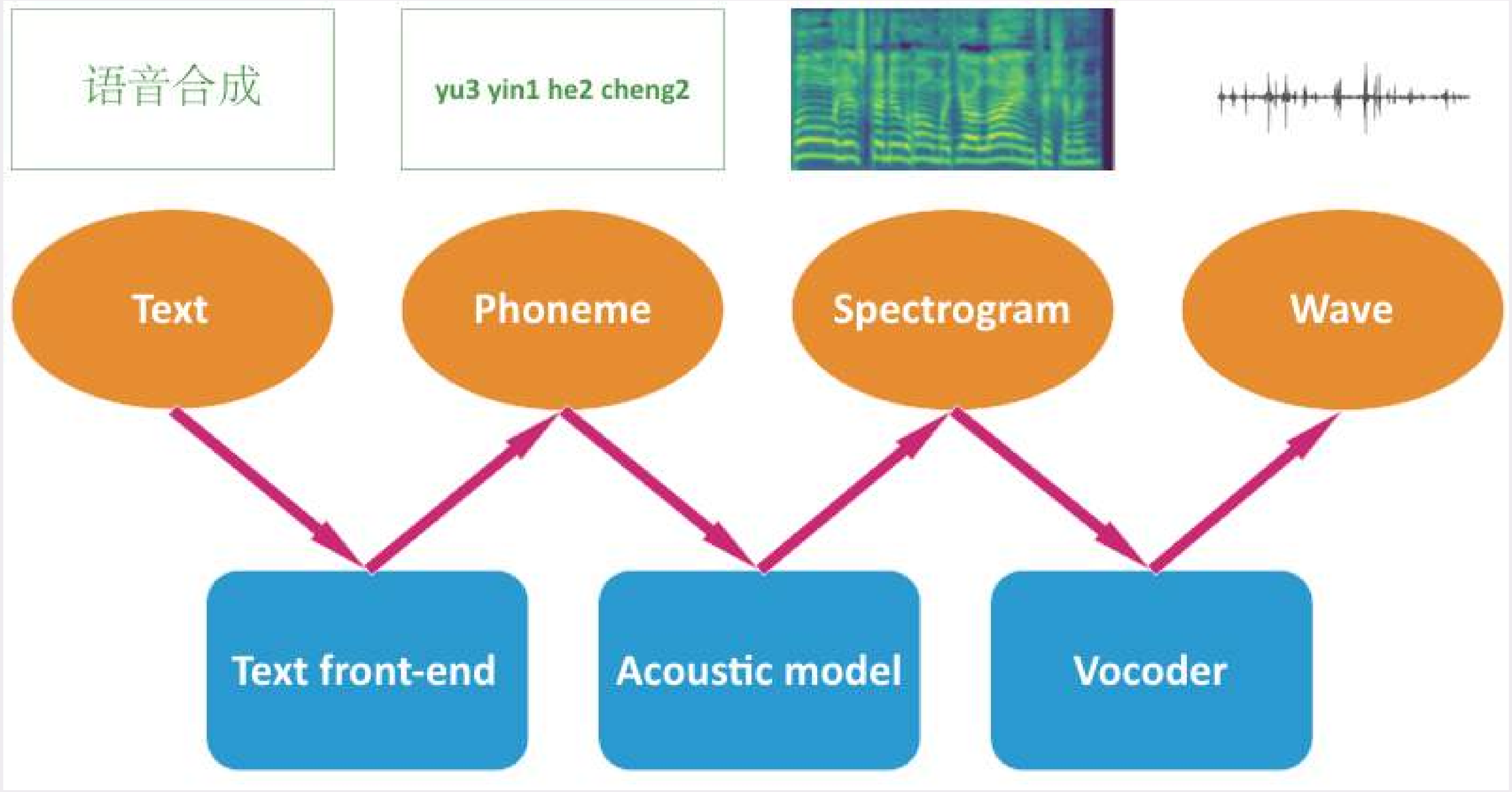
- **Oscilograma:** mostra a amplitude da onda no decorrer do tempo
- **Espectrogramas:** são representações visuais das frequências do som, ou seja, o som no domínio da frequência.

O QUE É?

- **Amostragem:** consiste em medir a amplitude do sinal analógico em intervalos de tempo regulares, resultando em uma sequência de valores discretos que representam o comportamento do sinal ao longo do tempo.
- **Quantização** é o processo de mapear essas amplitudes contínuas para um conjunto finito de níveis discretos, definidos pelo número de *bits* (valores binários) utilizados na representação digital.

TAREFAS

- **Text-to-speech (TTS):** síntese de fala baseavam-se em concatenação de segmentos de áudio ou em modelos estatísticos paramétricos, resultando em áudio com baixa naturalidade e fluidez.


语音合成
yu3 yin1 he2 cheng2
Text
Phoneme
Spectrogram
Wave
Text front-end
Acoustic model
Vocoder

TAREFAS

- **Melhoramento de voz (speech enhancer):** envolve técnicas para aprimorar a qualidade de gravações de áudio, reduzindo ruídos, reverberações e outras imperfeições

TAREFAS

- **Conversão de fala (voice conversion):** refere-se à transformação de características da voz, como timbre, sotaque ou idioma, mantendo o conteúdo semântico intacto.

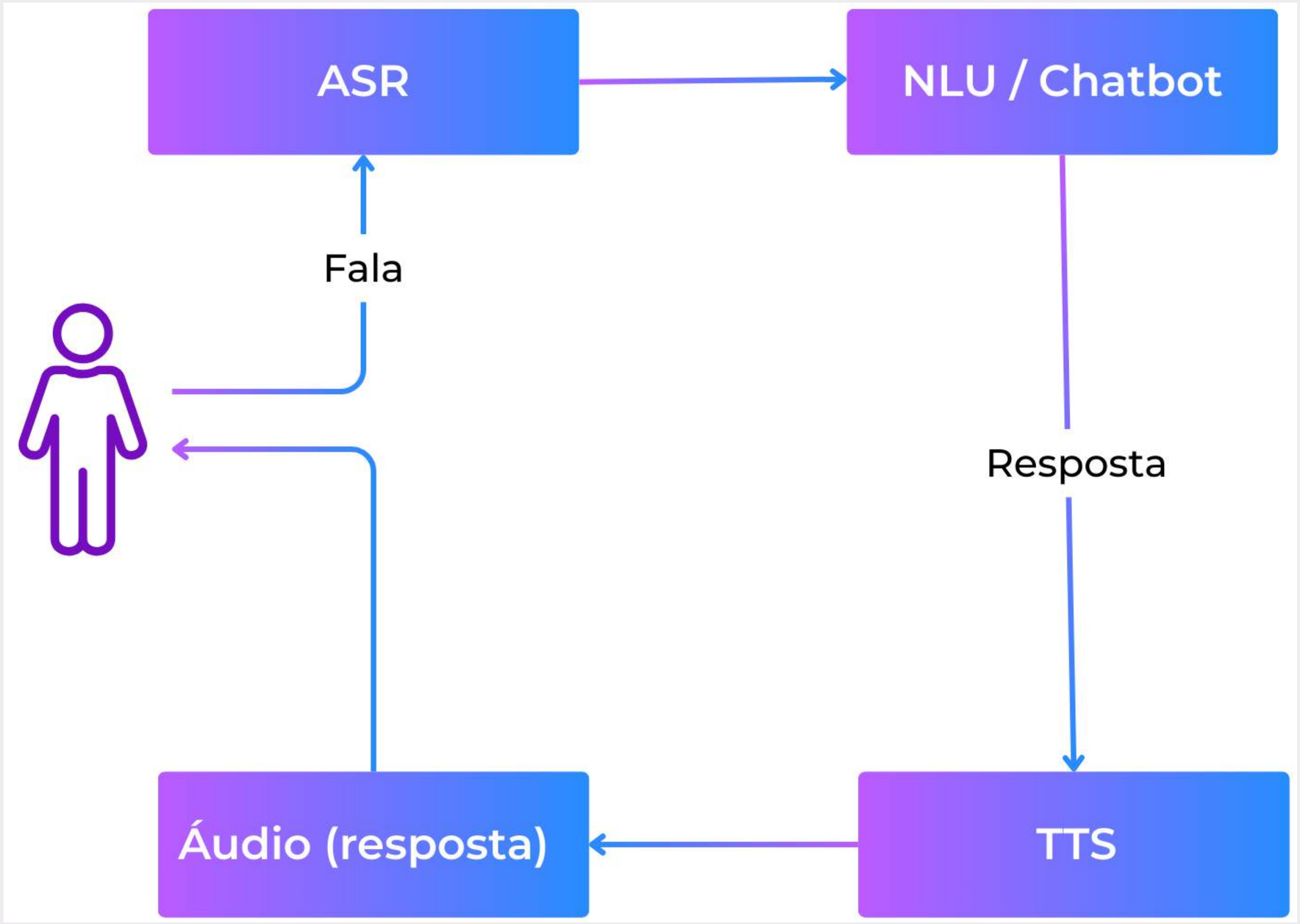
TAREFAS

Conversação: envolve a criação de sistemas capazes de manter diálogos naturais com usuários, compreendendo e gerando respostas em linguagem natural.

ESTRATÉGIAS PARA CONVERSÃO

Sistemas cascata: nesta abordagem, o processamento da linguagem é dividido em várias etapas sequenciais, cada uma responsável por uma função específica. As etapas típicas incluem:

- **Reconhecimento de fala (Automatic Speech Recognition [ASR]):** converte a fala do usuário em texto;
- **Compreensão da linguagem natural (Natural Language Understanding [NLU]):** interpreta o significado do texto;
- **Gerenciamento de diálogo:** determina a resposta apropriada com base no contexto;
- **Geração de linguagem natural (Natural Language Generation [NLG]):** cria a resposta em texto;
- **Síntese de fala (TTS):** converte o texto de resposta em fala audível.


ASR
NLU / Chatbot
Fala
Resposta
Áudio (resposta)
TTS

ESTRATÉGIAS PARA CONVERSÃO

- **Sistemas end-to-end**: nesta abordagem, um único modelo é treinado para mapear diretamente a entrada do usuário (fala ou texto) para a resposta apropriada, sem subdividir o processo em etapas distintas. Essa abordagem simplifica o pipeline e pode melhorar a fluidez das respostas, mas exige grandes volumes de dados para treinamento e pode ser menos interpretável.


Sistema de
conversação

Cenário	Requisitos	Desafios
Síntese de Fala para Assistentes Virtuais	• Naturalidade na fala • Personalização de voz • **Baixa latência**	• Necessidade de grandes volumes de dados para treinamento • Alta demanda por recursos computacionais
Conversão de Timbre de Voz para Aplicações Musicais	• **Alta qualidade de áudio** • Capacidade de generalização • Preservação da expressividade original	• Complexidade das arquiteturas para capturar nuances vocais
Sistema de Conversação em Tempo Real para Atendimento Telefônico	• Compreensão e geração de linguagem natural • Integração com sistemas de telefonia • Respostas contextualmente relevantes • **Baixa latência**	• Garantir baixa latência para fluidez na conversa • Necessidade de dados específicos do domínio para treinamento • Manutenção contínua para adaptação a novos contextos e consultas

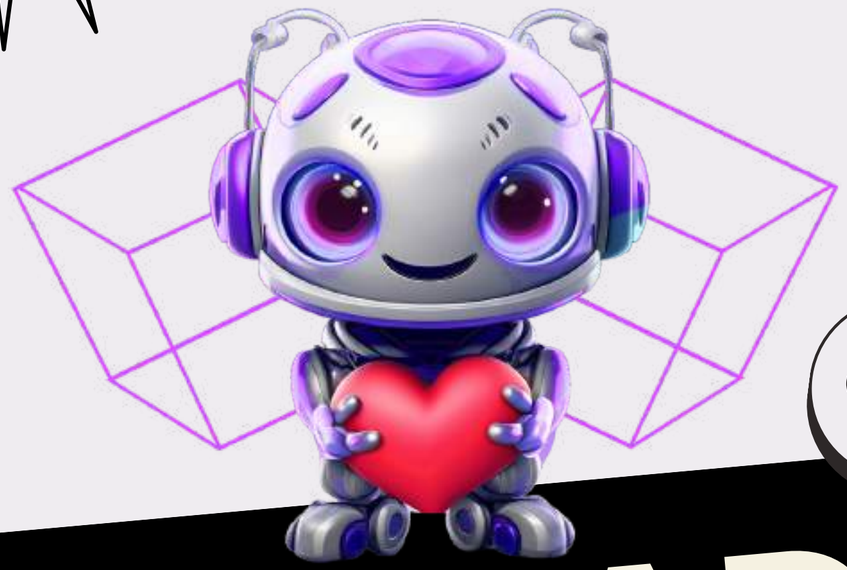
Geração de Voz para Audiolivros	• Expressividade na narração • Personalização de vozes para diferentes gêneros literários • Alta qualidade de áudio	• Processamento consistente de textos longos • Considerações de licenciamento ao utilizar vozes específicas
Tradução Simultânea com Síntese de Voz	• Precisão na tradução • Baixa latência para comunicação em tempo real • **Suporte a múltiplos idiomas**	• Gerenciamento de ruídos de fundo que afetam o reconhecimento de fala • Compreensão de variações dialetais e sotaques • Necessidade de infraestrutura robusta para processamento em tempo real

Ferramenta	Características	Vantagens	Desvantagens	Gratuita/Paga
Google Cloud Text-to-Speech	Oferece vozes de alta fidelidade com suporte a mais de 30 idiomas e variantes.	Integração com o Google Cloud®; vozes naturais; suporte a múltiplos idiomas.	Pode ser caro para grandes volumes de uso; dependência de conectividade com a internet.	Gratuita (limitada)/paga
Amazon Polly	Gera vozes naturais em diversos idiomas com opções de personalização.	Integração com *Amazon Web Services* (AWS); variedade de vozes; custo acessível.	Qualidade da voz pode variar entre idiomas; recursos avançados podem requerer plano pago.	Gratuita (limitada)/paga
IBM Watson Text to Speech	Oferece vozes neurais de alta qualidade com suporte a diversos idiomas.	Personalização avançada; integração com outros serviços de inteligência artificial (IA) da IBM®.	Pode ser complexo para iniciantes; custos podem ser elevados para uso extensivo.	Gratuita (limitada)/paga
Microsoft Azure Speech Service	Fornece vozes de alta qualidade com opções de personalização e suporte a vários idiomas.	Integração com o ecossistema Microsoft®; possibilidade de criar vozes personalizadas.	Curva de aprendizado para novas integrações; custos podem ser altos para grandes volumes.	Gratuita (limitada)/paga

Ferramenta	Características	Vantagens	Desvantagens	Gratuita/Paga
Murf.AI	Transforma texto em áudio realista e natural, disponível em mais de 20 idiomas.	Ajuste de entonação, ritmo e tom; suporte a múltiplas vozes e emoções.	Pode ser necessário plano pago para acesso completo aos recursos.	Gratuita (limitada)/Paga
Speechify	Oferece mais de 200 vozes de IA naturais em mais de 60 idiomas.	Integração com Google Docs®, notícias, e-mails, livros, PDFs; leitura até 4,5 vezes mais rápida.	Focado em conversão de textos e livros; algumas vozes podem soar artificiais; recursos avançados podem	Gratuita (limitada)/Paga
Synthesia	Combina voz com avatar de IA para criar vídeos com narração automatizada.	Criação de vídeos com avatares realistas; suporte a múltiplos idiomas e sotaques.	Focada em vídeos; pode não ser ideal para uso apenas de áudio.	Paga
Ondoku	Converte texto em fala com voz de IA, suporta aproximadamente 50 idiomas.	Uso online sem instalação; até 5.000 caracteres gratuitos por mês; download fácil de MP3.	Limite de caracteres na versão gratuita; recursos avançados podem requerer plano pago.	Gratuita (limitada)/paga

Ferramenta	Características	Vantagens	Desvantagens	Gratuita/Paga
Speechelo	Gera locuções de IA em mais de 30 idiomas e estilos de voz.	Plataforma baseada em nuvem; vozes que soam como humanas; fácil de usar.	Algumas vozes ainda podem soar robóticas; recursos completos disponíveis apenas na versão paga.	Paga
toVoice	Transforma texto em fala realista; oferece tradução automática e fala para texto.	Plataforma completa com múltiplas funcionalidades; suporte a vários idiomas.	Pode exigir plano pago para acesso completo aos recursos; qualidade da voz pode variar.	Paga
GSpeech	Solução online de TTS para *sites*, aplicativos móveis, *ebooks* e mais.	Fácil integração; suporte a múltiplas plataformas; opções de personalização.	Pode requerer conhecimento técnico para integração; recursos avançados podem ter custo adicional.	Gratuita (limitada)/paga

bm borcelle media



OBRIGADA

LOG IN