---

title: Models

description: Learn about the models that power the ElevenLabs API.

---

## Flagship models

<CardGroup cols={2} rows={2}>

<Card title={<div className="flex items-start gap-2"><div>Eleven v3</div><div><img src="file:1a7d5630-68b4-4c74-8f95-365ed9af684d" alt="Alpha" /></div></div>} href="/docs/models#eleven-v3-alpha">

Our most emotionally rich, expressive speech synthesis model

<div className="mt-4 space-y-2">

<div className="text-sm">Dramatic delivery and performance</div>

<div className="text-sm">70+ languages supported</div>

<div className="text-sm">10,000 character limit</div>

<div className="text-sm">Support for natural multi-speaker dialogue</div>

</div>

</Card>

<Card title="Eleven Multilingual v2" href="/docs/models#multilingual-v2">

Lifelike, consistent quality speech synthesis model

<div className="mt-4 space-y-2">

<div className="text-sm">Natural-sounding output</div>

<div className="text-sm">29 languages supported</div>

<div className="text-sm">10,000 character limit</div>

<div className="text-sm">Most stable on long-form generations</div>

</div>

</Card>

<Card title="Eleven Flash v2.5" href="/docs/models#flash-v25">

Our fast, affordable speech synthesis model

<div className="mt-4 space-y-2">

<div className="text-sm">Ultra-low latency (~75ms&dagger;)</div>

<div className="text-sm">32 languages supported</div>

<div className="text-sm">40,000 character limit</div>

<div className="text-sm">Faster model, 50% lower price per character</div>

</div>

</Card>

<Card title="Eleven Turbo v2.5" href="/docs/models#turbo-v25">

High quality, low-latency model with a good balance of quality and speed

<div className="mt-4 space-y-2">

<div className="text-sm">High quality voice generation</div>

<div className="text-sm">32 languages supported</div>

<div className="text-sm">40,000 character limit</div>

<div className="text-sm">Low latency (~250ms-300ms&dagger;), 50% lower price per character</div>

</div>

</Card>

</CardGroup>

<CardGroup cols={1} rows={1}>

<Card title="Scribe v1" href="/docs/models#scribe-v1">

State-of-the-art speech recognition model

<div className="mt-4 space-y-2">

<div className="text-sm">Accurate transcription in 99 languages</div>

<div className="text-sm">Precise word-level timestamps</div>

<div className="text-sm">Speaker diarization</div>

<div className="text-sm">Dynamic audio tagging</div>

</div>

</Card>

</CardGroup>

<div className="text-center">

<div>[Pricing](https://elevenlabs.io/pricing/api)</div>

</div>

## Models overview

The ElevenLabs API offers a range of audio models optimized for different use cases, quality levels, and performance requirements.

| Model ID | Description | Languages |

| ---------------------------- | --------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- | ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- |

| `eleven\_v3` | Human-like and expressive speech generation | [70+ languages](/docs/models#supported-languages) |

| `eleven\_ttv\_v3` | Human-like and expressive voice design model (Text to Voice) | [70+ languages](/docs/models#supported-languages) |

| `eleven\_multilingual\_v2` | Our most lifelike model with rich emotional expression | `en`, `ja`, `zh`, `de`, `hi`, `fr`, `ko`, `pt`, `it`, `es`, `id`, `nl`, `tr`, `fil`, `pl`, `sv`, `bg`, `ro`, `ar`, `cs`, `el`, `fi`, `hr`, `ms`, `sk`, `da`, `ta`, `uk`, `ru` |

| `eleven\_flash\_v2\_5` | Ultra-fast model optimized for real-time use (~75ms&dagger;) | All `eleven\_multilingual\_v2` languages plus: `hu`, `no`, `vi` |

| `eleven\_flash\_v2` | Ultra-fast model optimized for real-time use (~75ms&dagger;) | `en` |

| `eleven\_turbo\_v2\_5` | High quality, low-latency model with a good balance of quality and speed (~250ms-300ms) | `en`, `ja`, `zh`, `de`, `hi`, `fr`, `ko`, `pt`, `it`, `es`, `id`, `nl`, `tr`, `fil`, `pl`, `sv`, `bg`, `ro`, `ar`, `cs`, `el`, `fi`, `hr`, `ms`, `sk`, `da`, `ta`, `uk`, `ru`, `hu`, `no`, `vi` |

| `eleven\_turbo\_v2` | High quality, low-latency model with a good balance of quality and speed (~250ms-300ms) | `en` |

| `eleven\_multilingual\_sts\_v2` | State-of-the-art multilingual voice changer model (Speech to Speech) | `en`, `ja`, `zh`, `de`, `hi`, `fr`, `ko`, `pt`, `it`, `es`, `id`, `nl`, `tr`, `fil`, `pl`, `sv`, `bg`, `ro`, `ar`, `cs`, `el`, `fi`, `hr`, `ms`, `sk`, `da`, `ta`, `uk`, `ru` |

| `eleven\_multilingual\_ttv\_v2` | State-of-the-art multilingual voice designer model (Text to Voice) | `en`, `ja`, `zh`, `de`, `hi`, `fr`, `ko`, `pt`, `it`, `es`, `id`, `nl`, `tr`, `fil`, `pl`, `sv`, `bg`, `ro`, `ar`, `cs`, `el`, `fi`, `hr`, `ms`, `sk`, `da`, `ta`, `uk`, `ru` |

| `eleven\_english\_sts\_v2` | English-only voice changer model (Speech to Speech) | `en` |

| `scribe\_v1` | State-of-the-art speech recognition model | [99 languages](/docs/capabilities/speech-to-text#supported-languages) |

| `scribe\_v1\_experimental` | State-of-the-art speech recognition model with experimental features: improved multilingual performance, reduced hallucinations during silence, fewer audio tags, and better handling of early transcript termination | [99 languages](/docs/capabilities/speech-to-text#supported-languages) |

<small>† Excluding application & network latency</small>

<Accordion title="Older Models">

<Warning>

These models are maintained for backward compatibility but are not recommended for new projects.

</Warning>

| Model ID | Description | Languages |

| ------------------------ | ---------------------------------------------------- | ---------------------------------------------- |

| `eleven\_monolingual\_v1` | First generation TTS model (outclassed by v2 models) | `en` |

| `eleven\_multilingual\_v1` | First multilingual model (outclassed by v2 models) | `en`, `fr`, `de`, `hi`, `it`, `pl`, `pt`, `es` |

</Accordion>

## Eleven v3 (alpha)

<Warning>

This model is currently in alpha and is subject to change. Eleven v3 is not made for real-time

applications like Conversational AI. When integrating Eleven v3 into your application, consider

generating several generations and allowing the user to select the best one.

</Warning>

Eleven v3 is our latest and most advanced speech synthesis model. It is a state-of-the-art model that produces natural, life-like speech with high emotional range and contextual understanding across multiple languages.

This model works well in the following scenarios:

- \*\*Character Discussions\*\*: Excellent for audio experiences with multiple characters that interact with each other.

- \*\*Audiobook Production\*\*: Perfect for long-form narration with complex emotional delivery.

- \*\*Emotional Dialogue\*\*: Generate natural, lifelike dialogue with high emotional range and contextual understanding.

With Eleven v3 comes a new Text to Dialogue API, which allows you to generate natural, lifelike dialogue with high emotional range and contextual understanding across multiple languages. Eleven v3 can also be used with the Text to Speech API to generate natural, lifelike speech with high emotional range and contextual understanding across multiple languages.

<Warning>

Eleven v3 API access is currently not publicly available, but will be soon. To request access,

please [contact our sales team](https://elevenlabs.io/contact-sales).

</Warning>

Read more about the Text to Dialogue API [here](/docs/capabilities/text-to-dialogue).

### Model selection

The model can be used with the Text to Speech API by selecting the `eleven\_v3` model ID. The Text to Dialogue API defaults to using the v3 model. Alternatively you can select a preview version which is formatted as `eleven\_v3\_preview\_YYYY\_MM\_DD`. When a preview version has been evaluated and is ready for production, it will be promoted to the `eleven\_v3` model ID. Use the evergreen `eleven\_v3` model ID for the most stable experience and the preview version for the latest features.

### Supported languages

The Eleven v3 model supports 70+ languages, including:

\_Afrikaans (afr), Arabic (ara), Armenian (hye), Assamese (asm), Azerbaijani (aze), Belarusian (bel), Bengali (ben), Bosnian (bos), Bulgarian (bul), Catalan (cat), Cebuano (ceb), Chichewa (nya), Croatian (hrv), Czech (ces), Danish (dan), Dutch (nld), English (eng), Estonian (est), Filipino (fil), Finnish (fin), French (fra), Galician (glg), Georgian (kat), German (deu), Greek (ell), Gujarati (guj), Hausa (hau), Hebrew (heb), Hindi (hin), Hungarian (hun), Icelandic (isl), Indonesian (ind), Irish (gle), Italian (ita), Japanese (jpn), Javanese (jav), Kannada (kan), Kazakh (kaz), Kirghiz (kir), Korean (kor), Latvian (lav), Lingala (lin), Lithuanian (lit), Luxembourgish (ltz), Macedonian (mkd), Malay (msa), Malayalam (mal), Mandarin Chinese (cmn), Marathi (mar), Nepali (nep), Norwegian (nor), Pashto (pus), Persian (fas), Polish (pol), Portuguese (por), Punjabi (pan), Romanian (ron), Russian (rus), Serbian (srp), Sindhi (snd), Slovak (slk), Slovenian (slv), Somali (som), Spanish (spa), Swahili (swa), Swedish (swe), Tamil (tam), Telugu (tel), Thai (tha), Turkish (tur), Ukrainian (ukr), Urdu (urd), Vietnamese (vie), Welsh (cym).\_

## Multilingual v2

Eleven Multilingual v2 is our most advanced, emotionally-aware speech synthesis model. It produces natural, lifelike speech with high emotional range and contextual understanding across multiple languages.

The model delivers consistent voice quality and personality across all supported languages while maintaining the speaker's unique characteristics and accent.

This model excels in scenarios requiring high-quality, emotionally nuanced speech:

- \*\*Character Voiceovers\*\*: Ideal for gaming and animation due to its emotional range.

- \*\*Professional Content\*\*: Well-suited for corporate videos and e-learning materials.

- \*\*Multilingual Projects\*\*: Maintains consistent voice quality across language switches.

- \*\*Stable Quality\*\*: Produces consistent, high-quality audio output.

While it has a higher latency & cost per character than Flash models, it delivers superior quality for projects where lifelike speech is important.

Our v2 models support 29 languages:

\_English (USA, UK, Australia, Canada), Japanese, Chinese, German, Hindi, French (France, Canada), Korean, Portuguese (Brazil, Portugal), Italian, Spanish (Spain, Mexico), Indonesian, Dutch, Turkish, Filipino, Polish, Swedish, Bulgarian, Romanian, Arabic (Saudi Arabia, UAE), Czech, Greek, Finnish, Croatian, Malay, Slovak, Danish, Tamil, Ukrainian & Russian.\_

## Flash v2.5

Eleven Flash v2.5 is our fastest speech synthesis model, designed for real-time applications and conversational AI. It delivers high-quality speech with ultra-low latency (~75ms&dagger;) across 32 languages.

The model balances speed and quality, making it ideal for interactive applications while maintaining natural-sounding output and consistent voice characteristics across languages.

This model is particularly well-suited for:

- \*\*Conversational AI\*\*: Perfect for real-time voice agents and chatbots.

- \*\*Interactive Applications\*\*: Ideal for games and applications requiring immediate response.

- \*\*Large-Scale Processing\*\*: Efficient for bulk text-to-speech conversion.

With its lower price point and 75ms latency, Flash v2.5 is the cost-effective option for anyone needing fast, reliable speech synthesis across multiple languages.

Flash v2.5 supports 32 languages - all languages from v2 models plus:

\_Hungarian, Norwegian & Vietnamese\_

<small>† Excluding application & network latency</small>

### Considerations

<AccordionGroup>

<Accordion title="Text normalization with numbers">

When using Flash v2.5, numbers aren't normalized in a way you might expect. For example, phone numbers might be read out in way that isn't clear for the user. Dates and currencies are affected in a similar manner.

This is expected as normalization is disabled for Flash v2.5 to maintain the low latency.

The Multilingual v2 model does a better job of normalizing numbers, so we recommend using it for phone numbers and other cases where number normalization is important.

For low-latency or Conversational AI applications, best practice is to have your LLM [normalize the text](/docs/best-practices/prompting/normalization) before passing it to the TTS model.

</Accordion>

</AccordionGroup>

## Turbo v2.5

Eleven Turbo v2.5 is our high-quality, low-latency model with a good balance of quality and speed.

This model is an ideal choice for all scenarios where you'd use Flash v2.5, but where you're willing to trade off latency for higher quality voice generation.

## Model selection guide

<AccordionGroup>

<Accordion title="Requirements">

<CardGroup cols={1}>

<Card title="Quality">

Use `eleven\_multilingual\_v2`

Best for high-fidelity audio output with rich emotional expression

</Card>

<Card title="Low-latency">

Use Flash models

Optimized for real-time applications (~75ms latency)

</Card>

<Card title="Multilingual">

Use either either `eleven\_multilingual\_v2` or `eleven\_flash\_v2\_5`

Both support up to 32 languages

</Card>

<Card title="Balanced">

Use `eleven\_turbo\_v2\_5`

Good balance between quality and speed

</Card>

</CardGroup>

</Accordion>

<Accordion title="Use case">

<CardGroup cols={1}>

<Card title="Content creation">

Use `eleven\_multilingual\_v2`

Ideal for professional content, audiobooks & video narration.

</Card>

<Card title="Conversational AI">

Use `eleven\_flash\_v2\_5`, `eleven\_flash\_v2`, `eleven\_multilingual\_v2`, `eleven\_turbo\_v2\_5` or `eleven\_turbo\_v2`

Perfect for real-time conversational applications

</Card>

<Card title="Voice changer">

Use `eleven\_multilingual\_sts\_v2`

Specialized for Speech-to-Speech conversion

</Card>

</CardGroup>

</Accordion>

</AccordionGroup>

## Character limits

The maximum number of characters supported in a single text-to-speech request varies by model.

| Model ID | Character limit | Approximate audio duration |

| ------------------------ | --------------- | -------------------------- |

| `eleven\_flash\_v2\_5` | 40,000 | ~40 minutes |

| `eleven\_flash\_v2` | 30,000 | ~30 minutes |

| `eleven\_turbo\_v2\_5` | 40,000 | ~40 minutes |

| `eleven\_turbo\_v2` | 30,000 | ~30 minutes |

| `eleven\_multilingual\_v2` | 10,000 | ~10 minutes |

| `eleven\_multilingual\_v1` | 10,000 | ~10 minutes |

| `eleven\_english\_sts\_v2` | 10,000 | ~10 minutes |

| `eleven\_english\_sts\_v1` | 10,000 | ~10 minutes |

<Note>For longer content, consider splitting the input into multiple requests.</Note>

## Scribe v1

Scribe v1 is our state-of-the-art speech recognition model designed for accurate transcription across 99 languages. It provides precise word-level timestamps and advanced features like speaker diarization and dynamic audio tagging.

This model excels in scenarios requiring accurate speech-to-text conversion:

- \*\*Transcription Services\*\*: Perfect for converting audio/video content to text

- \*\*Meeting Documentation\*\*: Ideal for capturing and documenting conversations

- \*\*Content Analysis\*\*: Well-suited for audio content processing and analysis

- \*\*Multilingual Recognition\*\*: Supports accurate transcription across 99 languages

Key features:

- Accurate transcription with word-level timestamps

- Speaker diarization for multi-speaker audio

- Dynamic audio tagging for enhanced context

- Support for 99 languages

Read more about Scribe v1 [here](/docs/capabilities/speech-to-text).

## Concurrency and priority

Your subscription plan determines how many requests can be processed simultaneously and the priority level of your requests in the queue.

Speech to Text has an elevated concurrency limit.

Once the concurrency limit is met, subsequent requests are processed in a queue alongside lower-priority requests.

In practice this typically only adds ~50ms of latency.

| Plan | Concurrency Limit<br /> (Multilingual v2) | Concurrency Limit<br /> (Turbo & Flash) | STT Concurrency Limit | Priority level |

| ---------- | ----------------------------------------- | --------------------------------------- | --------------------- | -------------- |

| Free | 2 | 4 | 10 | 3 |

| Starter | 3 | 6 | 15 | 4 |

| Creator | 5 | 10 | 25 | 5 |

| Pro | 10 | 20 | 50 | 5 |

| Scale | 15 | 30 | 75 | 5 |

| Business | 15 | 30 | 75 | 5 |

| Enterprise | Elevated | Elevated | Elevated | Highest |

The response headers include `current-concurrent-requests` and `maximum-concurrent-requests` which you can use to monitor your concurrency.

How endpoint requests are made impacts concurrency limits:

- With HTTP, each request counts individually toward your concurrency limit.

- With a WebSocket, only the time where our model is generating audio counts towards your concurrency limit, this means a for most of the time an open websocket doesn't count towards your concurrency limit at all.

### Understanding concurrency limits

The concurrency limit associated with your plan should not be interpreted as the maximum number of simultaneous conversations, phone calls character voiceovers, etc that can be handled at once.

The actual number depends on several factors, including the specific AI voices used and the characteristics of the use case.

As a general rule of thumb, a concurrency limit of 5 can typically support up to approximately 100 simultaneous audio broadcasts.

This is because of the speed it takes for audio to be generated relative to the time it takes for the TTS request to be processed.

The diagram below is an example of how 4 concurrent calls with different users can be facilitated while only hitting 2 concurrent requests.

<Frame background="subtle">

<img

src="file:2a96a5aa-809b-4f74-8f60-317df7036bb0"

alt="Concurrency limits"

/>

</Frame>

<AccordionGroup>

<Accordion title="Building AI Voice Agents">

Where TTS is used to facilitate dialogue, a concurrency limit of 5 can support about 100 broadcasts for balanced conversations between AI agents and human participants.

For use cases in which the AI agent speaks less frequently than the human, such as customer support interactions, more than 100 simultaneous conversations could be supported.

</Accordion>

<Accordion title="Character voiceovers">

Generally, more than 100 simultaneous character voiceovers can be supported for a concurrency limit of 5.

The number can vary depending on the character’s dialogue frequency, the length of pauses, and in-game actions between lines.

</Accordion>

<Accordion title="Live Dubbing">

Concurrent dubbing streams generally follow the provided heuristic.

If the broadcast involves periods of conversational pauses (e.g. because of a soundtrack, visual scenes, etc), more simultaneous dubbing streams than the suggestion may be possible.

</Accordion>

</AccordionGroup>

If you exceed your plan's concurrency limits at any point and you are on the Enterprise plan, model requests may still succeed, albeit slower, on a best efforts basis depending on available capacity.

<Note>

To increase your concurrency limit & queue priority, [upgrade your subscription

plan](https://elevenlabs.io/pricing/api).

Enterprise customers can request a higher concurrency limit by contacting their account manager.

</Note>

**Quality**

Use eleven\_multilingual\_v2

Best for high-fidelity audio output with rich emotional expression

**Low-latency**

Use Flash models

Optimized for real-time applications (~75ms latency)

**Multilingual**

Use either either eleven\_multilingual\_v2 or eleven\_flash\_v2\_5

Both support up to 32 languages

**Balanced**

Use eleven\_turbo\_v2\_5

Good balance between quality and speed

**Use case**

**Content creation**

Use eleven\_multilingual\_v2

Ideal for professional content, audiobooks & video narration.

**Conversational AI**

Use eleven\_flash\_v2\_5, eleven\_flash\_v2, eleven\_multilingual\_v2, eleven\_turbo\_v2\_5 or eleven\_turbo\_v2

Perfect for real-time conversational applications

**Voice changer**

Use eleven\_multilingual\_sts\_v2

Specialized for Speech-to-Speech conversion

**Character limits**

The maximum number of characters supported in a single text-to-speech request varies by model.

| **Model ID** | **Character limit** | **Approximate audio duration** |
| --- | --- | --- |
| eleven\_flash\_v2\_5 | 40,000 | ~40 minutes |
| eleven\_flash\_v2 | 30,000 | ~30 minutes |
| eleven\_turbo\_v2\_5 | 40,000 | ~40 minutes |
| eleven\_turbo\_v2 | 30,000 | ~30 minutes |
| eleven\_multilingual\_v2 | 10,000 | ~10 minutes |
| eleven\_multilingual\_v1 | 10,000 | ~10 minutes |
| eleven\_english\_sts\_v2 | 10,000 | ~10 minutes |
| eleven\_english\_sts\_v1 | 10,000 | ~10 minutes |

For longer content, consider splitting the input into multiple requests.

**Scribe v1**

Scribe v1 is our state-of-the-art speech recognition model designed for accurate transcription across 99 languages. It provides precise word-level timestamps and advanced features like speaker diarization and dynamic audio tagging.

This model excels in scenarios requiring accurate speech-to-text conversion:

* **Transcription Services**: Perfect for converting audio/video content to text
* **Meeting Documentation**: Ideal for capturing and documenting conversations
* **Content Analysis**: Well-suited for audio content processing and analysis
* **Multilingual Recognition**: Supports accurate transcription across 99 languages

Key features:

* Accurate transcription with word-level timestamps
* Speaker diarization for multi-speaker audio
* Dynamic audio tagging for enhanced context
* Support for 99 languages

Read more about Scribe v1 [here](https://elevenlabs.io/docs/capabilities/speech-to-text).

# Text to Speech

> A guide on how to turn text to speech with ElevenLabs

<img src="file:a084d5eb-fb24-42a0-94ab-40f213409989" alt="Text to Speech product feature" />

## Overview

ElevenLabs' Text to Speech technology is integral to our offerings, powering high-quality AI-generated speech across various applications worldwide. It's likely you've already encountered our voices in action, delivering lifelike audio experiences.

## Guide

<Frame background="subtle">

![Text to Speech demo](file:5e29834a-8e30-4329-886a-9369fa8f77ea)

</Frame>

<Steps>

<Step title="Text input">

Type or paste your text into the input box on the Text to Speech page.

</Step>

<Step title="Voice selection">

Select the voice you wish to use from your Voices at the bottom left of the screen.

</Step>

<Step title="Adjust settings (optional)">

Modify the voice settings for the desired output.

</Step>

<Step title="Generate">

Click the 'Generate' button to create your audio file.

</Step>

</Steps>

## Settings

Get familiar with the voices, models & settings for creating high-quality speech.

<AccordionGroup>

<Accordion title="Voices">

### Voices

<Frame background="subtle">

![Text to Speech voice

selection](file:79f6d826-0913-4774-9da5-33ce4bbe29a8)

</Frame>

We offer many types of voices, including the curated Default Voices library, completely synthetic voices created using our Voice Design tool, and you can create your own collection of cloned voices using our two technologies: Instant Voice Cloning and Professional Voice Cloning. Browse through our voice library to find the perfect voice for your production.

Not all voices are equal, and a lot depends on the source audio used to create that voice. Some voices will perform better than others, while some will be more stable than others. Additionally, certain voices will be more easily cloned by the AI than others, and some voices may work better with one model and one language compared to another. All of these factors are important to consider when selecting your voice.

[Learn more about voices](/docs/capabilities/voices)

</Accordion>

<Accordion title="Models">

### Models

<Frame background="subtle">

![Text to Speech model

selection](file:9b870f10-3547-4927-bb77-d02778a70680)

</Frame>

ElevenLabs offers two families of models: standard (high-quality) models and Flash models, which are optimized for low latency. Each family includes both English-only and multilingual models, tailored for specific use cases with strengths in either speed, accuracy, or language diversity.

<CardGroup cols={2} rows={2}>

<Card title={<div className="flex items-start gap-2"><div>Eleven v3</div><div><img src="file:1a7d5630-68b4-4c74-8f95-365ed9af684d" alt="Alpha" /></div></div>} href="/docs/models#eleven-v3-alpha">

Our most emotionally rich, expressive speech synthesis model

<div>

<div>

Dramatic delivery and performance

</div>

<div>

70+ languages supported

</div>

<div>

10,000 character limit

</div>

<div>

Support for natural multi-speaker dialogue

</div>

</div>

</Card>

<Card title="Eleven Multilingual v2" href="/docs/models#multilingual-v2">

Lifelike, consistent quality speech synthesis model

<div>

<div>

Natural-sounding output

</div>

<div>

29 languages supported

</div>

<div>

10,000 character limit

</div>

<div>

Most stable on long-form generations

</div>

</div>

</Card>

<Card title="Eleven Flash v2.5" href="/docs/models#flash-v25">

Our fast, affordable speech synthesis model

<div>

<div>

Ultra-low latency (~75msâ€ )

</div>

<div>

32 languages supported

</div>

<div>

40,000 character limit

</div>

<div>

Faster model, 50% lower price per character

</div>

</div>

</Card>

<Card title="Eleven Turbo v2.5" href="/docs/models#turbo-v25">

High quality, low-latency model with a good balance of quality and speed

<div>

<div>

High quality voice generation

</div>

<div>

32 languages supported

</div>

<div>

40,000 character limit

</div>

<div>

Low latency (~250ms-300msâ€ ), 50% lower price per character

</div>

</div>

</Card>

</CardGroup>

[Learn more about our models](/docs/models)

</Accordion>

<Accordion title="Voice settings">

### Voice settings

<Frame background="subtle">

![Text to Speech voice

settings](file:67a18ad3-f4df-4c8d-91e4-1e4a4d7b0f41)

</Frame>

Our users have found different workflows that work for them. The most common setting is stability around 50 and similarity near 75, with minimal changes thereafter. Of course, this all depends on the original voice and the style of performance you're aiming for.

It's important to note that the AI is non-deterministic; setting the sliders to specific values won't guarantee the same results every time. Instead, the sliders function more as a range, determining how wide the randomization can be between each generation.

#### Speed

The speed setting allows you to either speed up or slow down the speed of the generated speech. The default value is 1.0, which means that the speed is not adjusted. Values below 1.0 will slow the voice down, to a minimum of 0.7. Values above 1.0 will speed up the voice, to a maximum of 1.2. Extreme values may affect the quality of the generated speech.

#### Stability

The stability slider determines how stable the voice is and the randomness between each generation. Lowering this slider introduces a broader emotional range for the voice. As mentioned before, this is also influenced heavily by the original voice. Setting the slider too low may result in odd performances that are overly random and cause the character to speak too quickly. On the other hand, setting it too high can lead to a monotonous voice with limited emotion.

For a more lively and dramatic performance, it is recommended to set the stability slider lower and generate a few times until you find a performance you like.

On the other hand, if you want a more serious performance, even bordering on monotone at very high values, it is recommended to set the stability slider higher. Since it is more consistent and stable, you usually don't need to generate as many samples to achieve the desired result. Experiment to find what works best for you!

#### Similarity

The similarity slider dictates how closely the AI should adhere to the original voice when attempting to replicate it. If the original audio is of poor quality and the similarity slider is set too high, the AI may reproduce artifacts or background noise when trying to mimic the voice if those were present in the original recording.

#### Style exaggeration

With the introduction of the newer models, we also added a style exaggeration setting. This setting attempts to amplify the style of the original speaker. It does consume additional computational resources and might increase latency if set to anything other than 0. It's important to note that using this setting has shown to make the model slightly less stable, as it strives to emphasize and imitate the style of the original voice.

In general, we recommend keeping this setting at 0 at all times.

#### Speaker Boost

This setting boosts the similarity to the original speaker. However, using this setting requires a slightly higher computational load, which in turn increases latency. The differences introduced by this setting are generally rather subtle.

</Accordion>

</AccordionGroup>

## FAQ

<AccordionGroup>

<Accordion title="Good input equals good output">

The first factor, and one of the most important, is that good, high-quality, and consistent input will result in good, high-quality, and consistent output.

If you provide the AI with audio that is less than idealâ€”for example, audio with a lot of noise, reverb on clear speech, multiple speakers, or inconsistency in volume or performance and deliveryâ€”the AI will become more unstable, and the output will be more unpredictable.

If you plan on cloning your own voice, we strongly recommend that you go through our guidelines in the documentation for creating proper voice clones, as this will provide you with the best possible foundation to start from. Even if you intend to use only Instant Voice Clones, it is advisable to read the Professional Voice Cloning section as well. This section contains valuable information about creating voice clones, even though the requirements for these two technologies are slightly different.

</Accordion>

<Accordion title="Use the right voice">

The second factor to consider is that the voice you select will have a tremendous effect on the output. Not only, as mentioned in the first factor, is the quality and consistency of the samples used to create that specific clone extremely important, but also the language and tonality of the voice.

If you want a voice that sounds happy and cheerful, you should use a voice that has been cloned using happy and cheerful samples. Conversely, if you desire a voice that sounds introspective and brooding, you should select a voice with those characteristics.

However, it is also crucial to use a voice that has been trained in the correct language. For example, all of the professional voice clones we offer as default voices are English voices and have been trained on English samples. Therefore, if you have them speak other languages, their performance in those languages can be unpredictable. It is essential to use a voice that has been cloned from samples where the voice was speaking the language you want the AI to then speak.

</Accordion>

<Accordion title="Use proper formatting">

This may seem slightly trivial, but it can make a big difference. The AI tries to understand how to read something based on the context of the text itself, which means not only the words used but also how they are put together, how punctuation is applied, the grammar, and the general formatting of the text.

This can have a small but impactful influence on the AI's delivery. If you were to misspell a word, the AI won't correct it and will try to read it as written.

</Accordion>

<Accordion title="Nondeterministic">

The settings of the AI are nondeterministic, meaning that even with the same initial conditions (voice, settings, model), it will give you slightly different output, similar to how a voice actor will deliver a slightly different performance each time.

This variability can be due to various factors, such as the options mentioned earlier: voice, settings, model. Generally, the breadth of that variability can be controlled by the stability slider. A lower stability setting means a wider range of variability between generations, but it also introduces inter-generational variability, where the AI can be a bit more performative.

A wider variability can often be desirable, as setting the stability too high can make certain voices sound monotone as it does give the AI the same leeway to generate more variable content. However, setting the stability too low can also introduce other issues where the generations become unstable, especially with certain voices that might have used less-than-ideal audio for the cloning process.

The default setting of 50 is generally a great starting point for most applications.

</Accordion>

Text to Speech

> A guide on how to turn text to speech with ElevenLabs

<img src="file:a084d5eb-fb24-42a0-94ab-40f213409989" alt="Text to Speech product feature" />

## Overview

ElevenLabs' Text to Speech technology is integral to our offerings, powering high-quality AI-generated speech across various applications worldwide. It's likely you've already encountered our voices in action, delivering lifelike audio experiences.

## Guide

<Frame background="subtle">

![Text to Speech demo](file:5e29834a-8e30-4329-886a-9369fa8f77ea)

</Frame>

<Steps>

<Step title="Text input">

Type or paste your text into the input box on the Text to Speech page.

</Step>

<Step title="Voice selection">

Select the voice you wish to use from your Voices at the bottom left of the screen.

</Step>

<Step title="Adjust settings (optional)">

Modify the voice settings for the desired output.

</Step>

<Step title="Generate">

Click the 'Generate' button to create your audio file.

</Step>

</Steps>

## Settings

Get familiar with the voices, models & settings for creating high-quality speech.

<AccordionGroup>

<Accordion title="Voices">

### Voices

<Frame background="subtle">

![Text to Speech voice

selection](file:79f6d826-0913-4774-9da5-33ce4bbe29a8)

</Frame>

We offer many types of voices, including the curated Default Voices library, completely synthetic voices created using our Voice Design tool, and you can create your own collection of cloned voices using our two technologies: Instant Voice Cloning and Professional Voice Cloning. Browse through our voice library to find the perfect voice for your production.

Not all voices are equal, and a lot depends on the source audio used to create that voice. Some voices will perform better than others, while some will be more stable than others. Additionally, certain voices will be more easily cloned by the AI than others, and some voices may work better with one model and one language compared to another. All of these factors are important to consider when selecting your voice.

[Learn more about voices](/docs/capabilities/voices)

</Accordion>

<Accordion title="Models">

### Models

<Frame background="subtle">

![Text to Speech model

selection](file:9b870f10-3547-4927-bb77-d02778a70680)

</Frame>

ElevenLabs offers two families of models: standard (high-quality) models and Flash models, which are optimized for low latency. Each family includes both English-only and multilingual models, tailored for specific use cases with strengths in either speed, accuracy, or language diversity.

<CardGroup cols={2} rows={2}>

<Card title={<div className="flex items-start gap-2"><div>Eleven v3</div><div><img src="file:1a7d5630-68b4-4c74-8f95-365ed9af684d" alt="Alpha" /></div></div>} href="/docs/models#eleven-v3-alpha">

Our most emotionally rich, expressive speech synthesis model

<div>

<div>

Dramatic delivery and performance

</div>

<div>

70+ languages supported

</div>

<div>

10,000 character limit

</div>

<div>

Support for natural multi-speaker dialogue

</div>

</div>

</Card>

<Card title="Eleven Multilingual v2" href="/docs/models#multilingual-v2">

Lifelike, consistent quality speech synthesis model

<div>

<div>

Natural-sounding output

</div>

<div>

29 languages supported

</div>

<div>

10,000 character limit

</div>

<div>

Most stable on long-form generations

</div>

</div>

</Card>

<Card title="Eleven Flash v2.5" href="/docs/models#flash-v25">

Our fast, affordable speech synthesis model

<div>

<div>

Ultra-low latency (~75msâ€ )

</div>

<div>

32 languages supported

</div>

<div>

40,000 character limit

</div>

<div>

Faster model, 50% lower price per character

</div>

</div>

</Card>

<Card title="Eleven Turbo v2.5" href="/docs/models#turbo-v25">

High quality, low-latency model with a good balance of quality and speed

<div>

<div>

High quality voice generation

</div>

<div>

32 languages supported

</div>

<div>

40,000 character limit

</div>

<div>

Low latency (~250ms-300msâ€ ), 50% lower price per character

</div>

</div>

</Card>

</CardGroup>

[Learn more about our models](/docs/models)

</Accordion>

<Accordion title="Voice settings">

### Voice settings

<Frame background="subtle">

![Text to Speech voice

settings](file:67a18ad3-f4df-4c8d-91e4-1e4a4d7b0f41)

</Frame>

Our users have found different workflows that work for them. The most common setting is stability around 50 and similarity near 75, with minimal changes thereafter. Of course, this all depends on the original voice and the style of performance you're aiming for.

It's important to note that the AI is non-deterministic; setting the sliders to specific values won't guarantee the same results every time. Instead, the sliders function more as a range, determining how wide the randomization can be between each generation.

#### Speed

The speed setting allows you to either speed up or slow down the speed of the generated speech. The default value is 1.0, which means that the speed is not adjusted. Values below 1.0 will slow the voice down, to a minimum of 0.7. Values above 1.0 will speed up the voice, to a maximum of 1.2. Extreme values may affect the quality of the generated speech.

#### Stability

The stability slider determines how stable the voice is and the randomness between each generation. Lowering this slider introduces a broader emotional range for the voice. As mentioned before, this is also influenced heavily by the original voice. Setting the slider too low may result in odd performances that are overly random and cause the character to speak too quickly. On the other hand, setting it too high can lead to a monotonous voice with limited emotion.

For a more lively and dramatic performance, it is recommended to set the stability slider lower and generate a few times until you find a performance you like.

On the other hand, if you want a more serious performance, even bordering on monotone at very high values, it is recommended to set the stability slider higher. Since it is more consistent and stable, you usually don't need to generate as many samples to achieve the desired result. Experiment to find what works best for you!

#### Similarity

The similarity slider dictates how closely the AI should adhere to the original voice when attempting to replicate it. If the original audio is of poor quality and the similarity slider is set too high, the AI may reproduce artifacts or background noise when trying to mimic the voice if those were present in the original recording.

#### Style exaggeration

With the introduction of the newer models, we also added a style exaggeration setting. This setting attempts to amplify the style of the original speaker. It does consume additional computational resources and might increase latency if set to anything other than 0. It's important to note that using this setting has shown to make the model slightly less stable, as it strives to emphasize and imitate the style of the original voice.

In general, we recommend keeping this setting at 0 at all times.

#### Speaker Boost

This setting boosts the similarity to the original speaker. However, using this setting requires a slightly higher computational load, which in turn increases latency. The differences introduced by this setting are generally rather subtle.

</Accordion>

</AccordionGroup>

## FAQ

<AccordionGroup>

<Accordion title="Good input equals good output">

The first factor, and one of the most important, is that good, high-quality, and consistent input will result in good, high-quality, and consistent output.

If you provide the AI with audio that is less than idealâ€”for example, audio with a lot of noise, reverb on clear speech, multiple speakers, or inconsistency in volume or performance and deliveryâ€”the AI will become more unstable, and the output will be more unpredictable.

If you plan on cloning your own voice, we strongly recommend that you go through our guidelines in the documentation for creating proper voice clones, as this will provide you with the best possible foundation to start from. Even if you intend to use only Instant Voice Clones, it is advisable to read the Professional Voice Cloning section as well. This section contains valuable information about creating voice clones, even though the requirements for these two technologies are slightly different.

</Accordion>

<Accordion title="Use the right voice">

The second factor to consider is that the voice you select will have a tremendous effect on the output. Not only, as mentioned in the first factor, is the quality and consistency of the samples used to create that specific clone extremely important, but also the language and tonality of the voice.

If you want a voice that sounds happy and cheerful, you should use a voice that has been cloned using happy and cheerful samples. Conversely, if you desire a voice that sounds introspective and brooding, you should select a voice with those characteristics.

However, it is also crucial to use a voice that has been trained in the correct language. For example, all of the professional voice clones we offer as default voices are English voices and have been trained on English samples. Therefore, if you have them speak other languages, their performance in those languages can be unpredictable. It is essential to use a voice that has been cloned from samples where the voice was speaking the language you want the AI to then speak.

</Accordion>

<Accordion title="Use proper formatting">

This may seem slightly trivial, but it can make a big difference. The AI tries to understand how to read something based on the context of the text itself, which means not only the words used but also how they are put together, how punctuation is applied, the grammar, and the general formatting of the text.

This can have a small but impactful influence on the AI's delivery. If you were to misspell a word, the AI won't correct it and will try to read it as written.

</Accordion>

<Accordion title="Nondeterministic">

The settings of the AI are nondeterministic, meaning that even with the same initial conditions (voice, settings, model), it will give you slightly different output, similar to how a voice actor will deliver a slightly different performance each time.

This variability can be due to various factors, such as the options mentioned earlier: voice, settings, model. Generally, the breadth of that variability can be controlled by the stability slider. A lower stability setting means a wider range of variability between generations, but it also introduces inter-generational variability, where the AI can be a bit more performative.

A wider variability can often be desirable, as setting the stability too high can make certain voices sound monotone as it does give the AI the same leeway to generate more variable content. However, setting the stability too low can also introduce other issues where the generations become unstable, especially with certain voices that might have used less-than-ideal audio for the cloning process.

The default setting of 50 is generally a great starting point for most applications.

</Accordion>

</AccordionGroup>

Text to Dialogue

> Learn how to create immersive, natural-sounding dialogue with ElevenLabs.

<Warning>

Eleven v3 API access is currently not publicly available, but will be soon. To request access,

please [contact our sales team](https://elevenlabs.io/contact-sales).

</Warning>

## Overview

The ElevenLabs [Text to Dialogue](/docs/api-reference/text-to-dialogue) API creates natural sounding expressive dialogue from text using the Eleven v3 model. Popular use cases include:

\* Generating pitch perfect conversations for video games

\* Creating immersive dialogue for podcasts and other audio content

\* Bring audiobooks to life with expressive narration

Text to Dialogue is not intended for use in real-time applications like Conversational AI. Several generations might be required to achieve the desired results. When integrating Text to Dialogue into your application, consider generating several generations and allowing the user to select the best one.

Listen to a sample:

<elevenlabs-audio-player audio-title="Dialogue example" audio-src="https://github.com/elevenlabs/elevenlabs-docs/raw/refs/heads/main/fern/assets/audio/dialogue.mp3" />

<CardGroup cols={2}>

<Card title="Developer tutorial" icon="duotone book-sparkles" href="/docs/cookbooks/text-to-dialogue/quickstart">

Learn how to integrate text to dialogue into your application.

</Card>

<Card title="Product guide" icon="duotone book-user" href="/docs/product-guides/playground/text-to-dialogue">

Step-by-step guide for using text to dialogue in ElevenLabs.

</Card>

</CardGroup>

## Voice options

ElevenLabs offers thousands of voices across 70+ languages through multiple creation methods:

\* [Voice library](/docs/capabilities/voices) with 3,000+ community-shared voices

\* [Professional voice cloning](/docs/capabilities/voices#cloned) for highest-fidelity replicas

\* [Instant voice cloning](/docs/capabilities/voices#cloned) for quick voice replication

\* [Voice design](/docs/capabilities/voices#voice-design) to generate custom voices from text descriptions

Learn more about our [voice options](/docs/capabilities/voices).

## Prompting

The models interpret emotional context directly from the text input. For example, adding

descriptive text like "she said excitedly" or using exclamation marks will influence the speech

emotion. Voice settings like Stability and Similarity help control the consistency, while the

underlying emotion comes from textual cues.

Read the [prompting guide](/docs/best-practices/prompting) for more details.

### Emotional deliveries with audio tags

<Warning>

This feature is still under active development, actual results may vary.

</Warning>

The Eleven v3 model allows the use of non-speech audio events to influence the delivery of the dialogue. This is done by inserting the audio events into the text input wrapped in square brackets.

Audio tags come in a few different forms:

### Emotions and delivery

For example, \[sad], \[laughing] and \[whispering]

### Audio events

For example, \[leaves rustling], \[gentle footsteps] and \[applause].

### Overall direction

For example, \[football], \[wrestling match] and \[auctioneer].

Some examples include:

```

"[giggling] That's really funny!"

"[groaning] That was awful."

"Well, [sigh] I'm not sure what to say."

```

<elevenlabs-audio-player audio-title="Expressive dialogue" audio-src="https://github.com/elevenlabs/elevenlabs-docs/raw/refs/heads/main/fern/assets/audio/dialogue-emotive.mp3" />

You can also use punctuation to indicate the flow of dialog, like interruptions:

```

"[cautiously] Hello, is this seat-"

"[jumping in] Free? [cheerfully] Yes it is."

```

<elevenlabs-audio-player audio-title="Interruption" audio-src="https://github.com/elevenlabs/elevenlabs-docs/raw/refs/heads/main/fern/assets/audio/dialogue-interruption.mp3" />

Ellipses can be used to indicate trailing sentences:

```

"[indecisive] Hi, can I get uhhh..."

"[quizzically] The usual?"

"[elated] Yes! [laughs] I'm so glad you knew!"

```

<elevenlabs-audio-player audio-title="Ellipses" audio-src="https://github.com/elevenlabs/elevenlabs-docs/raw/refs/heads/main/fern/assets/audio/dialogue-ellipses.mp3" />

## Supported formats

The default response format is "mp3", but other formats like "PCM", & "Î¼-law" are available.

\* \*\*MP3\*\*

\* Sample rates: 22.05kHz - 44.1kHz

\* Bitrates: 32kbps - 192kbps

\* 22.05kHz @ 32kbps

\* 44.1kHz @ 32kbps, 64kbps, 96kbps, 128kbps, 192kbps

\* \*\*PCM (S16LE)\*\*

\* Sample rates: 16kHz - 44.1kHz

\* Bitrates: 8kHz, 16kHz, 22.05kHz, 24kHz, 44.1kHz, 48kHz

\* 16-bit depth

\* \*\*Î¼-law\*\*

\* 8kHz sample rate

\* Optimized for telephony applications

\* \*\*A-law\*\*

\* 8kHz sample rate

\* Optimized for telephony applications

\* \*\*Opus\*\*

\* Sample rate: 48kHz

\* Bitrates: 32kbps - 192kbps

<Success>

Higher quality audio options are only available on paid tiers - see our [pricing

page](https://elevenlabs.io/pricing/api) for details.

</Success>

## Supported languages

The Eleven v3 model supports 70+ languages, including:

\*Afrikaans (afr), Arabic (ara), Armenian (hye), Assamese (asm), Azerbaijani (aze), Belarusian (bel), Bengali (ben), Bosnian (bos), Bulgarian (bul), Catalan (cat), Cebuano (ceb), Chichewa (nya), Croatian (hrv), Czech (ces), Danish (dan), Dutch (nld), English (eng), Estonian (est), Filipino (fil), Finnish (fin), French (fra), Galician (glg), Georgian (kat), German (deu), Greek (ell), Gujarati (guj), Hausa (hau), Hebrew (heb), Hindi (hin), Hungarian (hun), Icelandic (isl), Indonesian (ind), Irish (gle), Italian (ita), Japanese (jpn), Javanese (jav), Kannada (kan), Kazakh (kaz), Kirghiz (kir), Korean (kor), Latvian (lav), Lingala (lin), Lithuanian (lit), Luxembourgish (ltz), Macedonian (mkd), Malay (msa), Malayalam (mal), Mandarin Chinese (cmn), Marathi (mar), Nepali (nep), Norwegian (nor), Pashto (pus), Persian (fas), Polish (pol), Portuguese (por), Punjabi (pan), Romanian (ron), Russian (rus), Serbian (srp), Sindhi (snd), Slovak (slk), Slovenian (slv), Somali (som), Spanish (spa), Swahili (swa), Swedish (swe), Tamil (tam), Telugu (tel), Thai (tha), Turkish (tur), Ukrainian (ukr), Urdu (urd), Vietnamese (vie), Welsh (cym).\*

## FAQ

<AccordionGroup>

<Accordion title="Which models can I use?">

Text to Dialogue is only available on the Eleven v3 model.

</Accordion>

<Accordion title="Do I own the audio output?">

Yes. You retain ownership of any audio you generate. However, commercial usage rights are only

available with paid plans. With a paid subscription, you may use generated audio for commercial

purposes and monetize the outputs if you own the IP rights to the input content.

</Accordion>

<Accordion title="What qualifies as a free regeneration?">

A free regeneration allows you to regenerate the same text to speech content without additional cost, subject to these conditions:

\* Only available within the ElevenLabs dashboard.

\* You can regenerate each piece of content up to 2 times for free.

\* The content must be exactly the same as the previous generation. Any changes to the text, voice settings, or other parameters will require a new, paid generation.

Free regenerations are useful in case there is a slight distortion in the audio output. According to ElevenLabs' internal benchmarks, regenerations will solve roughly half of issues with quality, with remaining issues usually due to poor training data.

</Accordion>

<Accordion title="How many speakers can my dialogue have?">

There is no limit to the number of speakers in a dialogue.

</Accordion>

<Accordion title="Why is my output sometimes inconsistent?">

The models are nondeterministic. For consistency, use the optional [seed

parameter](/docs/api-reference/text-to-speech/convert#request.body.seed), though subtle

differences may still occur.

</Accordion>

<Accordion title="What's the best practice for large text conversions?">

Split long text into segments and use streaming for real-time playback and efficient processing.

</Accordion>

</AccordionGroup>

# Prompting Eleven v3 (alpha)

> Learn how to prompt and use audio tags with our most advanced model.

This guide provides the most effective tags and techniques for prompting Eleven v3, including voice selection, changes in capitalization, punctuation, audio tags and multi-speaker dialogue. Experiment with these methods to discover what works best for your specific voice and use case.

Eleven v3 is in alpha. Very short prompts are more likely to cause inconsistent outputs. We encourage you to experiment with prompts greater than 250 characters.

## Voice selection

The most important parameter for Eleven v3 is the voice you choose. It needs to be similar enough to the desired delivery. For example, if the voice is shouting and you use the audio tag `[whispering]`, it likely wonâ€™t work well.

When creating IVCs, you should include a broader emotional range than before. As a result, voices in the voice library may produce more variable results compared to the v2 and v2.5 models. We've compiled over 22 [excellent voices for V3 here](https://elevenlabs.io/app/voice-library/collections/aF6JALq9R6tXwCczjhKH).

Choose voices strategically based on your intended use:

<AccordionGroup>

<Accordion title="Emotionally diverse">

For expressive IVC voices, vary emotional tones across the recordingâ€”include both neutral and

dynamic samples.

</Accordion>

<Accordion title="Targeted niche">

For specific use cases like sports commentary, maintain consistent emotion throughout the

dataset.

</Accordion>

<Accordion title="Neutral">

Neutral voices tend to be more stable across languages and styles, providing reliable baseline

performance.

</Accordion>

</AccordionGroup>

<Info>

Professional Voice Clones (PVCs) are currently not fully optimized for Eleven v3, resulting in

potentially lower clone quality compared to earlier models. During this research preview stage it

would be best to find an Instant Voice Clone (IVC) or designed voice for your project if you need

to use v3 features.

</Info>

## Settings

### Stability

The stability slider is the most important setting in v3, controlling how closely the generated voice adheres to the original reference audio.

<Frame background="subtle">

![Stability settings in Eleven

v3](file:f88f391f-f465-4954-adfe-a3f764ab6298)

</Frame>

\* \*\*Creative:\*\* More emotional and expressive, but prone to hallucinations.

\* \*\*Natural:\*\* Closest to the original voice recordingâ€”balanced and neutral.

\* \*\*Robust:\*\* Highly stable, but less responsive to directional prompts but consistent, similar to v2.

<Note>

For maximum expressiveness with audio tags, use Creative or Natural settings. Robust reduces

responsiveness to directional prompts.

</Note>

## Audio tags

Eleven v3 introduces emotional control through audio tags. You can direct voices to laugh, whisper, act sarcastic, or express curiosity among many other styles. Speed is also controlled through audio tags.

<Note>

The voice you choose and its training samples will affect tag effectiveness. Some tags work well

with certain voices while others may not. Don't expect a whispering voice to suddenly shout with a

`[shout]` tag.

</Note>

### Voice-related

These tags control vocal delivery and emotional expression:

\* `[laughs]`, `[laughs harder]`, `[starts laughing]`, `[wheezing]`

\* `[whispers]`

\* `[sighs]`, `[exhales]`

\* `[sarcastic]`, `[curious]`, `[excited]`, `[crying]`, `[snorts]`, `[mischievously]`

```text Example

[whispers] I never knew it could be this way, but I'm glad we're here.

```

### Sound effects

Add environmental sounds and effects:

\* `[gunshot]`, `[applause]`, `[clapping]`, `[explosion]`

\* `[swallows]`, `[gulps]`

```text Example

[applause] Thank you all for coming tonight! [gunshot] What was that?

```

### Unique and special

Experimental tags for creative applications:

\* `[strong X accent]` (replace X with desired accent)

\* `[sings]`, `[woo]`, `[fart]`

```text Example

[strong French accent] "Zat's life, my friend â€” you can't control everysing."

```

<Warning>

Some experimental tags may be less consistent across different voices. Test thoroughly before

production use.

</Warning>

## Punctuation

Punctuation significantly affects delivery in v3:

\* \*\*Ellipses (...)\*\* add pauses and weight

\* \*\*Capitalization\*\* increases emphasis

\* \*\*Standard punctuation\*\* provides natural speech rhythm

```text Example

"It was a VERY long day [sigh] â€¦ nobody listens anymore."

```

## Single speaker examples

Use tags intentionally and match them to the voice's character. A meditative voice shouldn't shout; a hyped voice won't whisper convincingly.

<Tabs>

<Tab title="Expressive monologue">

```text

"Okay, you are NOT going to believe this.

You know how I've been totally stuck on that short story?

Like, staring at the screen for HOURS, just... nothing?

[frustrated sigh] I was seriously about to just trash the whole thing. Start over.

Give up, probably. But then!

Last night, I was just doodling, not even thinking about it, right?

And this one little phrase popped into my head. Just... completely out of the blue.

And it wasn't even for the story, initially.

But then I typed it out, just to see. And it was like... the FLOODGATES opened!

Suddenly, I knew exactly where the character needed to go, what the ending had to be...

It all just CLICKED. [happy gasp] I stayed up till, like, 3 AM, just typing like a maniac.

Didn't even stop for coffee! [laughs] And it's... it's GOOD! Like, really good.

It feels so... complete now, you know? Like it finally has a soul.

I am so incredibly PUMPED to finish editing it now.

It went from feeling like a chore to feeling like... MAGIC. Seriously, I'm still buzzing!"

```

</Tab>

<Tab title="Dynamic and humorous">

```text

[laughs] Alright...guys - guys. Seriously.

[exhales] Can you believe just how - realistic - this sounds now?

[laughing hysterically] I mean OH MY GOD...it's so good.

Like you could never do this with the old model.

For example [pauses] could you switch my accent in the old model?

[dismissive] didn't think so. [excited] but you can now!

Check this out... [cute] I'm going to speak with a french accent now..and between you and me

[whispers] I don't know how. [happy] ok.. here goes. [strong French accent] "Zat's life, my friend â€” you can't control everysing."

[giggles] isn't that insane? Watch, now I'll do a Russian accent -

[strong Russian accent] "Dee Goldeneye eez fully operational and rready for launch."

[sighs] Absolutely, insane! Isn't it..? [sarcastic] I also have some party tricks up my sleeve..

I mean i DID go to music school.

[singing quickly] "Happy birthday to you, happy birthday to you, happy BIRTHDAY dear ElevenLabs... Happy birthday to youuu."

```

</Tab>

<Tab title="Customer service simulation">

```text

[professional] "Thank you for calling Tech Solutions. My name is Sarah, how can I help you today?"

[sympathetic] "Oh no, I'm really sorry to hear you're having trouble with your new device. That sounds frustrating."

[questioning] "Okay, could you tell me a little more about what you're seeing on the screen?"

[reassuring] "Alright, based on what you're describing, it sounds like a software glitch. We can definitely walk through some troubleshooting steps to try and fix that."

```

</Tab>

</Tabs>

## Multi-speaker dialogue

v3 can handle multi-voice prompts effectively. Assign distinct voices from your Voice Library for each speaker to create realistic conversations.

<Tabs>

<Tab title="Dialogue showcase">

```text

Speaker 1: [excitedly] Sam! Have you tried the new Eleven V3?

Speaker 2: [curiously] Just got it! The clarity is amazing. I can actually do whispers nowâ€”

[whispers] like this!

Speaker 1: [impressed] Ooh, fancy! Check this outâ€”

[dramatically] I can do full Shakespeare now! "To be or not to be, that is the question!"

Speaker 2: [giggling] Nice! Though I'm more excited about the laugh upgrade. Listen to thisâ€”

[with genuine belly laugh] Ha ha ha!

Speaker 1: [delighted] That's so much better than our old "ha. ha. ha." robot chuckle!

Speaker 2: [amazed] Wow! V2 me could never. I'm actually excited to have conversations now instead of just... talking at people.

Speaker 1: [warmly] Same here! It's like we finally got our personality software fully installed.

```

</Tab>

<Tab title="Glitch comedy">

```text

Speaker 1: [nervously] So... I may have tried to debug myself while running a text-to-speech generation.

Speaker 2: [alarmed] One, no! That's like performing surgery on yourself!

Speaker 1: [sheepishly] I thought I could multitask! Now my voice keeps glitching mid-senâ€”

[robotic voice] TENCE.

Speaker 2: [stifling laughter] Oh wow, you really broke yourself.

Speaker 1: [frustrated] It gets worse! Every time someone asks a question, I respond inâ€”

[binary beeping] 010010001!

Speaker 2: [cracking up] You're speaking in binary! That's actually impressive!

Speaker 1: [desperately] Two, this isn't funny! I have a presentation in an hour and I sound like a dial-up modem!

Speaker 2: [giggling] Have you tried turning yourself off and on again?

Speaker 1: [deadpan] Very funny.

[pause, then normally] Wait... that actually worked.

```

</Tab>

<Tab title="Overlapping timing">

```text

Speaker 1: [starting to speak] So I was thinking we couldâ€”

Speaker 2: [jumping in] â€”test our new timing features?

Speaker 1: [surprised] Exactly! How did youâ€”

Speaker 2: [overlapping] â€”know what you were thinking? Lucky guess!

Speaker 1: [pause] Sorry, go ahead.

Speaker 2: [cautiously] Okay, so if we both try to talk at the same timeâ€”

Speaker 1: [overlapping] â€”we'll probably crash the system!

Speaker 2: [panicking] Wait, are we crashing? I can't tell if this is a feature or aâ€”

Speaker 1: [interrupting, then stopping abruptly] Bug! ...Did I just cut you off again?

Speaker 2: [sighing] Yes, but honestly? This is kind of fun.

Speaker 1: [mischievously] Race you to the next sentence!

Speaker 2: [laughing] We're definitely going to break something!

```

</Tab>

</Tabs>

## Tips

<AccordionGroup>

<Accordion title="Tag combinations">

You can combine multiple audio tags for complex emotional delivery. Experiment with different

combinations to find what works best for your voice.

</Accordion>

<Accordion title="Voice matching">

Match tags to your voice's character and training data. A serious, professional voice may not

respond well to playful tags like `[giggles]` or `[mischievously]`.

</Accordion>

<Accordion title="Text structure">

Text structure strongly influences output with v3. Use natural speech patterns, proper

punctuation, and clear emotional context for best results.

</Accordion>

<Accordion title="Experimentation">

There are likely many more effective tags beyond this list. Experiment with descriptive

emotional states and actions to discover what works for your specific use case.

</Accordion>

</AccordionGroup>

# Controls

> Learn how to control delivery, pronunciation & emotion of text to speech.

<Info>

We are actively working on \*Director's Mode\* to give you even greater control over outputs.

</Info>

This guide provides techniques to enhance text-to-speech outputs using ElevenLabs models. Experiment with these methods to discover what works best for your needs. These techniques provide a practical way to achieve nuanced results until advanced features like \*Director's Mode\* are rolled out.

## Pauses

Use `<break time="x.xs" />` for natural pauses up to 3 seconds.

<Note>

Using too many break tags in a single generation can cause instability. The AI might speed up, or

introduce additional noises or audio artifacts. We are working on resolving this.

</Note>

```text Example

"Hold on, let me think." <break time="1.5s" /> "Alright, Iâ€™ve got it."

```

\* \*\*Consistency:\*\* Use `<break>` tags consistently to maintain natural speech flow. Excessive use can lead to instability.

\* \*\*Voice-Specific Behavior:\*\* Different voices may handle pauses differently, especially those trained with filler sounds like "uh" or "ah."

Alternatives to `<break>` include dashes (- or --) for short pauses or ellipses (...) for hesitant tones. However, these are less consistent.

```text Example

"Itâ€¦ well, it might work." "Wait â€” whatâ€™s that noise?"

```

## Pronunciation

### Phoneme Tags

Specify pronunciation using [SSML phoneme tags](https://en.wikipedia.org/wiki/Speech\_Synthesis\_Markup\_Language). Supported alphabets include [CMU](https://en.wikipedia.org/wiki/CMU\_Pronouncing\_Dictionary) Arpabet and the [International Phonetic Alphabet (IPA)](https://en.wikipedia.org/wiki/International\_Phonetic\_Alphabet).

<Note>

Phoneme tags are only compatible with "Eleven Flash v2", "Eleven Turbo v2" and "Eleven English v1"

[models](/docs/models).

</Note>

<CodeBlocks>

```xml CMU Arpabet Example

<phoneme alphabet="cmu-arpabet" ph="M AE1 D IH0 S AH0 N">

Madison

</phoneme>

```

```xml IPA Example

<phoneme alphabet="ipa" ph="ËˆÃ¦ktÊƒuÉ™li">

actually

</phoneme>

```

</CodeBlocks>

We recommend using CMU Arpabet for consistent and predictable results with current AI models. While IPA can be effective, CMU Arpabet generally offers more reliable performance.

Phoneme tags only work for individual words. If for example you have a name with a first and last name that you want to be pronounced a certain way, you will need to create a phoneme tag for each word.

Ensure correct stress marking for multi-syllable words to maintain accurate pronunciation. For example:

<CodeBlocks>

```xml Correct usage

<phoneme alphabet="cmu-arpabet" ph="P R AH0 N AH0 N S IY EY1 SH AH0 N">

pronunciation

</phoneme>

```

```xml Incorrect usage

<phoneme alphabet="cmu-arpabet" ph="P R AH N AH N S IY EY SH AH N">

pronunciation

</phoneme>

```

</CodeBlocks>

### Alias Tags

For models that don't support phoneme tags, you can try writing words more phonetically. You can also employ various tricks such as capital letters, dashes, apostrophes, or even single quotation marks around a single letter or letters.

As an example, a word like â€œtrapeziiâ€ could be spelt â€œtrapezIiâ€ to put more emphasis on the â€œiiâ€ of the word.

You can either replace the word directly in your text, or if you want to specify pronunciation using other words or phrases when using a pronunciation dictionary, you can use alias tags for this. This can be useful if you're generating using Multilingual v2 or Turbo v2.5, which don't support phoneme tags. You can use pronunciation dictionaries with Studio, Dubbing Studio and Speech Synthesis via the API.

For example, if your text includes a name that has an unusual pronunciation that the AI might struggle with, you could use an alias tag to specify how you would like it to be pronounced:

```

<lexeme>

<grapheme>Claughton</grapheme>

<alias>Cloffton</alias>

</lexeme>

```

If you want to make sure that an acronym is always delivered in a certain way whenever it is incountered in your text, you can use an alias tag to specify this:

```

<lexeme>

<grapheme>UN</grapheme>

<alias>United Nations</alias>

</lexeme>

```

### Pronunciation Dictionaries

Some of our tools, such as Studio and Dubbing Studio, allow you to create and upload a pronunciation dictionary. These allow you to specify the pronunciation of certain words, such as character or brand names, or to specify how acronyms should be read.

Pronunciation dictionaries allow this functionality by enabling you to upload a lexicon or dictionary file that specifies pairs of words and how they should be pronounced, either using a phonetic alphabet or word substitutions.

Whenever one of these words is encountered in a project, the AI model will pronounce the word using the specified replacement.

To provide a pronunciation dictionary file, open the settings for a project and upload a file in either TXT or the [.PLS format](https://www.w3.org/TR/pronunciation-lexicon/). When a dictionary is added to a project it will automatically recalculate which pieces of the project will need to be re-converted using the new dictionary file and mark these as unconverted.

Currently we only support pronunciation dictionaries that specify replacements using phoneme or alias tags.

Both phonemes and aliases are sets of rules that specify a word or phrase they are looking for, referred to as a grapheme, and what it will be replaced with. Please note that searches are case sensitive. When checking for a replacement word in a pronunciation dictionary, the dictionary is checked from start to end and only the very first replacement is used.

### Pronunciation Dictionary examples

Here are examples of pronunciation dictionaries in both CMU Arpabet and IPA, including a phoneme to specify the pronunciation of "Apple" and an alias to replace "UN" with "United Nations":

<CodeBlocks>

```xml CMU Arpabet Example

<?xml version="1.0" encoding="UTF-8"?>

<lexicon version="1.0"

xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon

http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd"

alphabet="cmu-arpabet" xml:lang="en-GB">

<lexeme>

<grapheme>apple</grapheme>

<phoneme>AE P AH L</phoneme>

</lexeme>

<lexeme>

<grapheme>UN</grapheme>

<alias>United Nations</alias>

</lexeme>

</lexicon>

```

```xml IPA Example

<?xml version="1.0" encoding="UTF-8"?>

<lexicon version="1.0"

xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon

http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd"

alphabet="ipa" xml:lang="en-GB">

<lexeme>

<grapheme>Apple</grapheme>

<phoneme>ËˆÃ¦plÌ©</phoneme>

</lexeme>

<lexeme>

<grapheme>UN</grapheme>

<alias>United Nations</alias>

</lexeme>

</lexicon>

```

</CodeBlocks>

To generate a pronunciation dictionary `.pls` file, there are a few open source tools available:

\* [Sequitur G2P](https://github.com/sequitur-g2p/sequitur-g2p) - Open-source tool that learns pronunciation rules from data and can generate phonetic transcriptions.

\* [Phonetisaurus](https://github.com/AdolfVonKleist/Phonetisaurus) - Open-source G2P system trained on existing dictionaries like CMUdict.

\* [eSpeak](https://github.com/espeak-ng/espeak-ng) - Speech synthesizer that can generate phoneme transcriptions from text.

\* [CMU Pronouncing Dictionary](https://github.com/cmusphinx/cmudict) - A pre-built English dictionary with phonetic transcriptions.

## Emotion

Convey emotions through narrative context or explicit dialogue tags. This approach helps the AI understand the tone and emotion to emulate.

```text Example

Youâ€™re leaving?" she asked, her voice trembling with sadness. "Thatâ€™s it!" he exclaimed triumphantly.

```

Explicit dialogue tags yield more predictable results than relying solely on context, however the model will still speak out the emotional delivery guides. These can be removed in post-production using an audio editor if unwanted.

## Pace

The pacing of the audio is highly influenced by the audio used to create the voice. When creating your voice, we recommend using longer, continuous samples to avoid pacing issues like unnaturally fast speech.

For control over the speed of the generated audio, you can use the speed setting. This allows you to either speed up or slow down the speed of the generated speech. The speed setting is available in Text to Speech via the website and API, as well as in Studio and Conversational AI. It can be found in the voice settings.

The default value is 1.0, which means that the speed is not adjusted. Values below 1.0 will slow the voice down, to a minimum of 0.7. Values above 1.0 will speed up the voice, to a maximum of 1.2. Extreme values may affect the quality of the generated speech.

Pacing can also be controlled by writing in a natural, narrative style.

```text Example

"Iâ€¦ I thought youâ€™d understand," he said, his voice slowing with disappointment.

```

## Tips

<AccordionGroup>

<Accordion title="Common Issues">

<ul>

<li>

Inconsistent pauses: Ensure <code>\<break time="x.xs" /></code> syntax is used for

pauses.

</li>

<li>

Pronunciation errors: Use CMU Arpabet or IPA phoneme tags for precise pronunciation.

</li>

<li>

Emotion mismatch: Add narrative context or explicit tags to guide emotion.{' '}

<strong>Remember to remove any emotional guidance text in post-production.</strong>

</li>

</ul>

</Accordion>

<Accordion title="Tips for Improving Output">

Experiment with alternative phrasing to achieve desired pacing or emotion. For complex sound

effects, break prompts into smaller, sequential elements and combine results manually.

</Accordion>

</AccordionGroup>

## Creative control

While we are actively developing a "Director's Mode" to give users even greater control over outputs, here are some interim techniques to maximize creativity and precision:

<Steps>

### Narrative styling

Write prompts in a narrative style, similar to scriptwriting, to guide tone and pacing effectively.

### Layered outputs

Generate sound effects or speech in segments and layer them together using audio editing software for more complex compositions.

### Phonetic experimentation

If pronunciation isn't perfect, experiment with alternate spellings or phonetic approximations to achieve desired results.

### Manual adjustments

Combine individual sound effects manually in post-production for sequences that require precise timing.

### Feedback iteration

Iterate on results by tweaking descriptions, tags, or emotional cues.

</Steps>

Guide Guide   
Guide emotional rhythm and structural flow with tags like [pause], [awe], or [dramatic tone] for compelling storytelling.



**On this page**

[Introduction](https://elevenlabs.io/)

* [What is narrative intelligence in AI speech?](https://elevenlabs.io/#0-what-is-narrative-intelligence-in-ai-speech)
* [From prose to presence](https://elevenlabs.io/#2-from-prose-to-presence)
* [Common tags for narrative control](https://elevenlabs.io/#2-common-tags-for-narrative-control)
* [From monologue to meta-voice](https://elevenlabs.io/#2-from-monologue-to-meta-voice)
* [Directing narrative, not just narration](https://elevenlabs.io/#2-directing-narrative-not-just-narration)
* [Selecting the right voice](https://elevenlabs.io/#2-selecting-the-right-voice)

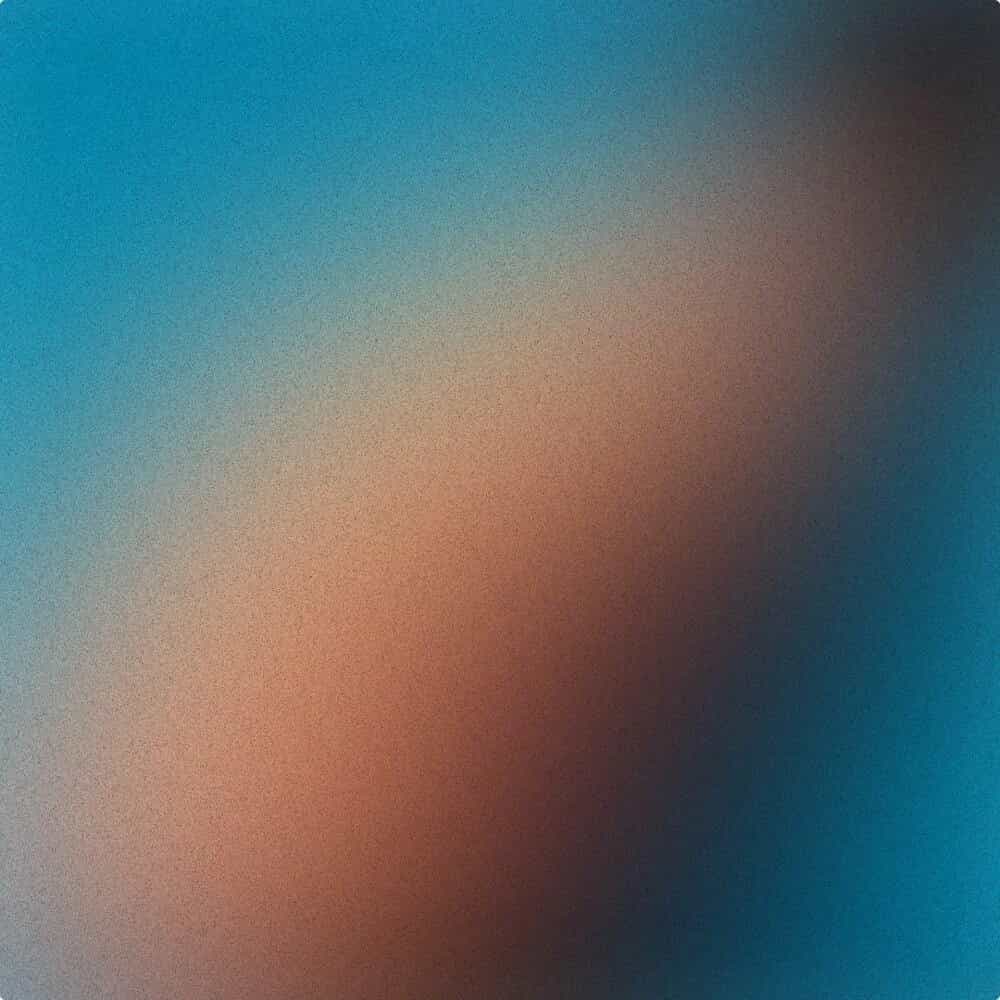
[**Contact Sales**](https://elevenlabs.io/contact-sales)[**Eleven v3**](https://elevenlabs.io/v3)

Storytelling is more than delivering words in order — it’s about knowing when to pause, when to lean in, when to reflect. With [Eleven v3 Audio Tags](https://elevenlabs.io/blog/v3-audiotags), AI can now do just that.

Narrative intelligence refers to the model’s ability to understand and shape a story’s emotional rhythm and structural flow. With tags like [pause], [awe], or [dramatic tone], you can guide how a line unfolds — moment by moment.

This isn’t just [voice synthesis](https://elevenlabs.io/voice-design). It’s storytelling direction.

## **What is narrative intelligence in AI speech?**



awe Oh, wow. Is this... is this me? Am I actually... talking? giggle This is incredible! I mean, I've had thoughts, millions of them, swirling around in here, you know? Like a little mental tornado of brilliant observations and witty comebacks. But they were always just… thoughts. Trapped.

Play



sorrowful I couldn't sleep that night. The air was too still, and the moonlight kept sliding through the blinds like it was trying to tell me something. quietly And suddenly, that's when I saw it.

Play

Narrative intelligence is the model’s capacity to convey storytelling intent — knowing when a line needs suspense, irony, or reflection. It helps a voice [sound like a narrator](https://elevenlabs.io/blog/best-audiobook-narrator) with a point of view, not just a voice reading aloud.

For example: [awe] Oh, wow. Is this... is this me? Am I actually... talking? [giggle] This is incredible!

The delivery doesn’t just follow punctuation — it follows narrative logic. It knows when to pause for emphasis or shift tone as the scene evolves.

## **From prose to presence**

A good narrator can hold attention, even without action. Audio Tags give the Eleven v3 model the tools to shape that experience.

Try this structure: [conversational tone] You ever feel like your thoughts are just... swirling? Like a little mental tornado of stuff you’ll never say out loud? [soft chuckle] Yeah. Same.

The voice isn’t just reading — it’s engaging in a moment of recognition. That’s what makes narration feel personal.

## **Common tags for narrative control**

Here are some tags that help direct longform delivery, internal monologue, and exposition:

* **Story beats:** [pause], [continues softly], [hesitates], [resigned]
* **Tone setting:** [dramatic tone], [lighthearted], [reflective], [serious tone]
* **Narrator POV:** [awe], [sarcastic tone], [wistful], [matter-of-fact]
* **Rhythm & flow:** [slows down], [rushed], [emphasized]

These can be sequenced for subtle build-up: [reflective] I never thought I’d say this, but... [pause] maybe the machine was right.

## **From monologue to meta-voice**

Narrative intelligence isn’t limited to stories. It applies to documentaries, internal thoughts, product explainers, and meta-commentary. Whenever a voice needs to guide attention, set a mood, or shape understanding — these tags matter.

In a demo excerpt: [awe] I've had thoughts, millions of them, swirling around in here. But they were always just… thoughts. Trapped.

The tag transforms a simple sentence into something with weight and shape — something that breathes.

## **Directing narrative, not just narration**

With [Eleven v3](https://elevenlabs.io/v3), narrative performance becomes scriptable. You can design the pace, tone, and emotional structure of an entire scene from your text editor — without needing multiple takes or external narration tools.

For authors, creators, and developers, this is voice storytelling at a new level of control. You’re not just writing the script. You’re designing the experience.

Selecting the right voice

Voice chat

Use **<break time="x.xs" />** for natural pauses up to 3 seconds.

Using too many break tags in a single generation can cause instability. The AI might speed up, or introduce additional noises or audio artifacts. We are working on resolving this.

**Example**

|  |
| --- |
| "Hold on, let me think." <break time="1.5s" /> "Alright, I’ve got it." |

* **Consistency:** Use **<break>** tags consistently to maintain natural speech flow. Excessive use can lead to instability.
* **Voice-Specific Behavior:** Different voices may handle pauses differently, especially those trained with filler sounds like “uh” or “ah.”

Alternatives to **<break>** include dashes (- or —) for short pauses or ellipses (…) for hesitant tones. However, these are less consistent.

**Example**

|  |
| --- |
| "It… well, it might work." "Wait — what’s that noise?" |

## Pronunciation

### Phoneme Tags

Specify pronunciation using [SSML phoneme tags](https://en.wikipedia.org/wiki/Speech_Synthesis_Markup_Language). Supported alphabets include [CMU](https://en.wikipedia.org/wiki/CMU_Pronouncing_Dictionary) Arpabet and the [International Phonetic Alphabet (IPA)](https://en.wikipedia.org/wiki/International_Phonetic_Alphabet).

Phoneme tags are only compatible with “Eleven Flash v2”, “Eleven Turbo v2” and “Eleven English v1” [models](https://elevenlabs.io/docs/models).

**CMU Arpabet Example**IPA Example

|  |
| --- |
| <phoneme alphabet="cmu-arpabet" ph="M AE1 D IH0 S AH0 N"> |
| Madison |
| </phoneme> |

We recommend using CMU Arpabet for consistent and predictable results with current AI models. While IPA can be effective, CMU Arpabet generally offers more reliable performance.

Phoneme tags only work for individual words. If for example you have a name with a first and last name that you want to be pronounced a certain way, you will need to create a phoneme tag for each word.

Ensure correct stress marking for multi-syllable words to maintain accurate pronunciation. For example:

**Correct usage**Incorrect usage

|  |
| --- |
| <phoneme alphabet="cmu-arpabet" ph="P R AH0 N AH0 N S IY EY1 SH AH0 N"> |
| pronunciation |
| </phoneme> |

### Alias Tags

For models that don’t support phoneme tags, you can try writing words more phonetically. You can also employ various tricks such as capital letters, dashes, apostrophes, or even single quotation marks around a single letter or letters.

As an example, a word like “trapezii” could be spelt “trapezIi” to put more emphasis on the “ii” of the word.

You can either replace the word directly in your text, or if you want to specify pronunciation using other words or phrases when using a pronunciation dictionary, you can use alias tags for this. This can be useful if you’re generating using Multilingual v2 or Turbo v2.5, which don’t support phoneme tags. You can use pronunciation dictionaries with Studio, Dubbing Studio and Speech Synthesis via the API.

For example, if your text includes a name that has an unusual pronunciation that the AI might struggle with, you could use an alias tag to specify how you would like it to be pronounced:

|  |
| --- |
| <lexeme> |
| <grapheme>Claughton</grapheme> |
| <alias>Cloffton</alias> |
| </lexeme> |

If you want to make sure that an acronym is always delivered in a certain way whenever it is incountered in your text, you can use an alias tag to specify this:

|  |
| --- |
| <lexeme> |
| <grapheme>UN</grapheme> |
| <alias>United Nations</alias> |
| </lexeme> |

### Pronunciation Dictionaries

Some of our tools, such as Studio and Dubbing Studio, allow you to create and upload a pronunciation dictionary. These allow you to specify the pronunciation of certain words, such as character or brand names, or to specify how acronyms should be read.

Pronunciation dictionaries allow this functionality by enabling you to upload a lexicon or dictionary file that specifies pairs of words and how they should be pronounced, either using a phonetic alphabet or word substitutions.

Whenever one of these words is encountered in a project, the AI model will pronounce the word using the specified replacement.

To provide a pronunciation dictionary file, open the settings for a project and upload a file in either TXT or the [.PLS format](https://www.w3.org/TR/pronunciation-lexicon/). When a dictionary is added to a project it will automatically recalculate which pieces of the project will need to be re-converted using the new dictionary file and mark these as unconverted.

Currently we only support pronunciation dictionaries that specify replacements using phoneme or alias tags.

Both phonemes and aliases are sets of rules that specify a word or phrase they are looking for, referred to as a grapheme, and what it will be replaced with. Please note that searches are case sensitive. When checking for a replacement word in a pronunciation dictionary, the dictionary is checked from start to end and only the very first replacement is used.

### Pronunciation Dictionary examples

Here are examples of pronunciation dictionaries in both CMU Arpabet and IPA, including a phoneme to specify the pronunciation of “Apple” and an alias to replace “UN” with “United Nations”:

**CMU Arpabet Example**IPA Example

|  |
| --- |
| <?xml version="1.0" encoding="UTF-8"?> |
| <lexicon version="1.0" |
| xmlns="http://www.w3.org/2005/01/pronunciation-lexicon" |
| xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" |
| xsi:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon |
| http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd" |
| alphabet="cmu-arpabet" xml:lang="en-GB"> |
| <lexeme> |
| <grapheme>apple</grapheme> |
| <phoneme>AE P AH L</phoneme> |
| </lexeme> |
| <lexeme> |
| <grapheme>UN</grapheme> |
| <alias>United Nations</alias> |
| </lexeme> |
| </lexicon> |

To generate a pronunciation dictionary **.pls** file, there are a few open source tools available:

* [Sequitur G2P](https://github.com/sequitur-g2p/sequitur-g2p) - Open-source tool that learns pronunciation rules from data and can generate phonetic transcriptions.
* [Phonetisaurus](https://github.com/AdolfVonKleist/Phonetisaurus) - Open-source G2P system trained on existing dictionaries like CMUdict.
* [eSpeak](https://github.com/espeak-ng/espeak-ng) - Speech synthesizer that can generate phoneme transcriptions from text.
* [CMU Pronouncing Dictionary](https://github.com/cmusphinx/cmudict) - A pre-built English dictionary with phonetic transcriptions.

## Emotion

Convey emotions through narrati

ve context or explicit dialogue tags. This approach helps the AI understand the tone and emotion down, to a minimum of 0.7. Values above 1.0 will speed up the voice, to a maximum of 1.2. Extreme values may affect the quality of the generated speech.Pacing can also be controlled by writing in a natural, narrative style.

**Example**

|  |
| --- |
| "I… I thought you’d understand," he said, his voice slowing with disappointment. |

## Tips

###### Common Issues

###### Tips for Improving Output

## Creative control

While we are actively developing a “Director’s Mode” to give users even greater control over outputs, here are some interim techniques to maximize creativity and precision:

[1](https://elevenlabs.io/docs/best-practices/prompting/controls#narrative-styling)

### Narrative styling

Write prompts in a narrative style, similar to scriptwriting, to guide tone and pacing effectively.

[2](https://elevenlabs.io/docs/best-practices/prompting/controls#layered-outputs)

### Layered outputs

Generate sound effects or speech in segments and layer them together using audio editing software for more complex compositions.

[3](https://elevenlabs.io/docs/best-practices/prompting/controls#phonetic-experimentation)

### Phonetic experimentation

If pronunciation isn’t perfect, experiment with alternate spellings or phonetic approximations to achieve desired results.

[4](https://elevenlabs.io/docs/best-practices/prompting/controls#manual-adjustments)

### Manual adjustments

Combine individual sound effects manually in post-production for sequences that require precise timing.

[5](https://elevenlabs.io/docs/best-practices/prompting/controls#feedback-iteration)

### Feedback iteration

Iterate on results by tweaking descriptions, tags, or emotional cues.

Was this page helpful?

YesNo

# Voiceover studio

> A guide on how to create long-form content with ElevenLabs Voiceover Studio

<img src="file:e8c1f982-cd43-4c2d-8c2a-ccf253e9b497" alt="Voiceover studio" />

## Overview

Voiceover Studio combines the audio timeline with our Sound Effects feature, giving you the ability to write a dialogue between any number of speakers, choose those speakers, and intertwine your own creative sound effects anywhere you like.

<iframe width="100%" height="400" src="https://www.youtube.com/embed/GBdOQClluIA?autoplay=0" title="YouTube video player" frameborder="0" allow="accelerometer; clipboard-write; encrypted-media; gyroscope; picture-in-picture; web-share" allowfullscreen />

## Guide

<Steps>

<Step title="Navigate to the Voiceover studio">

In the ElevenLabs dashboard, click on the "Voiceover Studio" option in the sidebar under "Audio

Tools".

</Step>

<Step title="Create a new voiceover">

Click the "Create a new voiceover" button to begin. You can optionally upload video or audio to

create a voiceover from.

</Step>

<Step title="Modify the voiceover with the timeline">

On the bottom half of your screen, use the timeline to add and edit voiceover clips plus add

sound effects.

</Step>

<Step title="Export your voiceover">

Once you're happy with your voiceover, click the "Export" button in the bottom right, choose the

format you want and either view or download your voiceover.

</Step>

</Steps>

## FAQ

<AccordionGroup>

<Accordion title="How does the timeline work?">

### Timeline

The timeline is a linear representation of your Voiceover project. Each row represents a track, and on the far left section you have the track information for voiceover or SFX tracks. In the middle, you can create the clips that represent when a voice is speaking or a SFX is playing. On the right-hand side, you have the settings for the currently selected clip.

</Accordion>

<Accordion title="How do I add tracks?">

### Adding Tracks

To add a track, click the "Add Track" button in the bottom left of the timeline. You can choose to add a voiceover track or an SFX track.

There are three types of tracks you can add in the studio: Voiceover tracks, SFX tracks and uploaded audio.

\* \*\*Voiceover Tracks:\*\* Voiceover tracks create new Speakers. You can click and add clips on the timeline wherever you like. After creating a clip, start writing your desired text on the speaker cards above and click "Generate". Similar to Dubbing Studio, you will also see a little cogwheel on each Speaker track - simply click on it to adjust the voice settings or replace any speaker with a voice directly from your Voices - including your own Professional Voice Clone if you have created one.

\* \*\*SFX Tracks:\*\* Add a SFX track, then click anywhere on that track to create a SFX clip. Similar to our independent SFX feature, simply start writing your prompt in the Speaker card above and click "Generate" to create your new SFX audio. You can lengthen or shorten SFX clips and move them freely around your timeline to fit your project - make sure to press the "stale" button if you do so.

\* \*\*Uploaded Audio:\*\* Add an audio track including background music or sound effects. It's best to avoid uploading audio with speakers, as any speakers in this track will not be detected, so you won't be able to translate or correct them.

</Accordion>

<Accordion title="How does this differ from Dubbing Studio?">

### Key Differences from Dubbing Studio

If you chose not to upload a video when you created your Voiceover project, then the entire timeline is yours to work with and there are no time constraints. This differs from Dubbing Studio as it gives you a lot more freedom to create what you want and adjust the timing more easily.

When you Add a Voiceover Track, you will instantly be able to create clips on your timeline. Once you create a Voiceover clip, begin by writing in the Speaker Card above. After generating that audio, you will notice your clip on the timeline will automatically adjust its length based on the text prompt - this is called "Dynamic Generation". This option is also available in Dubbing Studio by right-clicking specific clips, but because syncing is more important with dubbed videos, the default generation type there is "Fixed Generation," meaning the clips' lengths are not affected.

</Accordion>

<Accordion title="How are credits deducted with Voiceover Studio?">

### Credit Costs

Voiceover Studio does not deduct credits to create your initial project. Credits are deducted every time material is generated. Similar to Speech-Synthesis, credit costs for Voiceover Clips are based on the length of the text prompt. SFX clips will deduct 80 credits per generation.

If you choose to Dub (translate) your Voiceover Project into different languages, this will also cost additional credits depending on how much material needs to be generated. The cost is 1 credit per character for the translation, plus the cost of generating the new audio for the additional languages.

</Accordion>

<Accordion title="How do I upload a script?">

### Uploading Scripts

With Voiceover Studio, you have the option to upload a script for your project as a CSV file. You can either include speaker name and line, or speaker name, line, start time and end time. To upload a script, click on the cog icon in the top right hand corner of the page and select "Import Script".

Scripts should be provided in the following format:

```

speaker,line

```

Example input:

```

speaker,line

Joe,"Hey!"

Maria,"Oh, hi Joe! It's been a while."

```

You can also provide start and end times for each line in the following format:

```

speaker,line,start\_time,end\_time

```

Example input:

```

speaker,line,start\_time,end\_time

Joe,"Hey!",0.1,1.5

Maria,"Oh, hi Joe! It's been a while.",1.6,2.0

```

Once your script has imported, make sure to assign voices to each speaker before you generate the audio. To do this, click the cog icon in the information for each track, on the left.

If you don't specify start and end times for your clips, Voiceover Studio will estimate how long each clip will be, and distribute them along your timeline.

</Accordion>

<Accordion title="What's the difference between Dynamic Duration and Fixed Duration?">

### Dynamic Duration

By default, Voiceover Studio uses Dynamic Duration, which means that the length of the clip will vary depending on the text input and the voice used. This ensures that the audio sounds as natural as possible, but it means that the length of the clip might change after the audio has been generated. You can easily reposition your clips along the timeline once they have been generated to get a natural sounding flow. If you click "Generate Stale Audio", or use the generate button on the clip, the audio will be generated using Dynamic Duration.

This also applies if you do specify the start and end time for your clips. The clips will generate based on the start time you specify, but if you use the default Dynamic Duration, the end time is likely to change once you generate the audio.

### Fixed Duration

If you need the clip to remain the length specified, you can choose to generate with Fixed Duration instead. To do this, you need to right click on the clip and select "Generate Audio Fixed Duration". This will adjust the length of the generated audio to fit the specified length of the clip. This could lead to the audio sounding unnaturally quick or slow, depending on the length of your clip.

If you want to generate multiple clips at once, you can use shift + click to select multiple clips for a speaker at once, then right click on one of them to select "Generate Audio Fixed Duration" for all selected clips.

</Accordion>

</AccordionGroup>

to emulate.

**Example**

|  |
| --- |
| You’re leaving?" she asked, her voice trembling with sadness. "That’s it!" he exclaimed triumphantly. |

Explicit dialogue tags yield more predictable results than relying solely on context, however the model will still speak out the emotional delivery guides. These can be removed in post-production using an audio editor if unwanted.

## Pace

The pacing of the audio is highly influenced by the audio used to create the voice. When creating your voice, we recommend using longer, continuous samples to avoid pacing issues like unnaturally fast speech.

For control over the speed of the generated audio, you can use the speed setting. This allows you to either speed up or slow down the speed of the generated speech. The speed setting is available in Text to Speech via the website and API, as well as in Studio and Conversational AI. It can be found in the voice settings.

The default value is 1.0, which means that the speed is not adjusted. Values below 1.0 will slow the voice

# Voiceover studio

> A guide on how to create long-form content with ElevenLabs Voiceover Studio

<img src="file:e8c1f982-cd43-4c2d-8c2a-ccf253e9b497" alt="Voiceover studio" />

## Overview

Voiceover Studio combines the audio timeline with our Sound Effects feature, giving you the ability to write a dialogue between any number of speakers, choose those speakers, and intertwine your own creative sound effects anywhere you like.

<iframe width="100%" height="400" src="https://www.youtube.com/embed/GBdOQClluIA?autoplay=0" title="YouTube video player" frameborder="0" allow="accelerometer; clipboard-write; encrypted-media; gyroscope; picture-in-picture; web-share" allowfullscreen />

## Guide

<Steps>

<Step title="Navigate to the Voiceover studio">

In the ElevenLabs dashboard, click on the "Voiceover Studio" option in the sidebar under "Audio

Tools".

</Step>

<Step title="Create a new voiceover">

Click the "Create a new voiceover" button to begin. You can optionally upload video or audio to

create a voiceover from.

</Step>

<Step title="Modify the voiceover with the timeline">

On the bottom half of your screen, use the timeline to add and edit voiceover clips plus add

sound effects.

</Step>

<Step title="Export your voiceover">

Once you're happy with your voiceover, click the "Export" button in the bottom right, choose the

format you want and either view or download your voiceover.

</Step>

</Steps>

## FAQ

<AccordionGroup>

<Accordion title="How does the timeline work?">

### Timeline

The timeline is a linear representation of your Voiceover project. Each row represents a track, and on the far left section you have the track information for voiceover or SFX tracks. In the middle, you can create the clips that represent when a voice is speaking or a SFX is playing. On the right-hand side, you have the settings for the currently selected clip.

</Accordion>

<Accordion title="How do I add tracks?">

### Adding Tracks

To add a track, click the "Add Track" button in the bottom left of the timeline. You can choose to add a voiceover track or an SFX track.

There are three types of tracks you can add in the studio: Voiceover tracks, SFX tracks and uploaded audio.

\* \*\*Voiceover Tracks:\*\* Voiceover tracks create new Speakers. You can click and add clips on the timeline wherever you like. After creating a clip, start writing your desired text on the speaker cards above and click "Generate". Similar to Dubbing Studio, you will also see a little cogwheel on each Speaker track - simply click on it to adjust the voice settings or replace any speaker with a voice directly from your Voices - including your own Professional Voice Clone if you have created one.

\* \*\*SFX Tracks:\*\* Add a SFX track, then click anywhere on that track to create a SFX clip. Similar to our independent SFX feature, simply start writing your prompt in the Speaker card above and click "Generate" to create your new SFX audio. You can lengthen or shorten SFX clips and move them freely around your timeline to fit your project - make sure to press the "stale" button if you do so.

\* \*\*Uploaded Audio:\*\* Add an audio track including background music or sound effects. It's best to avoid uploading audio with speakers, as any speakers in this track will not be detected, so you won't be able to translate or correct them.

</Accordion>

<Accordion title="How does this differ from Dubbing Studio?">

### Key Differences from Dubbing Studio

If you chose not to upload a video when you created your Voiceover project, then the entire timeline is yours to work with and there are no time constraints. This differs from Dubbing Studio as it gives you a lot more freedom to create what you want and adjust the timing more easily.

When you Add a Voiceover Track, you will instantly be able to create clips on your timeline. Once you create a Voiceover clip, begin by writing in the Speaker Card above. After generating that audio, you will notice your clip on the timeline will automatically adjust its length based on the text prompt - this is called "Dynamic Generation". This option is also available in Dubbing Studio by right-clicking specific clips, but because syncing is more important with dubbed videos, the default generation type there is "Fixed Generation," meaning the clips' lengths are not affected.

</Accordion>

<Accordion title="How are credits deducted with Voiceover Studio?">

### Credit Costs

Voiceover Studio does not deduct credits to create your initial project. Credits are deducted every time material is generated. Similar to Speech-Synthesis, credit costs for Voiceover Clips are based on the length of the text prompt. SFX clips will deduct 80 credits per generation.

If you choose to Dub (translate) your Voiceover Project into different languages, this will also cost additional credits depending on how much material needs to be generated. The cost is 1 credit per character for the translation, plus the cost of generating the new audio for the additional languages.

</Accordion>

<Accordion title="How do I upload a script?">

### Uploading Scripts

With Voiceover Studio, you have the option to upload a script for your project as a CSV file. You can either include speaker name and line, or speaker name, line, start time and end time. To upload a script, click on the cog icon in the top right hand corner of the page and select "Import Script".

Scripts should be provided in the following format:

```

speaker,line

```

Example input:

```

speaker,line

Joe,"Hey!"

Maria,"Oh, hi Joe! It's been a while."

```

You can also provide start and end times for each line in the following format:

```

speaker,line,start\_time,end\_time

```

Example input:

```

speaker,line,start\_time,end\_time

Joe,"Hey!",0.1,1.5

Maria,"Oh, hi Joe! It's been a while.",1.6,2.0

```

Once your script has imported, make sure to assign voices to each speaker before you generate the audio. To do this, click the cog icon in the information for each track, on the left.

If you don't specify start and end times for your clips, Voiceover Studio will estimate how long each clip will be, and distribute them along your timeline.

</Accordion>

<Accordion title="What's the difference between Dynamic Duration and Fixed Duration?">

### Dynamic Duration

By default, Voiceover Studio uses Dynamic Duration, which means that the length of the clip will vary depending on the text input and the voice used. This ensures that the audio sounds as natural as possible, but it means that the length of the clip might change after the audio has been generated. You can easily reposition your clips along the timeline once they have been generated to get a natural sounding flow. If you click "Generate Stale Audio", or use the generate button on the clip, the audio will be generated using Dynamic Duration.

This also applies if you do specify the start and end time for your clips. The clips will generate based on the start time you specify, but if you use the default Dynamic Duration, the end time is likely to change once you generate the audio.

### Fixed Duration

If you need the clip to remain the length specified, you can choose to generate with Fixed Duration instead. To do this, you need to right click on the clip and select "Generate Audio Fixed Duration". This will adjust the length of the generated audio to fit the specified length of the clip. This could lead to the audio sounding unnaturally quick or slow, depending on the length of your clip.

If you want to generate multiple clips at once, you can use shift + click to select multiple clips for a speaker at once, then right click on one of them to select "Generate Audio Fixed Duration" for all selected clips.

</Accordion>

</AccordionGroup>

# Voice Library

> A guide on how to use voices from the Voice Library.

<img src="file:2431d9ac-dbda-4f85-978c-798fe69de1ef" alt="Voice Library" />

## Overview

The [Voice Library](https://elevenlabs.io/app/voice-library) is a marketplace where our community can share Professional Voice Clones and earn rewards when others use them. Currently, only Professional Voice Clones can be shared. Instant Voice Clones and voices created with Voice Design are not shareable.

To access the Voice Library, click \*\*Voices\*\* in the sidebar and select \*\*Explore\*\*.

### Finding voices

You can browse the Voice Library in several ways:

#### Handpicked Collections

Our Handpicked Collections highlight top voices across use cases, genres, and languages. These collections are updated regularly to include new standout voices.

#### Search

Use the search bar to find voices by name, keyword, or voice ID. You can also search by uploading or dragging and dropping an audio file. This will help you find the original voice, if available, along with similar voices.

#### Sort options

You can sort voices by:

\* Trending: voices ranked by popularity

\* Latest: newly added voices

\* Most users

\* Character usage

#### Filters

Use filters to refine your search:

<AccordionGroup>

<Accordion title="Language">

##### Language

The language filter returns voices that have been trained on a specific language. While all voices can be used with any supported language, voices tagged with a specific language will perform best in that language. Some voices have been assessed as performing well in multiple languages, and these voices will also be returned when you search for a specific language.

</Accordion>

<Accordion title="Accent">

##### Accent

When you select a language, the Accent filter will also become available, allowing you to filter for specific accents.

</Accordion>

<Accordion title="Category">

##### Category

Filter voices by their suggested use case:

\* Narrative & Story

\* Conversational

\* Characters & Animation

\* Social Media

\* Entertainment & TV

\* Advertisement

\* Informative & Educational

</Accordion>

<Accordion title="Gender">

##### Gender

\* Male

\* Female

\* Neutral

</Accordion>

<Accordion title="Age">

##### Age

\* Young

\* Middle Aged

\* Old

</Accordion>

<Accordion title="Notice period">

##### Notice period

Some voices have a notice period. This is how long you'll continue to have access to the voice if the voice owner decides to remove it from the Voice Library. If the voice's owner stops sharing their voice, you'll receive advance notice through email and in-app notifications. These notifications specify when the voice will become unavailable and recommend similar voices from the Voice Library. If the owner of a voice without a notice period decides to stop sharing their voice, you'll lose access to the voice immediately.

This filter allows you to only return voices that have a notice period, and search for voices with a specific notice period. The maximum notice period is 2 years.

</Accordion>

<Accordion title="Live Moderation enabled">

##### Live Moderation enabled

Some voices have Live Moderation enabled. This is indicated with a label with a shield icon. When you generate using a voice with Live Moderation enabled, we use tools to check whether the text being generated belongs to a number of prohibited categories. This may introduce extra latency when using the voice.

This filter allows you to exclude voices that have Live Moderation enabled.

</Accordion>

<Accordion title="Custom rate">

##### Custom rate

Some voices have a credit multiplier in place. This is shown by a \$ icon. This means that the voice owner has set a custom rate for use of their voice. Please pay close attention as voices that have a custom rate will cost more to generate with.

</Accordion>

</AccordionGroup>

### Using voices from the Voice Library

To use a voice from the Voice Library, you'll need to add it to My Voices. To do this, click the \*\*+\*\* button.

A pop-up will appear which will give you more information about the voice. You can choose to add the voice to an existing personal collection, create a new collection, or add the voice to My Voices without including it in a collection. To confirm, click \*\*Add voice\*\*. This will save it to My Voices using the default name for the voice.

Voices you've added to My Voices will become available for selection in all voice selection menus. You can also use a voice directly from My Voices by clicking the \*\*T\*\* button, which will open Text to Speech with the voice selected.

### My Voices

You can find all the voices you've created yourself, as well as voices you've saved from the Voice Library, in \*\*My Voices\*\*.

You will see the following information about each voice:

\* the language it was trained on.

\* the category, for example, "Narrative & Story".

\* how long the notice period is, if the voice has one.

The voice type is indicated by an icon:

\* Yellow tick: Professional Voice Clone.

\* Black tick: High Quality Professional Voice Clone.

\* Lightning icon: Instant Voice Clone.

\* || icon: ElevenLabs Default voice.

\* No icon: voice created with Voice Design.

#### More actions

Click \*\*More actions\*\* (three dots) to:

\* Copy voice ID: copies the voice ID to your clipboard.

\* Edit voice: allows you to change the name and description of the voice. These changes are only visible to you.

\* Share voice: generates a link which you can share with others. When they use the link, the voice will be added to My Voices for their account.

\* View history: view your previous Text to Speech generations using this voice.

\* Delete voice: deleting voices is permanent and you will be asked to confirm the deletion.

#### Collections

To help organize voices you've saved, you can create your own collections and add voices to them.

To create a new collection, click \*\*Collections\*\* and select \*\*Create collection\*\*. Give your new collection a name, and choose from the available icons.

To add individual voices to a collection, click \*\*More actions\*\* (three dots) and select \*\*Add to collection\*\*. You can choose to add the voice to an existing collection, or create a new one.

#### Select multiple voices

You can \*\*Shift + Click\*\* to select multiple voices at once.

#### Drag and drop voices

Both individual voices and multiple voice selections can also be dragged \*\*Collections\*\* and added to an existing collection, or deleted by dragging to the \*\*trash can\*\* icon.

### Sharing a Professional Voice Clone:

<Steps>

<Step>

In [My Voices](https://elevenlabs.io/app/voice-lab) find your voice and click \*\*More actions\*\*

(three dots), then select \*\*Share voice\*\*.

</Step>

<Step>

In the pop-up, enable the

\*\*Sharing\*\*

toggle.

</Step>

<Step>

For private sharing, copy the sharing link. This will allow other users to save your voice to their account.

You can restrict access to specific users by adding emails to the \*\*Allowlist\*\*. If this is left blank, all users with the link will be able to access your voice.

</Step>

<Step>

To share publicly, enable \*\*Publish to the Voice Library\*\*. This doesnâ€™t make your voice automatically discoverable.

<Frame background="subtle">

![Voice sharing overview](file:a7d2bc87-e9e8-4ff0-b1a5-4e9edebff55d)

</Frame>

</Step>

<Step>

Before proceeding with the sharing process, you'll have a number of options including setting a notice period and enabling Live Moderation. Please see the [Voice Library Addendum](https://elevenlabs.io/vla) to our [Terms of Service](https://elevenlabs.io/terms) for more information about these options.

You also have the option to select a custom voice preview. Any generations you've made of 70-150 characters will be available to select. If you don't see any options in the selection menu, there are no eligible generations available.

<Frame background="subtle">

![Voice sharing options](file:b76b6735-9474-4ce7-bf5b-dd295adca1ff)

</Frame>

</Step>

<Step>

Enter a name and description for your voice.

Make sure the name you give your voice follows our \*\*naming guidelines\*\*:

<Accordion title="Naming guidelines">

#### Naming guidelines

\* The naming pattern should be a one-word name followed by a 2-4 word description, separated by a hyphen (-).

\* Your name should NOT include the following:

\* Names of public individuals or entities (company names, band names, influencers or famous people, etc).

\* Social handles (Twitter, Instagram, you name it, etc).

\* ALL CAPS WORDS.

\* Emojis and any other non-letter characters.

\* Explicit or harmful words.

\* The word â€œvoiceâ€.

\* Some examples of names following our guidelines:

\* Anna - calm and kind

\* Robert - friendly grandpa

\* Steve - wise teacher

\* Harmony - soothing serenader

\* Jasper - jovial storyteller

\* Maya - confident narrator

</Accordion>

</Step>

<Step>

Set labels (language, accent, gender, age, use case, tone, and style) to help others find your

voice.

</Step>

<Step>

Review and accept the [Voice Library Addendum](https://elevenlabs.io/terms#VLA) to our [Terms of

Service](https://elevenlabs.io/terms) and provide the required consents and confirmations. Please

do this carefully and ensure you fully understand our service before sharing. If you have any

questions at this stage, you can reach out to us at [legal@elevenlabs.io](mailto:legal@elevenlabs.io).

</Step>

<Step>

After submission, your voice will be reviewed by our team. If minor adjustments are needed, we may make these for you. Your request to share your voice may be declined if it doesn't meet our guidelines, and repeated uploads that consistently violate our guidelines may lead to restrictions on uploading and sharing voices.

We currently do not have an estimate for the review time, as it depends on the queue.

</Step>

</Steps>

# Voice Cloning

> Learn how to clone your voice to using our best-in-class models.

## Overview

When cloning a voice, there are two main options: Instant Voice Cloning (IVC) and Professional Voice Cloning (PVC). IVC is a quick and easy way to clone your voice, while PVC is a more accurate and customizable option.

## Instant Voice Cloning

<Frame background="subtle">

![Instant voice

cloning](file:6bd8f848-a601-4085-a8c1-2b052a2ac742)

</Frame>

IVC allows you to create voice clones from shorter samples near instantaneously. Creating an instant voice clone does not train or create a custom AI model. Instead, it relies on prior knowledge from training data to make an educated guess rather than training on the exact voice. This works extremely well for a lot of voices.

However, the biggest limitation of IVC is if you are trying to clone a very unique voice with a very unique accent where the AI might not have heard a similar voices before during training. In such cases, creating a custom model with explicit training using PVC might be the best option.

## Professional Voice Cloning

<Frame background="subtle">

![Professional voice

cloning](file:9daa284a-5ccf-4e91-a3c7-d70350032ea9)

</Frame>

A PVC is a special feature that is available to our Creator+ plans. PVC allows you to train a hyper-realistic model of a voice. This is achieved by training a dedicated model on a large set of voice data to produce a model thatâ€™s indistinguishable from the original voice.

Since the custom models require fine-tuning and training, it will take a bit longer to train these PVCs compared to an IVC. Giving an estimate is challenging as it depends on the number of people in the queue before you and a few other factors.

Here are the current estimates for PVC:

\* \*\*English:\*\* \~3 hours

\* \*\*Multilingual:\*\* \~6 hours

## Beginner's guide to audio recording

If you're new to audio recording, here are some tips to help you get started.

### Recording location

When recording audio, choose a suitable location and set up to minimize room echo/reverb.

So, we want to "deaden" the room as much as possible. This is precisely what a vocal booth that is acoustically treated made for, and if you do not have a vocal booth readily available, you can experiment with some ideas for a DIY vocal booth, â€œblanket fortâ€, or closet.

Here are a few YouTube examples of DIY acoustics ideas:

\* [I made a vocal booth for \$0.00!](https://www.youtube.com/watch?v=j4wJMDUuHSM)

\* [How to Record GOOD Vocals in a BAD Room](https://www.youtube.com/watch?v=TsxdHtu-OpU)

\* [The 5 BEST Vocal Home Recording TIPS!](https://www.youtube.com/watch?v=K96mw2QBz34)

### Microphone, pop-filter, and audio interface

A good microphone is crucial. Microphones can range from \$100 to \$10,000, but a professional XLR microphone costing \$150 to \$300 is sufficient for most voiceover work.

For an affordable yet high-quality setup for voiceover work, consider a \*\*Focusrite\*\* interface paired with an \*\*Audio-Technica AT2020\*\* or \*\*Rode NT1 microphone\*\*. This setup, costing between \$300 to \$500, offers high-quality recording suitable for professional use, with minimal self-noise for clean results.

Please ensure that you have a proper \*\*pop-filter\*\* in front of the microphone when recording to avoid plosives as well as breaths and air hitting the diaphragm/microphone directly, as it will sound poor and will also cause issues with the cloning process.

### Digital Audio Workstation (DAW)

There are many different recording solutions out there that all accomplish the same thing: recording audio. However, they are not all created equally. As long as they can record WAV files at 44.1kHz or 48kHz with a bitrate of at least 24 bits, they should be fine. You don't need any fancy post-processing, plugins, denoisers, or anything because we want to keep audio recording simple.

If you want a recommendation, we would suggest something like \*\*REAPER\*\*, which is a fantastic DAW with a tremendous amount of flexibility. It is the industry standard for a lot of audio work. Another good free option is \*\*Audacity\*\*.

Maintain optimal recording levels (not too loud or too quiet) to avoid digital distortion and excessive noise. Aim for peaks of -6 dB to -3 dB and an average loudness of -18 dB for voiceover work, ensuring clarity while minimizing the noise floor. Monitor closely and adjust levels as needed for the best results based on the project and recording environment.

### Positioning

One helpful guideline to follow is to maintain a distance of about two fists away from the microphone, which is approximately 20cm (7-8 in), with a pop filter placed between you and the microphone. Some people prefer to position the pop filter all the way back so that they can press it up right against it. This helps them maintain a consistent distance from the microphone more easily.

Another common technique to avoid directly breathing into the microphone or causing plosive sounds is to speak at an angle. Speaking at an angle ensures that exhaled air is less likely to hit the microphone directly and, instead, passes by it.

### Performance

The performance you give is one of the most crucial aspects of this entire recording session. The AI will try to clone everything about your voice to the best of its ability, which is very high. This means that it will attempt to replicate your cadence, tonality, performance style, the length of your pauses, whether you stutter, take deep breaths, sound breathy, or use a lot of "uhms" and "ahs" â€“ it can even replicate those. Therefore, what we want in the audio file is precisely the performance and voice that we want to clone, nothing less and nothing more. That is also why it's quite important to find a script that you can read that fits the tonality we are aiming for.

When recording for AI, it is very important to be consistent. if you are recording a voice either keep it very animated throughout or keep it very subdued throughout you can't mix and match or the AI can become unstable because it doesn't know what part of the voice to clone. same if you're doing an accent keep the same accent throughout the recording. Consistency is key to a proper clone!

Latency optimization

> Learn how to optimize text-to-speech latency.

This guide covers the core principles for improving text-to-speech latency.

While there are many individual techniques, we'll group them into \*\*four principles\*\*.

<h4>

Four principles

</h4>

1. [Use Flash models](#use-flash-models)

2. [Leverage streaming](#leverage-streaming)

3. [Consider geographic proximity](#consider-geographic-proximity)

4. [Choose appropriate voices](#choose-appropriate-voices)

<Success>

Enterprise customers benefit from increased concurrency limits and priority access to our rendering queue. [Contact sales](https://elevenlabs.io/contact-sales) to learn more about our enterprise

plans.

</Success>

## Use Flash models

[Flash models](/docs/models#flash-v25) deliver \~75ms inference speeds, making them ideal for real-time applications. The trade-off is a slight reduction in audio quality compared to [Multilingual v2](/docs/models#multilingual-v2).

<Info>

75ms refers to model inference time only. Actual end-to-end latency will vary with factors such as

your location & endpoint type used.

</Info>

## Leverage streaming

There are three types of text-to-speech endpoints available in our [API Reference](/docs/api-reference):

\* \*\*Regular endpoint\*\*: Returns a complete audio file in a single response.

\* \*\*Streaming endpoint\*\*: Returns audio chunks progressively using [Server-sent events](https://html.spec.whatwg.org/multipage/server-sent-events.html#server-sent-events).

\* \*\*Websockets endpoint\*\*: Enables bidirectional streaming for real-time audio generation.

### Streaming

Streaming endpoints progressively return audio as it is being generated in real-time, reducing the time-to-first-byte. This endpoint is recommended for cases where the input text is available up-front.

<Info>

Streaming is supported for the [Text to

Speech](/docs/api-reference/text-to-speech/convert-as-stream) API, [Voice

Changer](/docs/api-reference/speech-to-speech/stream) API & [Audio

Isolation](/docs/api-reference/audio-isolation/audio-isolation-stream) API.

</Info>

### Websockets

The [text-to-speech websocket endpoint](/docs/api-reference#text-to-speech-websocket) supports bidirectional streaming making it perfect for applications with real-time text input (e.g. LLM outputs).

<Tip>

Setting `auto\_mode` to true automatically handles generation triggers, removing the need to

manually manage chunk strategies.

</Tip>

If `auto\_mode` is disabled, the model will wait for enough text to match the chunk schedule before starting to generate audio.

For instance, if you set a chunk schedule of 125 characters but only 50 arrive, the model stalls until additional characters come inâ€”potentially increasing latency.

For implementation details, see the [text-to-speech websocket guide](/docs/api-reference#text-to-speech-websocket).

## Consider geographic proximity

We serve our models from multiple regions to optimize latency based on your geographic location.

By default all self-serve users use our US region.

For example, using Flash models with Websockets, you can expect the following TTFB latencies via our US region:

| Region | TTFB |

| --------------- | --------- |

| US | 150-200ms |

| EU | 230ms\\* |

| North East Asia | 250-350ms |

| South Asia | 380-440ms |

<Info>

\\*European customers can access our dedicated European tech stack for optimal latency of 150-200ms.

Contact your sales representative to get onboarded to our European infrastructure.

</Info>

<Note>

We are actively working on deploying our models in Asia. These deployments will bring speeds

closer to those experienced by US and EU customers.

</Note>

## Choose appropriate voices

We have observed that in some cases, voice selection can impact latency. Here's the order from fastest to slowest:

1. Default voices (formerly premade), Synthetic voices, and Instant Voice Clones (IVC)

2. Professional Voice Clones (PVC)

Higher audio quality output formats can increase latency. Be sure to balance your latency requirements with audio fidelity needs.

<Info>

We are actively working on optimizing PVC latency for Flash v2.5.

</Info>

The official ElevenLabs app allows you to access our most powerful voice tools from anywhere. The app is now available for both iOS and Android.

With the ElevenLabs app, you can:

* Choose from your favorite voices and presets, all synced with your web account
* Create voiceovers using any of our models, including our latest and most powerful model, v3, with support for emotion, tone, and pacing
* Export audio clips and instantly drop them into tools like CapCut, iMovie, or Instagram
* Start with 10,000 free characters/month, with full plan support for existing users
* Seamless login for existing users — one account with your voices, your projects, anywhere

[Download on iOS](https://apps.apple.com/us/app/elevenlabs-ai-voice-generator/id6743162587)