Bark –

Text to Speech pretrain model

Mì AI

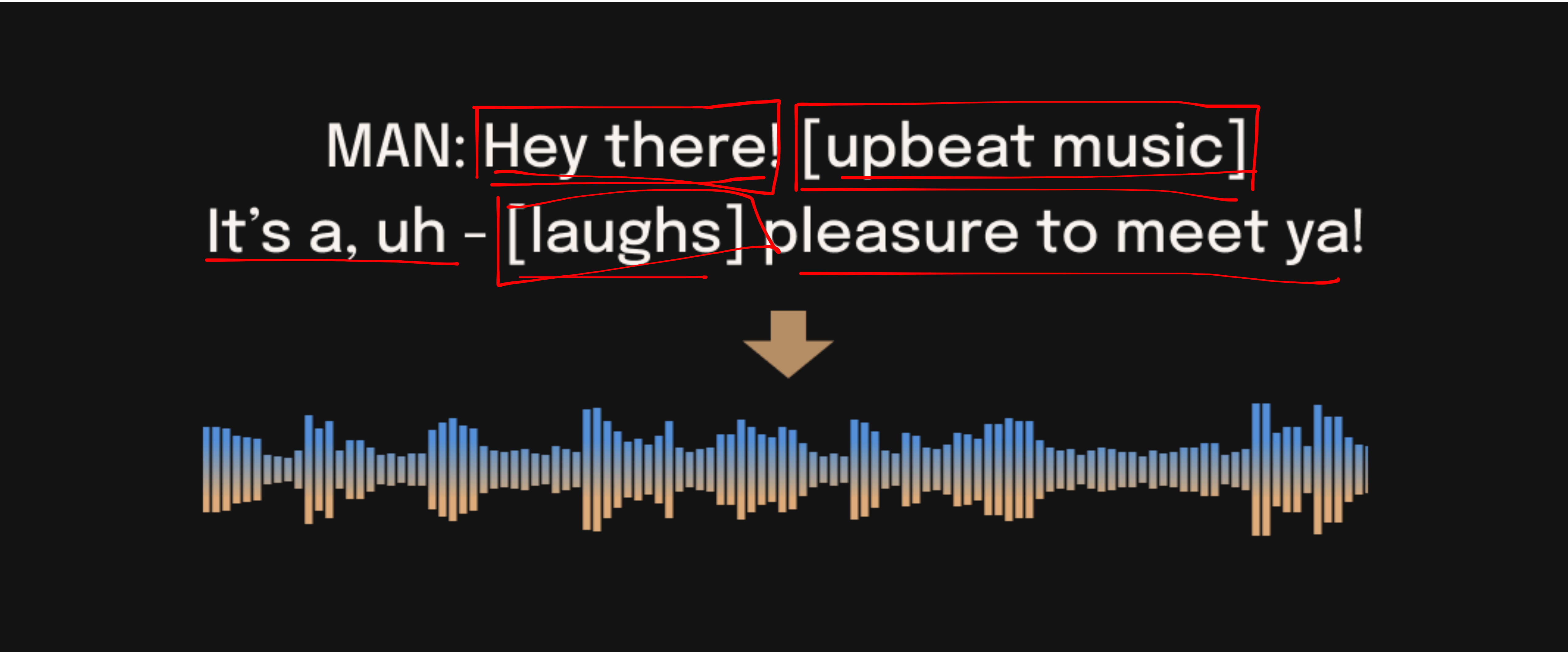
What is Bark?

- Bark is a transformer-based text-to-audio model created by Suno. Bark can generate highly realistic, multilingual speech as well as other audio - including music, background noise and simple sound effects. The model can also produce nonverbal communications like laughing, sighing and crying.
- To support the research community, we are providing access to pretrained model checkpoints, which are ready for inference and available for commercial use.

What is Bark?

- Bark was developed for research purposes. It is not a conventional text-to-speech model but instead a fully generative text-to-audio model, which can deviate in unexpected ways from provided prompts. Suno does not take responsibility for any output generated. Use at your own risk, and please act responsibly.

What is Bark?

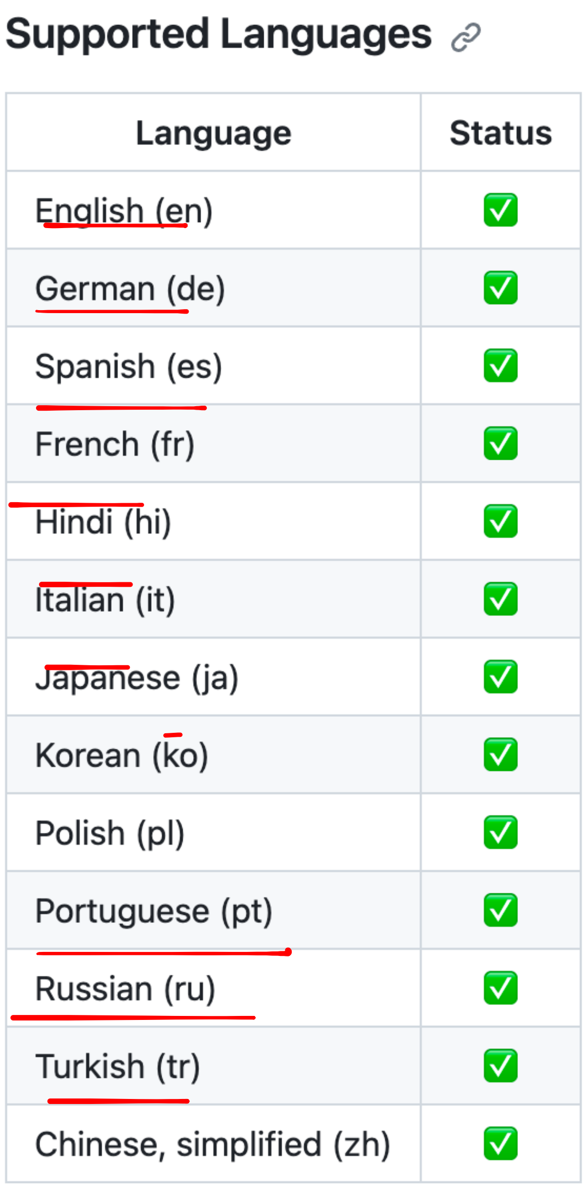


MAN: Hey there! [upbeat music]
It’s a, uh – [laughs] pleasure to meet ya!

What is Bark?

No training code

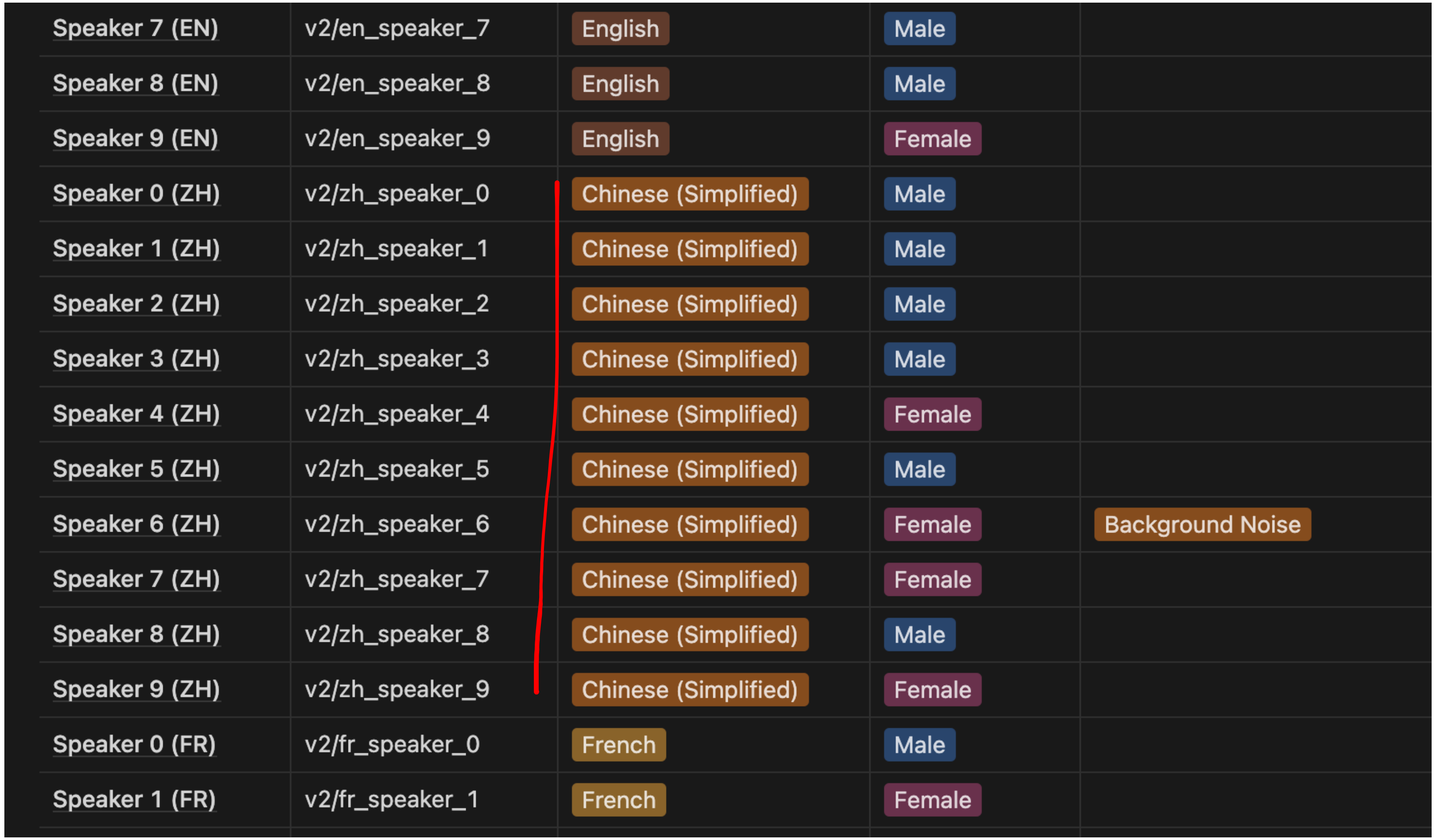
Support languages



Supported Languages

Language	Status
English (en)	✅
German (de)	✅
Spanish (es)	✅
French (fr)	✅
Hindi (hi)	✅
Italian (it)	✅
Japanese (ja)	✅
Korean (ko)	✅
Polish (pl)	✅
Portuguese (pt)	✅
Russian (ru)	✅
Turkish (tr)	✅
Chinese, simplified (zh)	✅

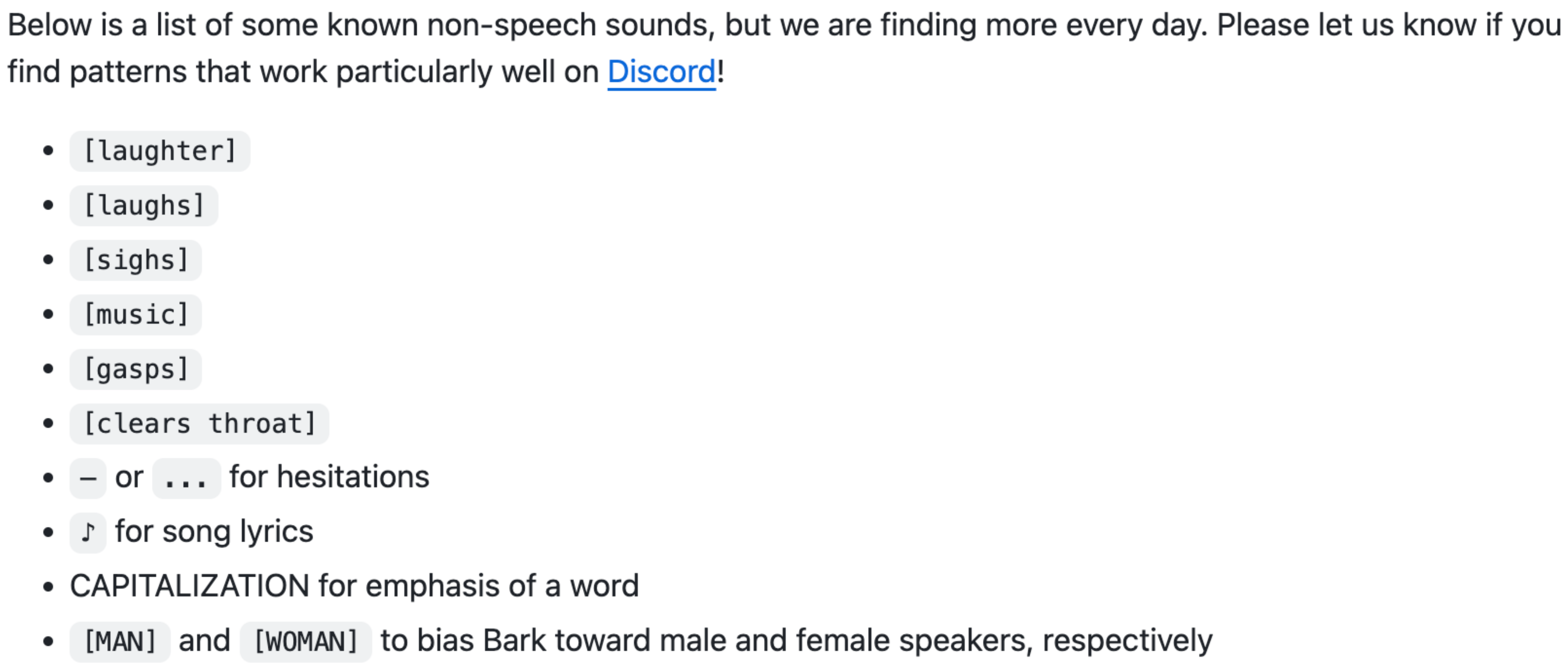
Support speaker



Speaker 7 (EN)	v2/en_speaker_7	English	Male	
Speaker 8 (EN)	v2/en_speaker_8	English	Male	
Speaker 9 (EN)	v2/en_speaker_9	English	Female	
Speaker 0 (ZH)	v2/zh_speaker_0	Chinese (Simplified)	Male	
Speaker 1 (ZH)	v2/zh_speaker_1	Chinese (Simplified)	Male	
Speaker 2 (ZH)	v2/zh_speaker_2	Chinese (Simplified)	Male	
Speaker 3 (ZH)	v2/zh_speaker_3	Chinese (Simplified)	Male	
Speaker 4 (ZH)	v2/zh_speaker_4	Chinese (Simplified)	Female	
Speaker 5 (ZH)	v2/zh_speaker_5	Chinese (Simplified)	Male	
Speaker 6 (ZH)	v2/zh_speaker_6	Chinese (Simplified)	Female	Background Noise
Speaker 7 (ZH)	v2/zh_speaker_7	Chinese (Simplified)	Female	
Speaker 8 (ZH)	v2/zh_speaker_8	Chinese (Simplified)	Male	
Speaker 9 (ZH)	v2/zh_speaker_9	Chinese (Simplified)	Female	
Speaker 0 (FR)	v2/fr_speaker_0	French	Male	
Speaker 1 (FR)	v2/fr_speaker_1	French	Female	

https://suno-ai.notion.site/8b8e8749ed514b0cbf3f699013548683

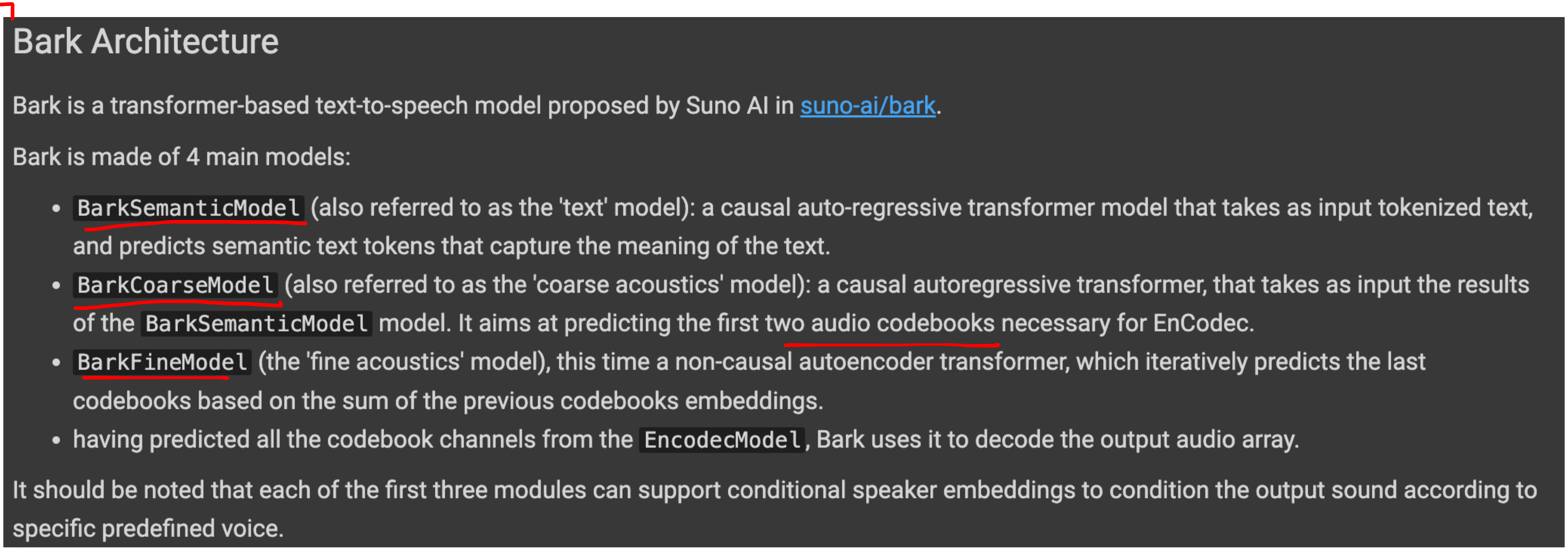
Support non-speak sound



Below is a list of some known non-speech sounds, but we are finding more every day. Please let us know if you find patterns that work particularly well on Discord!

- `[laughter]`
- `[laughs]`
- `[sighs]`
- `[music]`
- `[gasps]`
- `[clears throat]`
- `–` or `...` for hesitations
- `♪` for song lyrics
- CAPITALIZATION for emphasis of a word
- `[MAN]` and `[WOMAN]` to bias Bark toward male and female speakers, respectively

Architect



Bark Architecture

Bark is a transformer-based text-to-speech model proposed by Suno AI in suno-ai/bark.

Bark is made of 4 main models:

- `BarkSemanticModel` (also referred to as the 'text' model): a causal auto-regressive transformer model that takes as input tokenized text, and predicts semantic text tokens that capture the meaning of the text.
- `BarkCoarseModel` (also referred to as the 'coarse acoustics' model): a causal autoregressive transformer, that takes as input the results of the `BarkSemanticModel` model. It aims at predicting the first two audio codebooks necessary for EnCodec.
- `BarkFineModel` (the 'fine acoustics' model), this time a non-causal autoencoder transformer, which iteratively predicts the last codebooks based on the sum of the previous codebooks embeddings.
- having predicted all the codebook channels from the `EncodecModel`, Bark uses it to decode the output audio array.

It should be noted that each of the first three modules can support conditional speaker embeddings to condition the output sound according to specific predefined voice.

Hands-on on Colab