Overview of Zero-Shot Multi-speaker TTS Systems

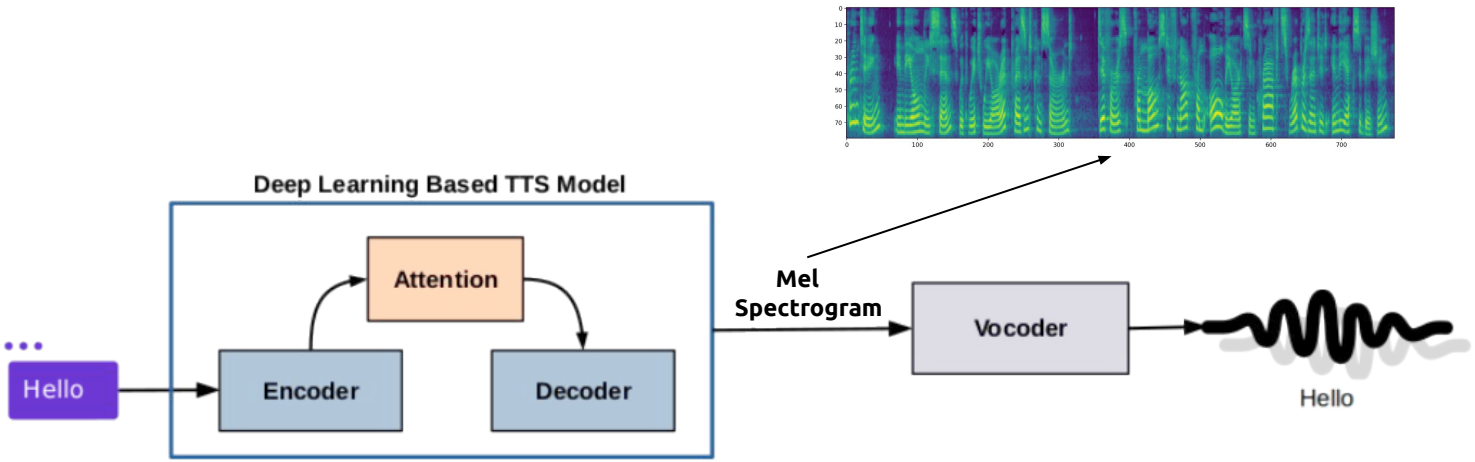
Edresson Casanova
edresson@coqui.ai
github.com/Edresson
linkedin.com/in/edresson

Agenda

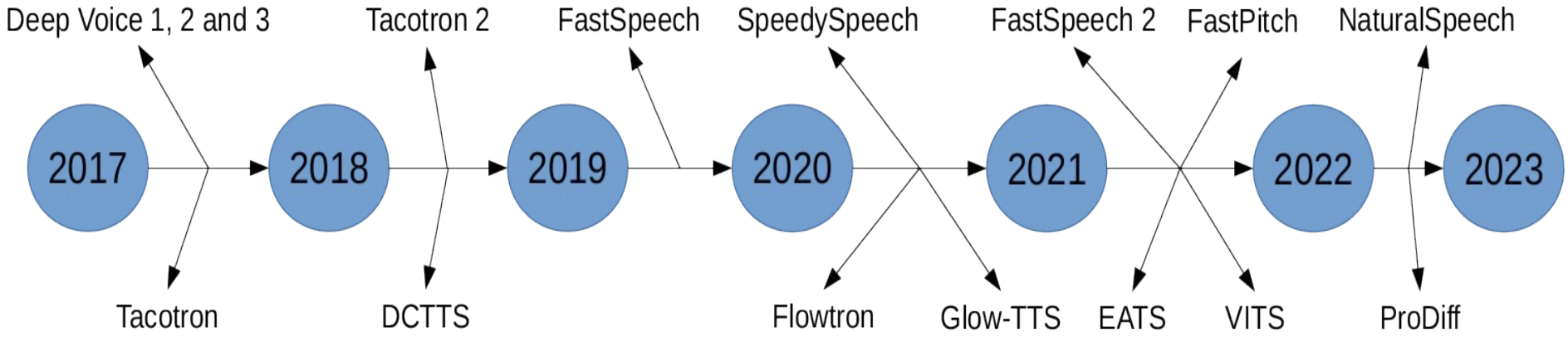
- Text to Speech (TTS)
- What is zero-shot multi-speaker TTS (ZS-TTS)?
- Multi-speaker TTS versus ZS-TTS
- How important is the Speaker Encoder for ZS-TTS?
- ZS-TTS main papers
- Main ZS-TTS open-source implementations
- ZS-TTS state-of-the-art audio samples

Text to Speech

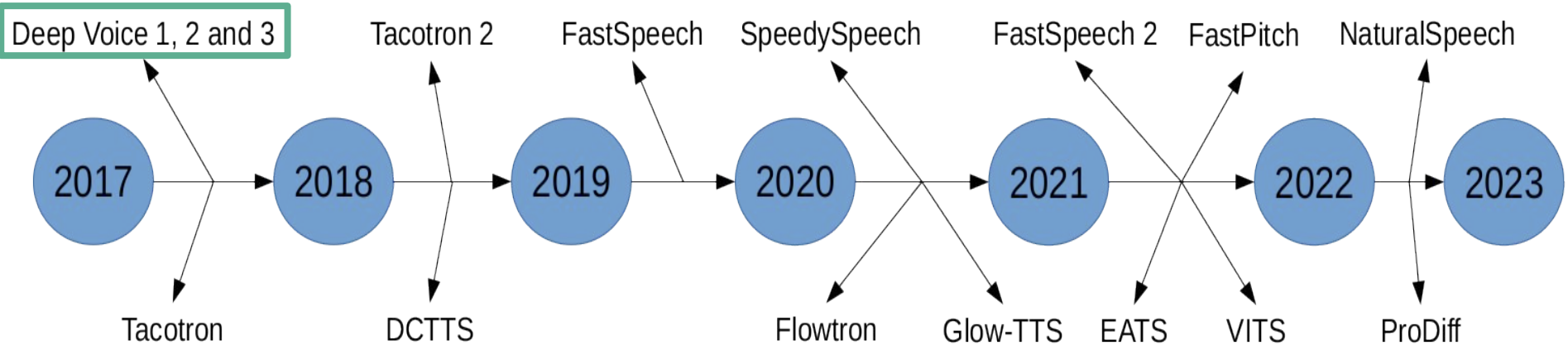
- From 2010s, neural network-based speech synthesis has gradually become the dominant method and achieved much better voice quality (Tan et al. 2021).



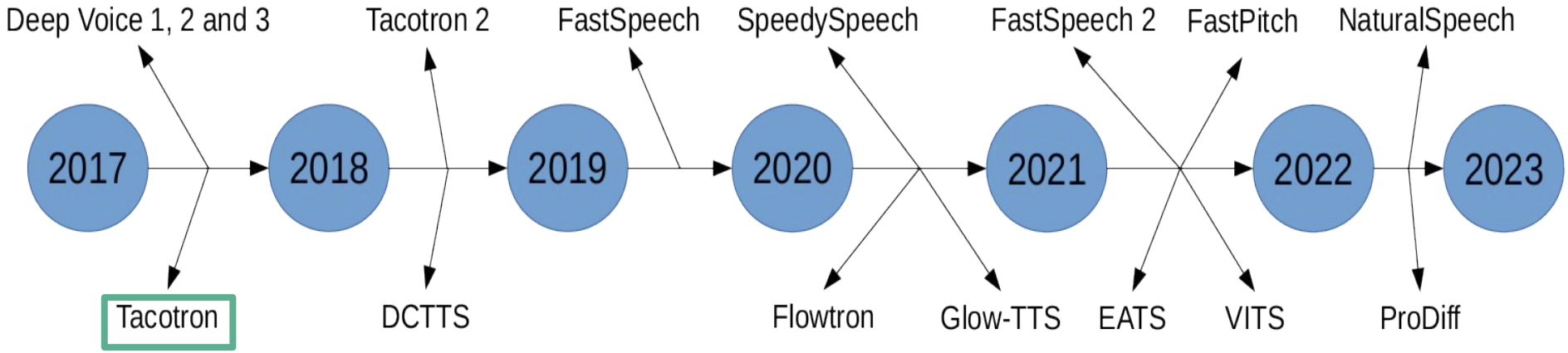
Text to Speech



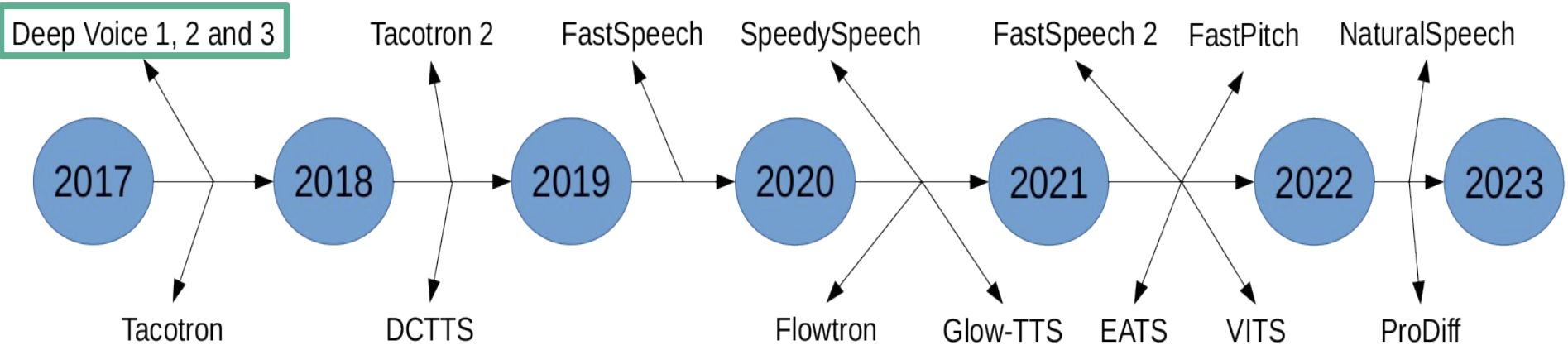
Text to Speech



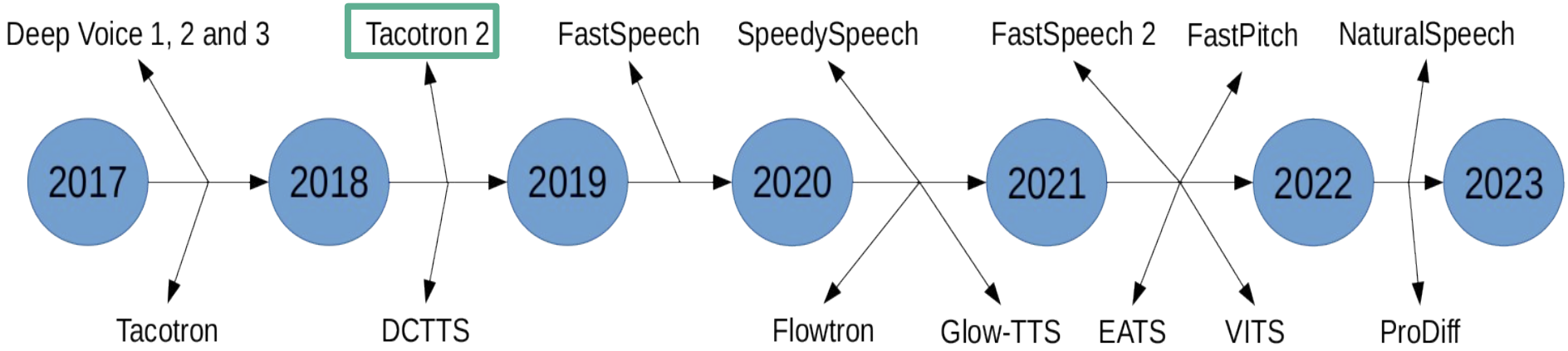
Text to Speech



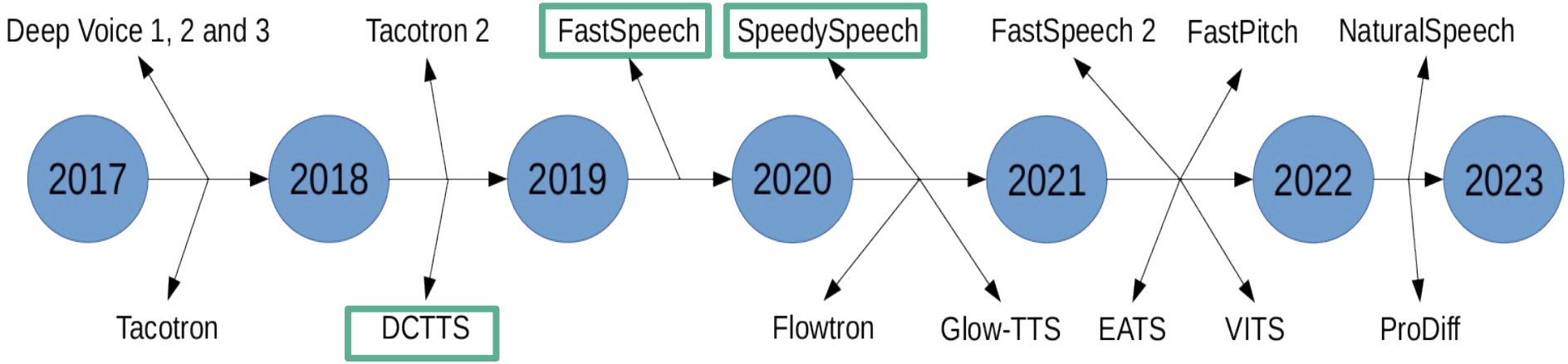
Text to Speech



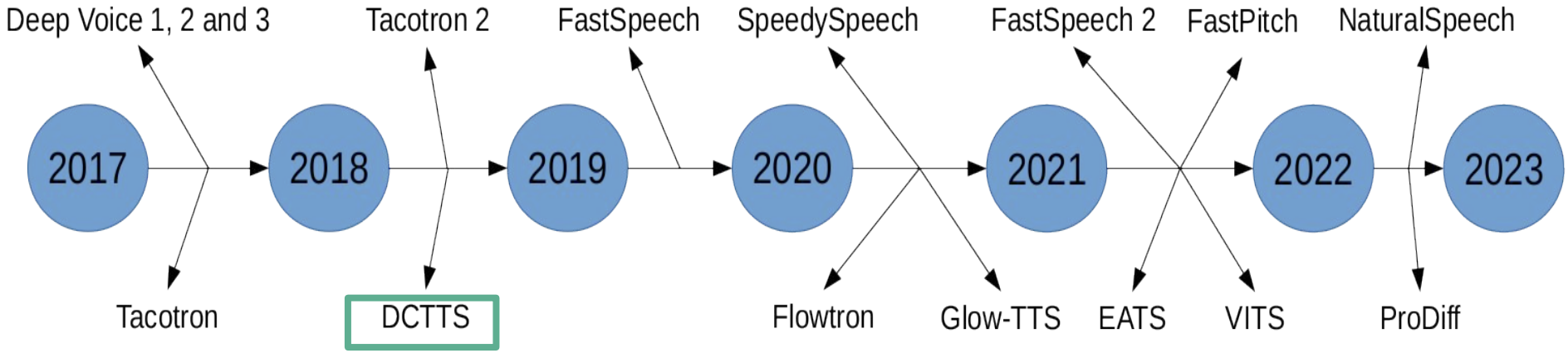
Text to Speech



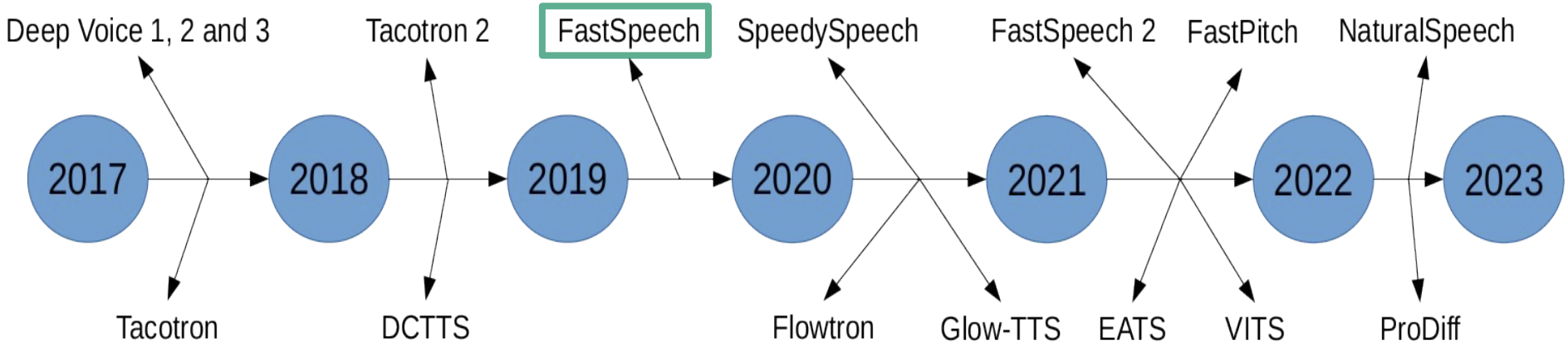
Text to Speech



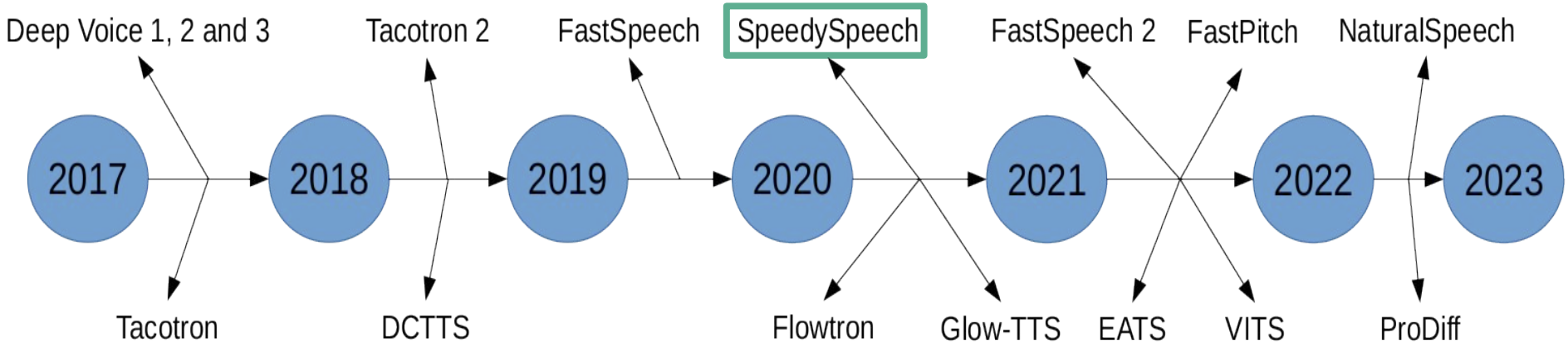
Text to Speech



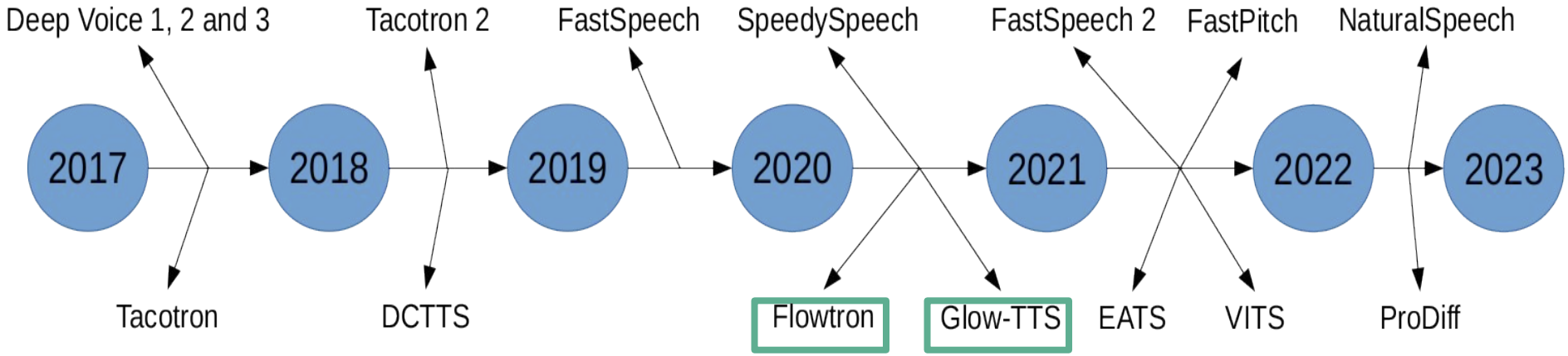
Text to Speech



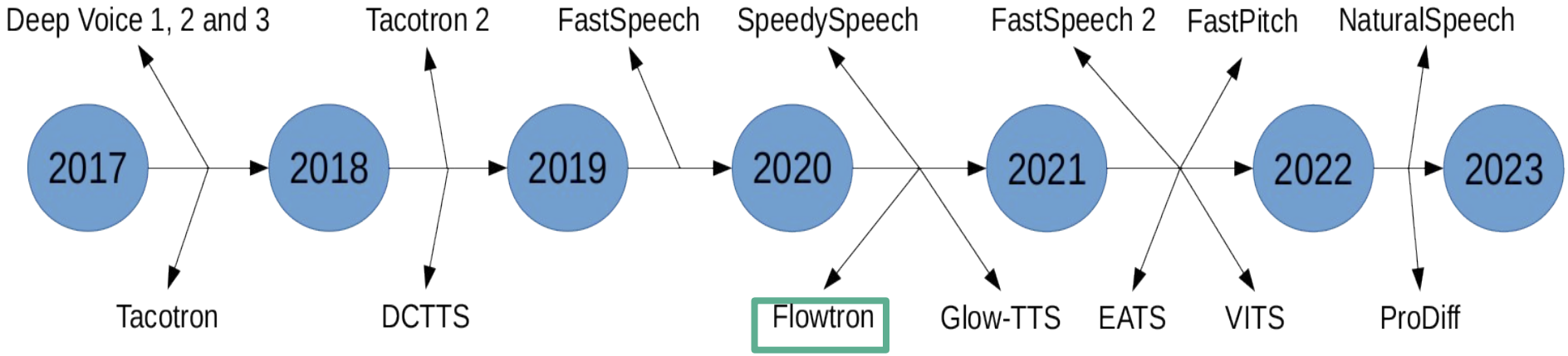
Text to Speech



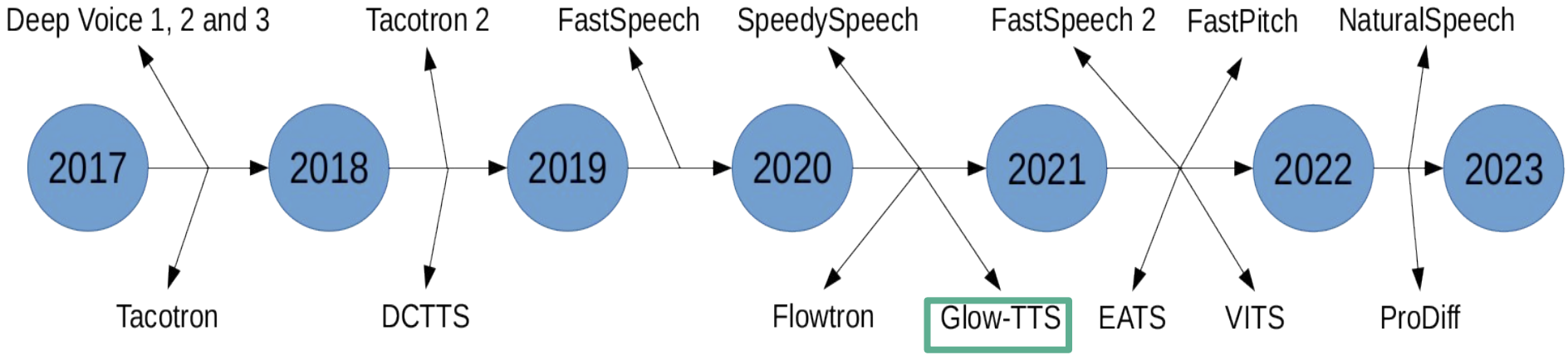
Text to Speech



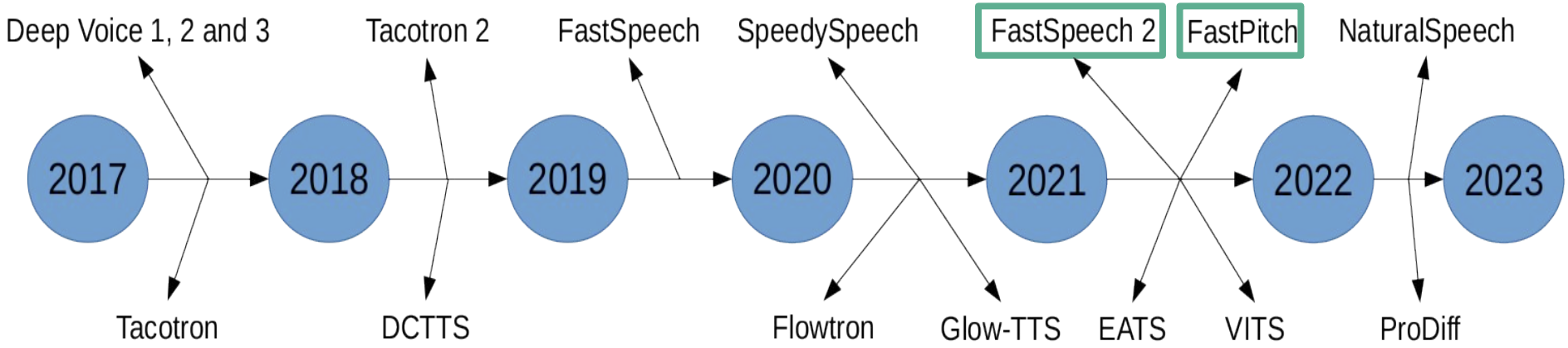
Text to Speech



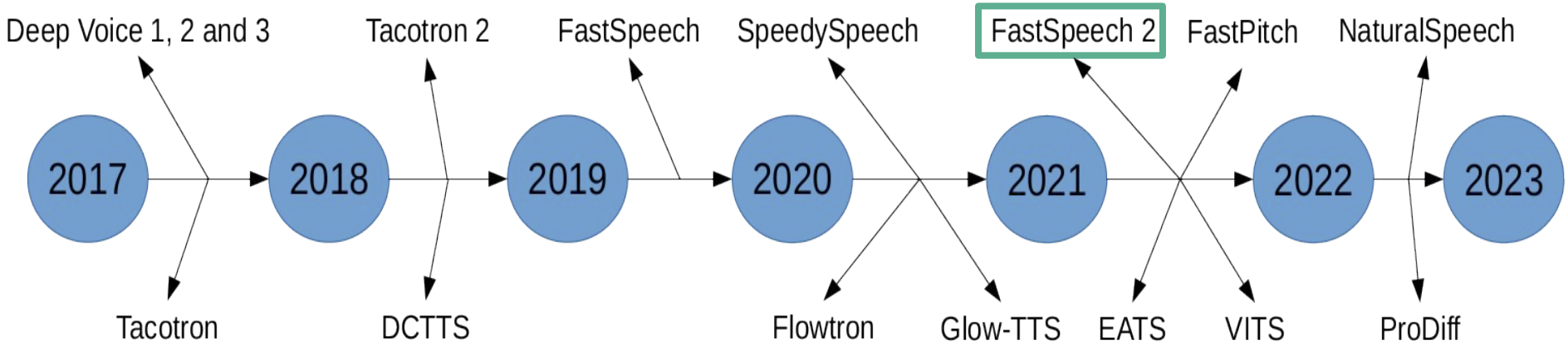
Text to Speech



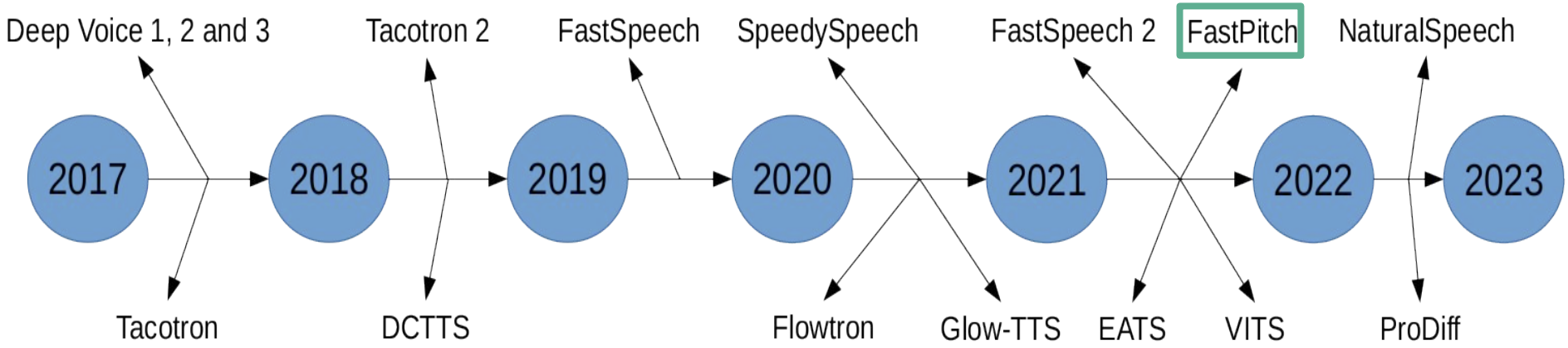
Text to Speech



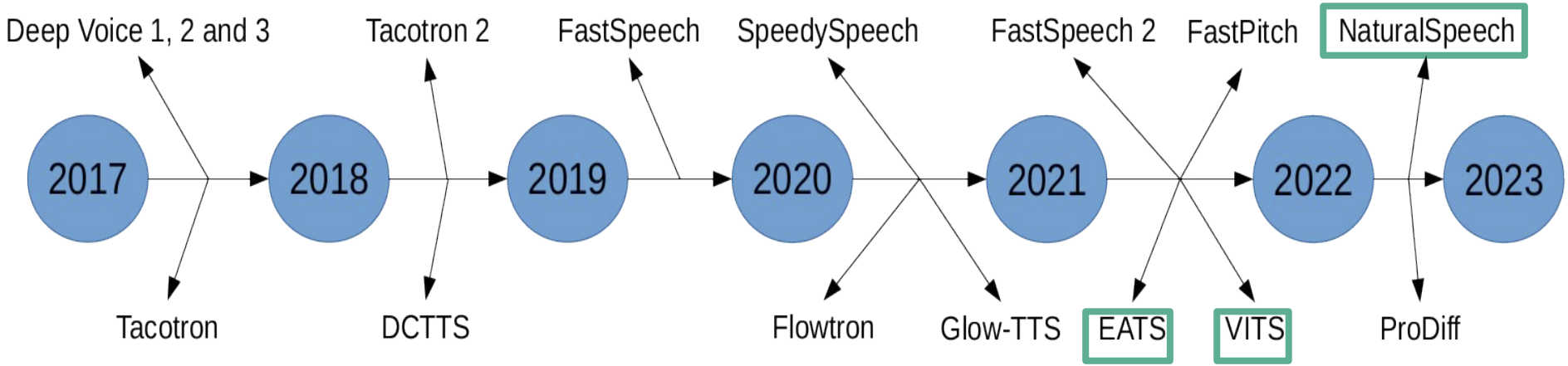
Text to Speech



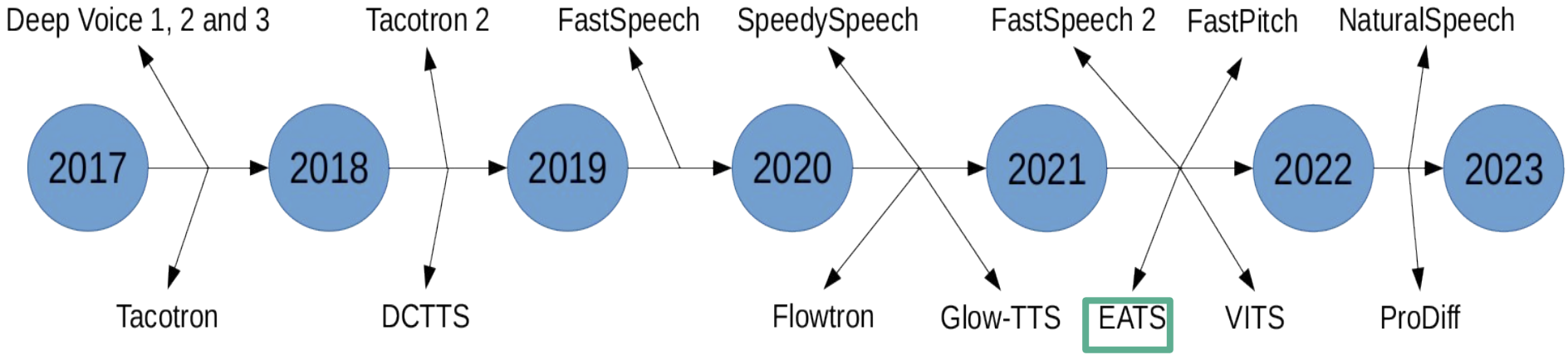
Text to Speech



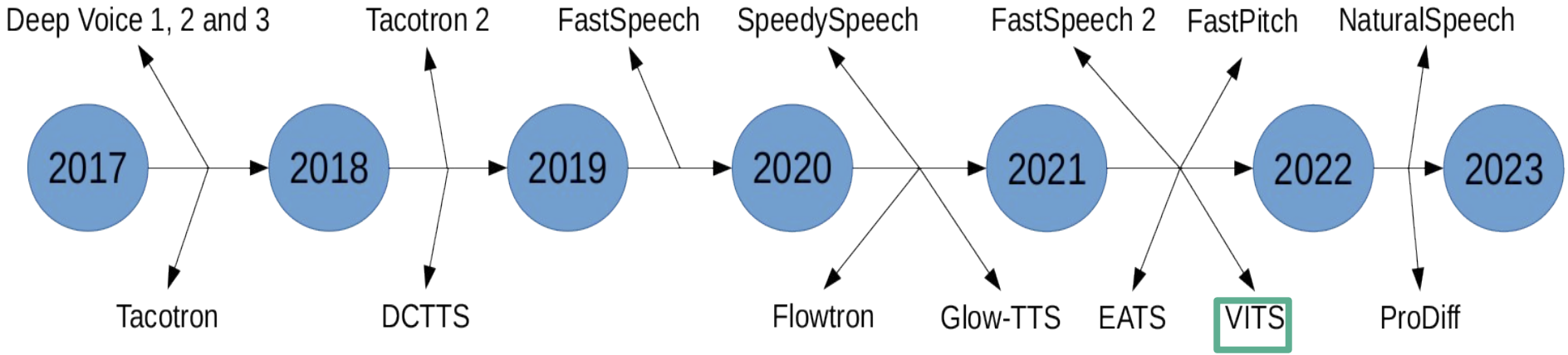
Text to Speech



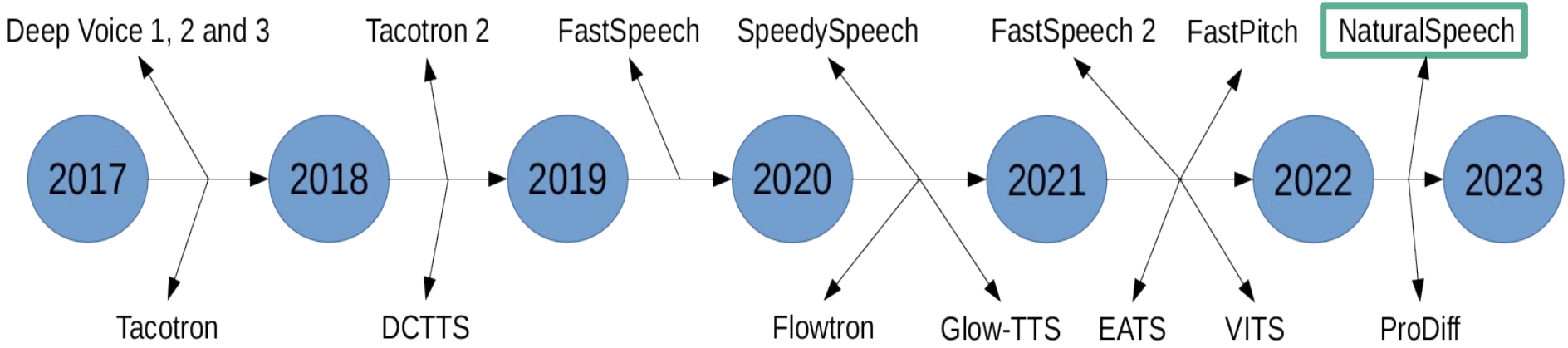
Text to Speech



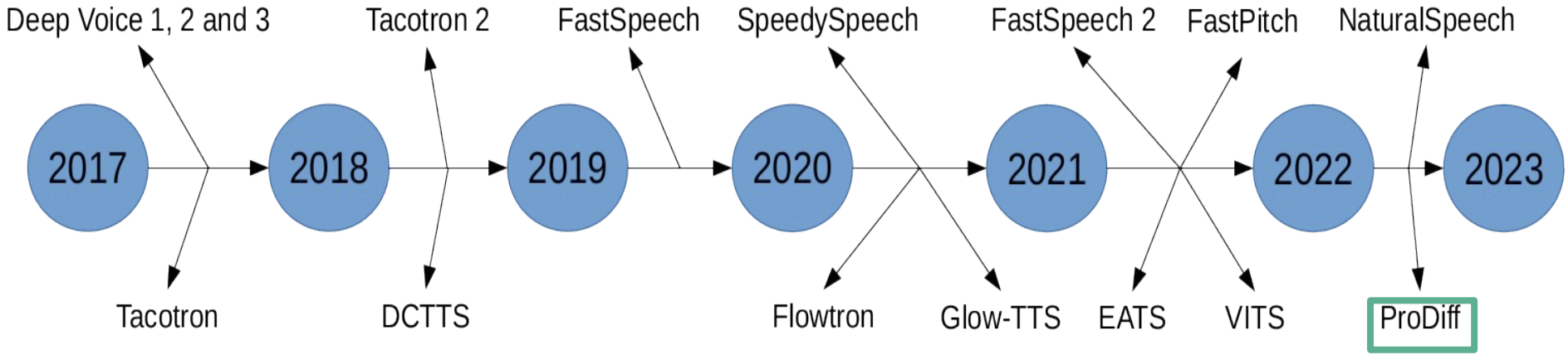
Text to Speech



Text to Speech

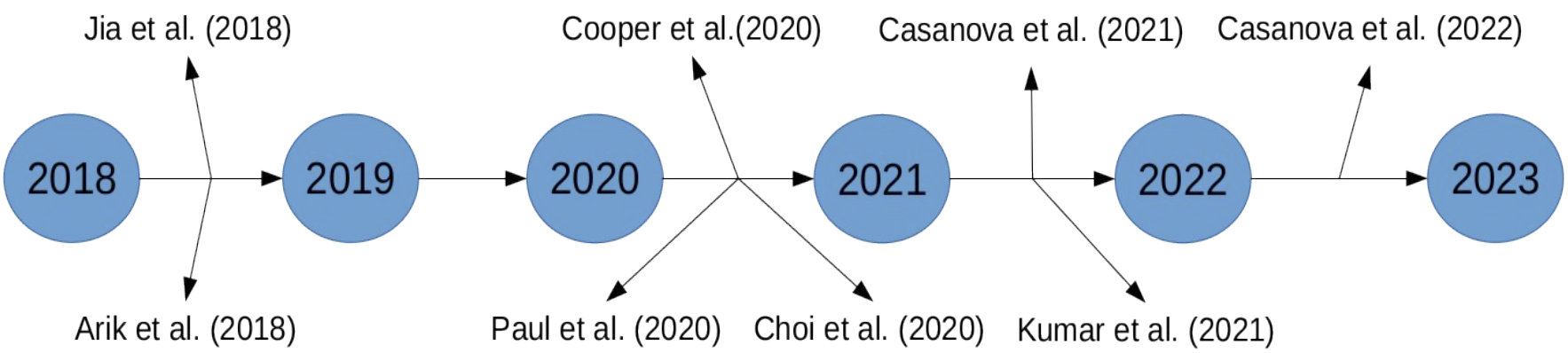


Text to Speech



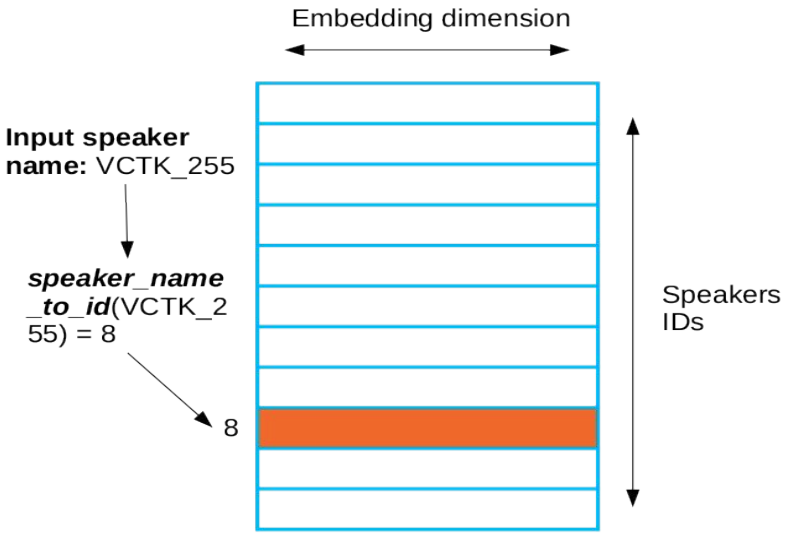
What is zero-shot multi-speaker TTS?

- In 2018, advances in speech synthesis motivated research that aimed to synthesize speech in the voice of a target speaker using just a few seconds of speech.



MULTI-SPEAKER TTS MODELS

- Uses a lookup table with learned fixed-size vectors to represent each speaker (speaker embedding).
- These fixed-size vectors are normally conditioned on the TTS model decoder.

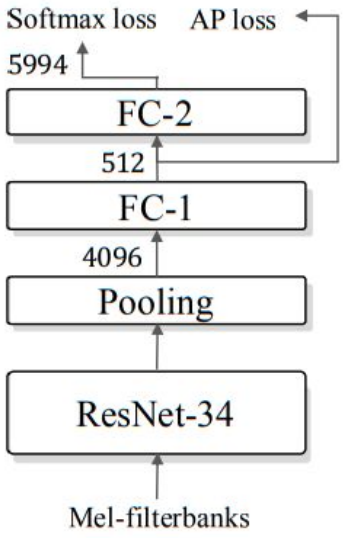


Zero-shot multi-speaker TTS

- Uses speaker embeddings extracted from an external speaker encoder/speaker verification system.
- Therefore, the speaker encoder is a very important module for the final quality of a zero-shot multi-speaker TTS model.

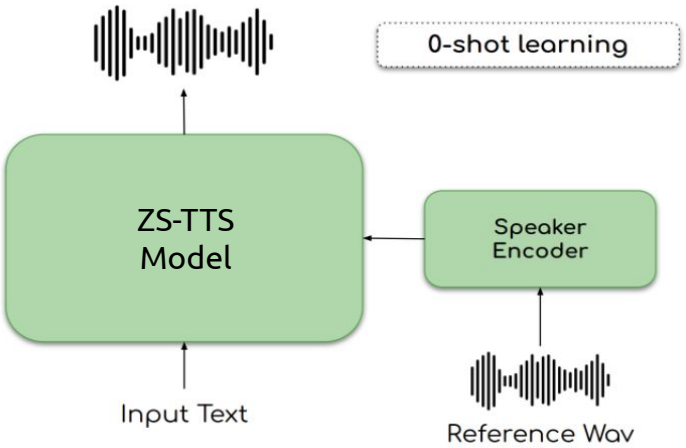
Speaker Encoder

- Losses:
  - Softmax
  - Angular Softmax
  - Angular Prototypical
  - Angular Margin Softmax
  - Generalized end2end loss
- Voxceleb Dataset:
  - 7k speakers
  - 145 different nationalities
- Produces good embeddings for new speakers

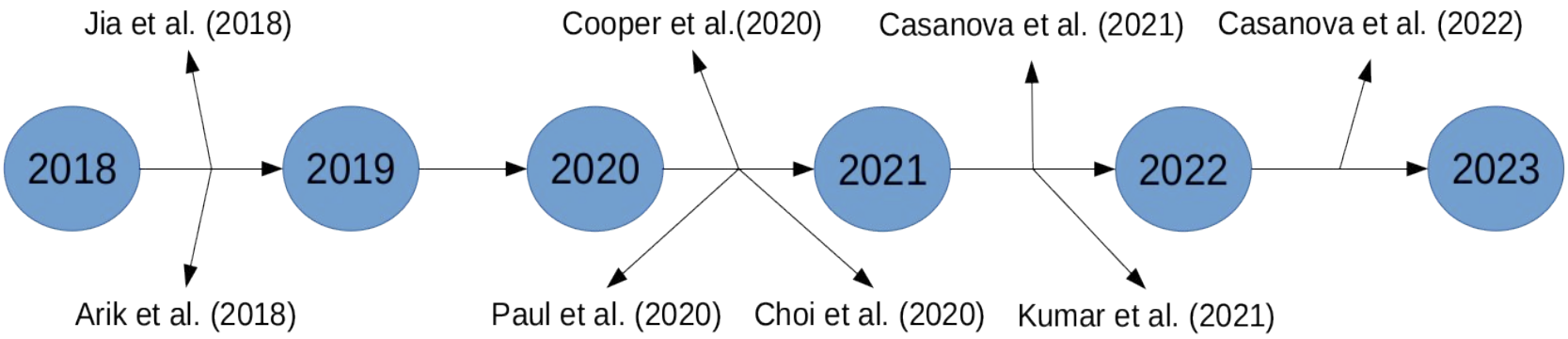


Zero-shot multi-speaker TTS

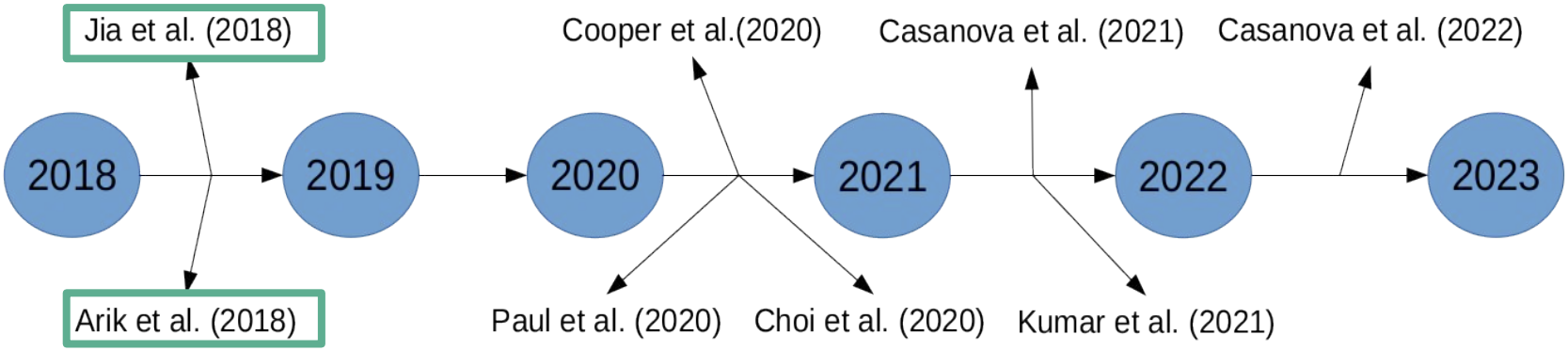
- On inference time it is able to clone a target speaker's voice not seen in training using just a few seconds of speech.
- Trained with thousand of speakers:
  - VCTK: 109 speakers
  - LibriTTS: 1151 speakers (train clean partitions)



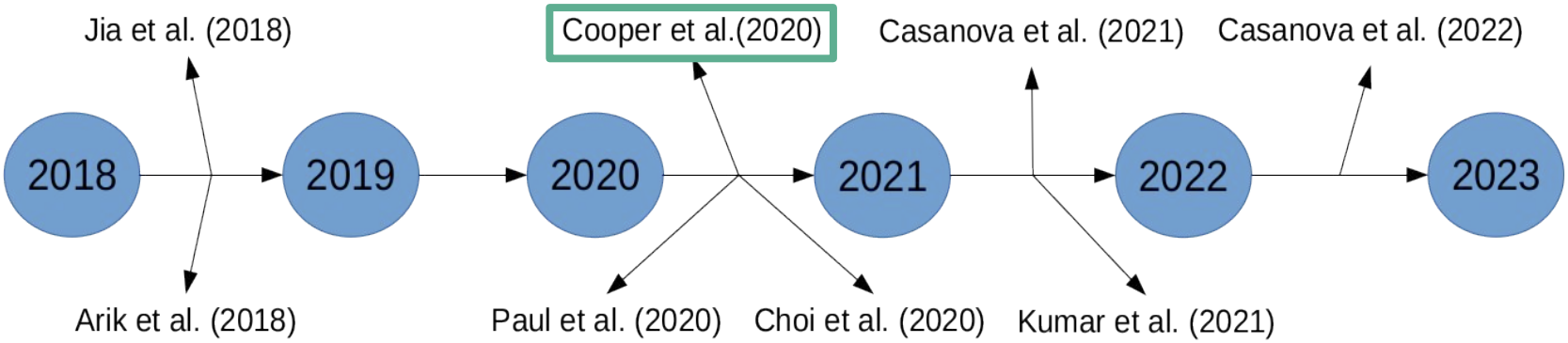
Zero-shot multi-speaker TTS



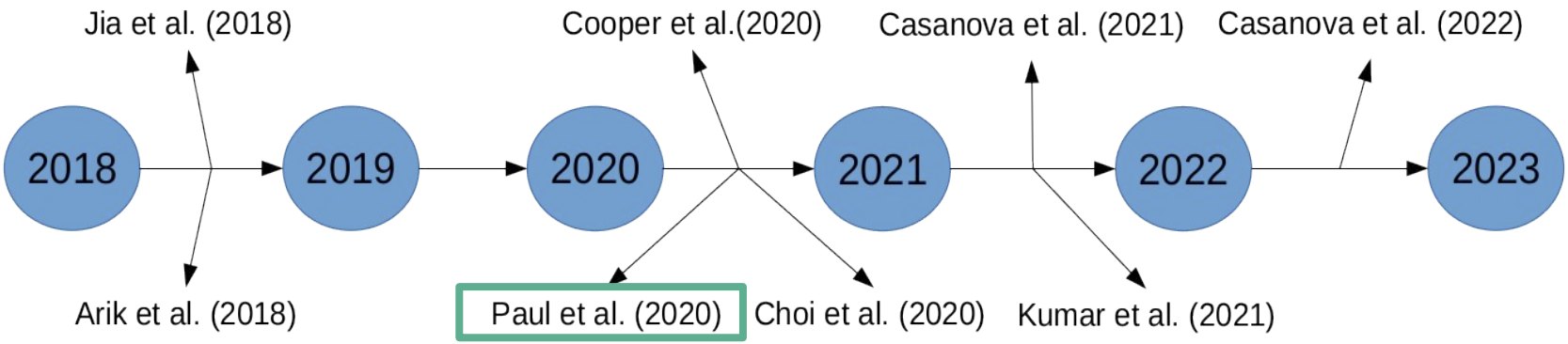
Zero-shot multi-speaker TTS



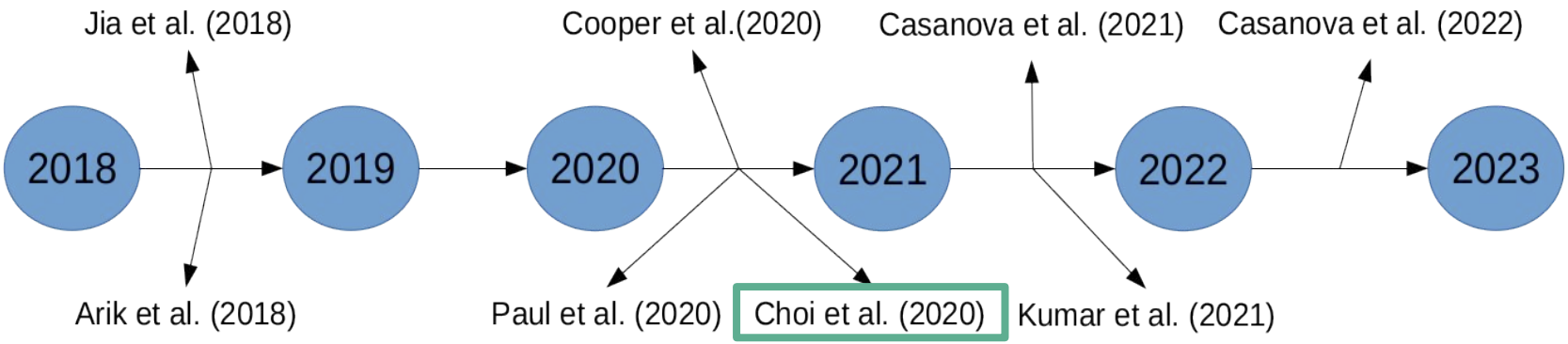
Zero-shot multi-speaker TTS



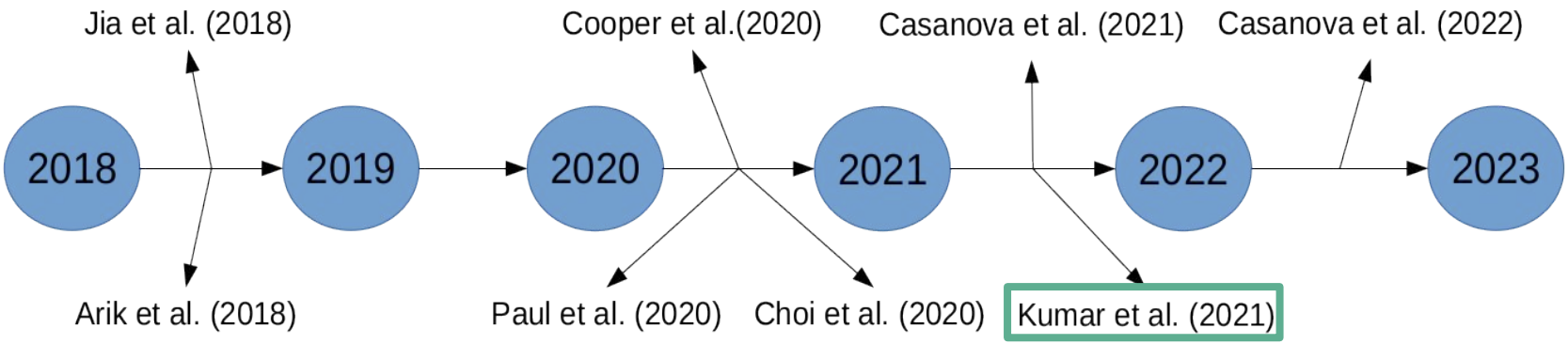
Zero-shot multi-speaker TTS



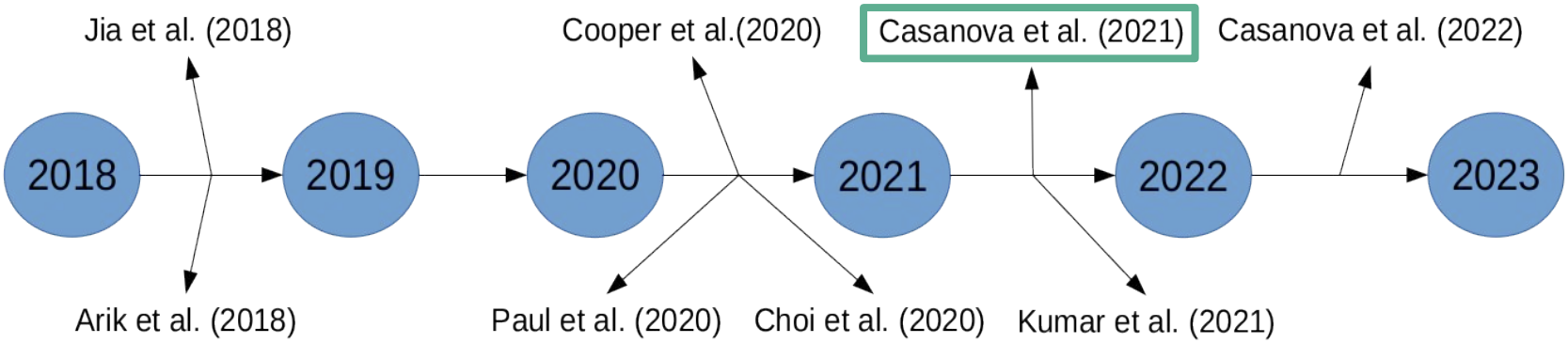
Zero-shot multi-speaker TTS



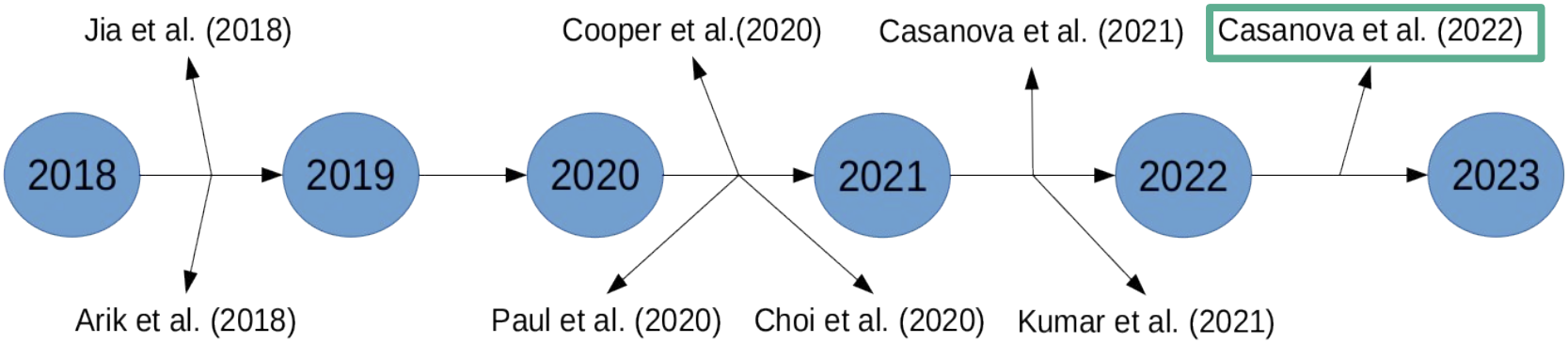
Zero-shot multi-speaker TTS



Zero-shot multi-speaker TTS



Zero-shot multi-speaker TTS



Open-source: Zero-shot multi-speaker TTS

- Cooper et al. 2020:
  - https://github.com/nii-yamagishilab/multi-speaker-tacotron
- Paul et al. 2020:
  - https://github.com/dipjyoti92/SC-WaveRNN
- Jia et al. 2018, SC-GlowTTS and YourTTS:
  - TTS: https://github.com/coqui-ai/TTS

multi-speaker-tacotron

VCTK multi-speaker tacotron for ICASSP 2020

Python 253 43

SC-WaveRNN

Official PyTorch implementation of Speaker Conditional WaveRNN

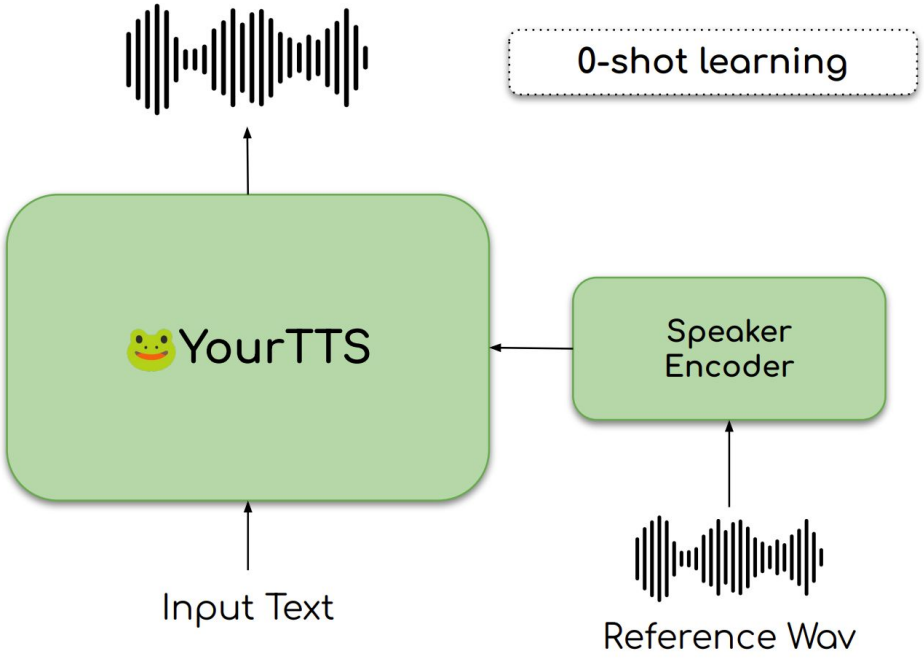
Python 104 18

TTS Public

- a deep learning toolkit for Text-to-Speech, battle-tested in research and production

Python 6.6k 692

YourTTS: Zero-shot and Multilingual TTS

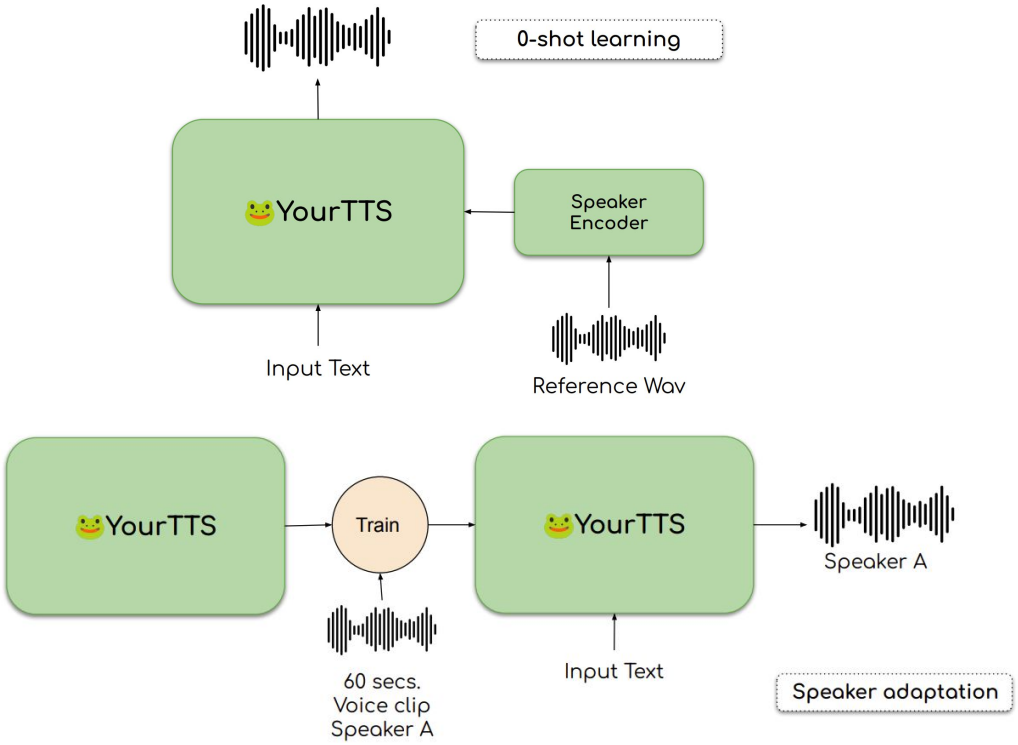


Julian reference

Julian zero-shot FR

Julian zero-shot EN

YourTTS: Zero-Shot and Speaker Adaptation - English

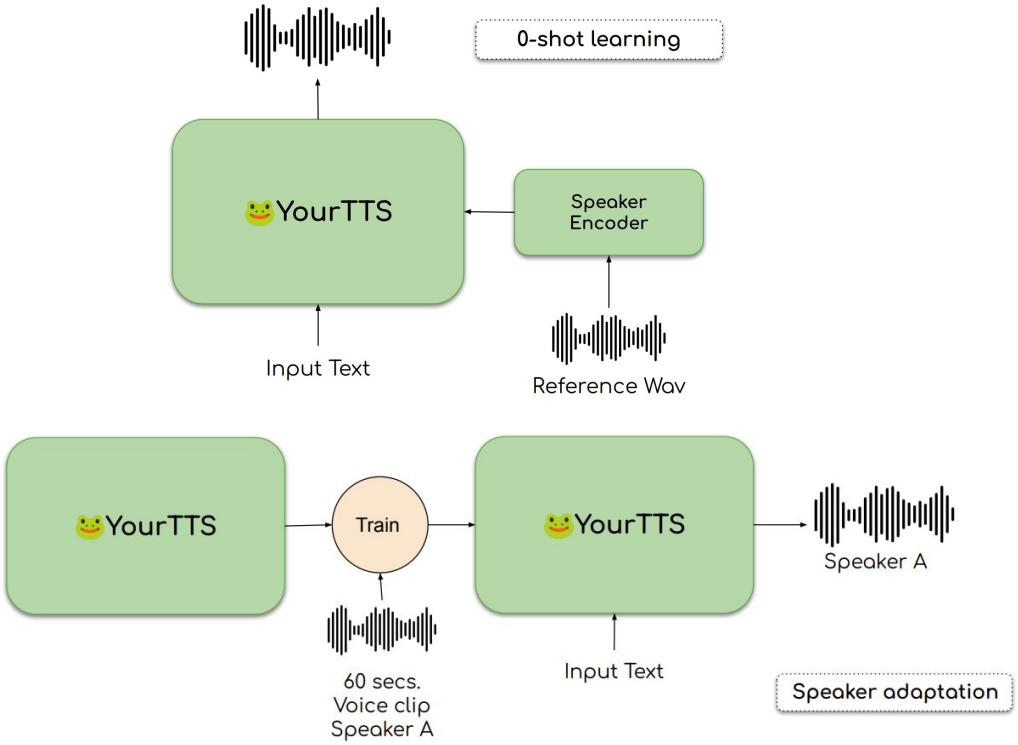


Chris reference

Chris Zero-Shot

Chris Fine-Tuned

YourTTS: Zero-Shot and Speaker Adaptation - Portuguese



Moacir reference

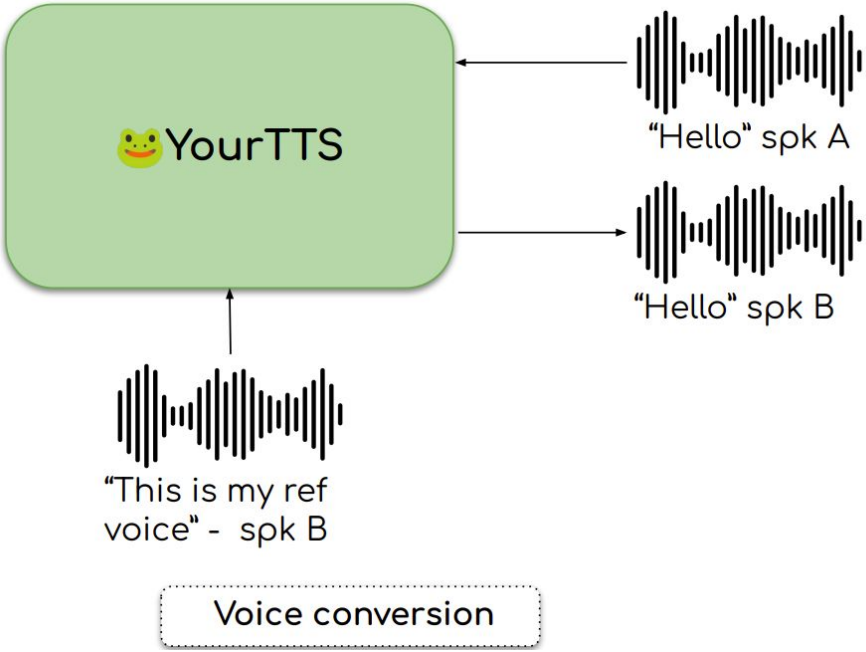
Moacir Zero-Shot

Moacir Fine-Tuned PT

Moacir Fine-Tuned EN

Moacir Fine-Tuned FR

YourTTS: Zero-shot Voice Conversion



p228 target speaker reference

p226 driving audio

p226 converted into p228

YourTTS: try it yourself

- TTS inference instructions:
  - https://github.com/Edresson/YourTTS#coqui-tts-released-model

- Demo of our latest YourTTS English only model:
  - https://coqui.ai/

References

- TAN, Xu et al. **A survey on neural speech synthesis.** arXiv preprint arXiv:2106.15561, 2021.
- ARIK, Sercan Ö. et al. **Deep voice: Real-time neural text-to-speech.** In: International Conference on Machine Learning. PMLR, 2017. p. 195-204.
- ARIK, Sercan et al. **Deep voice 2: Multi-speaker neural text-to-speech.** arXiv preprint arXiv:1705.08947, 2017.
- PING, Wei et al. **Deep voice 3: Scaling text-to-speech with convolutional sequence learning.** arXiv preprint arXiv:1710.07654, 2017.
- WANG, Yuxuan et al. **Tacotron: Towards end-to-end speech synthesis.** arXiv preprint arXiv:1703.10135, 2017.

References

- SHEN, Jonathan et al. **Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.** In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018. p. 4779-4783.
- TACHIBANA, Hideyuki; UENOYAMA, Katsuya; AIHARA, Shunsuke. **Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention.** In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. p. 4784-4788.
- REN, Yi et al. **Fastspeech: Fast, robust and controllable text to speech.** Advances in Neural Information Processing Systems, v. 32, 2019.
- VAINER, Jan; DUŠEK, Ondřej. **Speedyspeech: Efficient neural speech synthesis.** arXiv preprint arXiv:2008.03802, 2020.

References

- VALLE, Rafael et al. **Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis.** In: International Conference on Learning Representations. 2020.
- KIM, Jaehyeon et al. **Glow-tts: A generative flow for text-to-speech via monotonic alignment search.** Advances in Neural Information Processing Systems, v. 33, p. 8067-8077, 2020.
- REN, Yi et al. **Fastspeech 2: Fast and high-quality end-to-end text to speech.** arXiv preprint arXiv:2006.04558, 2020.
- ŁAŃCUCKI, Adrian. **Fastpitch: Parallel text-to-speech with pitch prediction.** In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021. p. 6588-6592.

References

- DONAHUE, Jeff et al. **End-to-end Adversarial Text-to-Speech.** In: International Conference on Learning Representations. 2020.
- KIM, Jaehyeon; KONG, Jungil; SON, Juhee. **Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech.** In: International Conference on Machine Learning. PMLR, 2021. p. 5530-5540.
- TAN, Xu et al. **NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality.** arXiv preprint arXiv:2205.04421, 2022.
- JIA, Ye et al. **Transfer learning from speaker verification to multispeaker text-to-speech synthesis.** Advances in neural information processing systems, v. 31, 2018.

few

References

- ARIK, Sercan et al. **Neural voice cloning with a few samples.** Advances in neural information processing systems, v. 31, 2018.
- COOPER, Erica et al. **Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings.** In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020. p. 6184-6188.
- PAUL, Dipjyoti; PANTAZIS, Yannis; STYLIANOU, Yannis. **Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions.** arXiv preprint arXiv:2008.05289, 2020.
- CHOI, Seungwoo et al. **Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding.** Proc. Interspeech 2020, p. 2007-2011, 2020.

References

- CASANOVA, Edresson et al. **SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model.** Proc. Interspeech 2021, p.3645-3649, 2021.
- CASANOVA, Edresson et al. **Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone.** In: International Conference on Machine Learning. PMLR, 2022. p. 2709-2720.
- HUANG, Rongjie et al. **Prodiff: Progressive fast diffusion model for high-quality text-to-speech.** arXiv preprint arXiv:2207.06389, 2022.

Overview of Zero-Shot Multi-speaker TTS Systems

Edresson Casanova
edresson@coqui.ai
github.com/Edresson
linkedin.com/in/edresson