**A. Generating Voice cloned from samples of the same voice and in different languages:**

Some AI solutions require only a couple of seconds of the sampling voice to start generating from it and some require a couple of hours to start generation. However, generating a voice in different languages that require speech translation and cross-language code-switching is still a work in progress in the speech generation domain.

Some research findings are as the following:

1. Tacotron and Tacotron 2 (<https://google.github.io/tacotron/publications/multilingual/> <https://github.com/NVIDIA/tacotron2>)
2. Play.ht: <https://play.ht/voice-cloning/>
3. <https://beebom.com/ai-mimics-voice-in-5-seconds/>
4. <https://www.resemble.ai/> (can also do translation between two different voices)
5. Speechify or Synthesia are multi-language TTS and there are similar products in market
6. VALL-E X <https://github.com/lifeiteng/vall-e> <https://plachtaa.github.io/>

This is used to synthesize high-quality personalized speech with around 3-second enrolled recording of an unseen speaker as an acoustic prompt.

1. IMS-Toucan (Multi-lingual and multi-speaker audios) <https://multilingualtoucan.github.io/> <https://github.com/DigitalPhonetics/IMS-Toucan>
2. Respeecher (<https://www.respeecher.com/product>) Speech-to-speech voice cloning product
3. Lovo (<https://lovo.ai/>) AI Voice Generator & Text to Speech platform
4. Overdub (<https://www.descript.com/overdub?lmref=hGOsGw>) lets you create a text-to-speech model of your voice or select one from a ultra-realistic stock voices selection.

List of solution to detect voice spoofing:

1. Dessa (<https://github.com/dessa-oss/fake-voice-detection>) Using temporal convolution to detect Audio Deepfakes
2. Voice Cloning Detection [[VCD](https://github.com/alep079/voice_cloning_detection)]
3. Cross-modal information fusion for voice spoofing detection [[Github](https://github.com/ibliever/Cross-modal-information-fusion-for-voice-spoofing-detection),[Paper](https://paperswithcode.com/paper/cross-modal-information-fusion-for-voice)]
4. Fake Audio Detector [[Github](https://github.com/UNICT-Fake-Audio/fake-audio-detector)]

Noteworthy, to benchmark the performance of tools in detecting spoofed and deepfake speech, there is a well-known database used in the community named [ASVspoof](https://arxiv.org/abs/2109.00537). It is being published as part of a continuous challenge to aid in the progress of automatic speaker verification.

General Notes about these models:

* Codec models are able to encode waveforms into discrete acoustic codes and reconstruct high-quality waveforms even if the speaker is unseen in training.
* Transfer Learning can be also used like from Speaker Verification to Multispeaker Text-To-Speech Synthesis (trained on a speaker verification task using an independent dataset of noisy speech without transcripts from thousands of speakers)

**B. Generating Music Songs of specific styles (melodies and rhythms) and translation between languages:**

1. Simple and Controllable Music Generation by Meta as MusicGen (<https://ai.honu.io/papers/musicgen/>) conditioned on textual description/prompt or melodic features, allowing better controls over the generated output.
2. Google MusicLM (<https://google-research.github.io/seanet/musiclm/examples/>)
3. RVC [https://github.com/RVC-Project](https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/blob/main/docs/README.en.md) (Voice data <= 10 mins can also be used to train a good VC model), tutorial guide: <https://medium.com/@futuretechie.ai/use-ai-to-clone-any-voices-sing-any-songs-for-free-a0ad2f0f7432>
4. Text-Prompted Generative Audio Model (<https://github.com/suno-ai/bark>)
5. Diffusion models for generating high quality music audio from text prompts <https://google-research.github.io/noise2music/>

* No to very limited research on generating songs with voice translated between different languages, however with MusicLM or similar models this can be achieved through direct prompting (e.g. adding up the selected voice in specific language with the requested melody of the generated music in the description/promot). VALL-E X or similar models alongside LLMs can be also utilized to generate music with voice in different languages through employing multiple internal codec and cloning models.