Topic 23: Speech & Audio Models

1. Meta AI's Wav2Vec 2.0: a self-supervised framework that learns rich speech representations from large amounts of unlabeled audio for ASR: https://research.facebook.com/publications/wav2vec-2-0-a-framework-for-self-supervised-learning-of-speech-representations/
2. OpenAI's Whisper: A high-performance multilingual ASR system: https://openai.com/index/whisper/
3. Meta AI's Massively Multilingual Speech: Supports ASR and TTS for 1,100+ languages and language identification for 4,000+ languages: https://ai.meta.com/blog/multilingual-model-speech-recognition/
4. Meta AI's SeamlessM4T: A unified multimodal model for speech recognition, speech-to-text translation, speech-to-speech translation, and text translation across ~100 languages: https://ai.meta.com/research/publications/seamlessm4t-massively-multilingual-multimodal-machine-translation/
5. Google's AudioLM: A framework that treats audio generation like language modeling, generating coherent speech (and music) without transcripts: https://research.google/blog/audiolm-a-language-modeling-approach-to-audio-generation/
6. Microsoft's VALL-E: A zero-shot TTS model that clones a voice from a 3-second sample and generates speech in that voice: https://arxiv.org/abs/2301.02111
7. Suno AI Bark: An open-source generative text-to-audio model that produces expressive and natural speech (and even music) from text prompts: https://github.com/suno-ai/bark
8. Meta AI's Voicebox: A flexible generative speech model supporting zero-shot TTS, noise removal, editing, and style transfer: https://ai.meta.com/blog/voicebox-generative-ai-model-speech/
9. Deepgram Blog on the Evolution of ASR: https://deepgram.com/learn/evolution-of-asr
10. Deepgram's Open-Source Speech Model Benchmark: A dive into three generations of open ASR models: https://deepgram.com/learn/best-speech-to-text-apis
11. What is Speaker Diarization: Beginner-friendly guide to diarization: https://www.assemblyai.com/blog/what-is-speaker-diarization-and-how-does-it-work
12. Pyannote: An open-source toolkit for diarization: https://pyannote.github.io/
13. A short video overview of OpenAI's Whisper model: https://www.youtube.com/watch?v=nE5iVtwKerA
14. NVIDIA NeMo T5-TTS: a text-to-speech approach using large language models: https://github.com/NVIDIA-NeMo/NeMo
15. Meta's AudioCraft: An open-source toolkit for generative audio: https://ai.meta.com/resources/models-and-libraries/audiocraft/
16. OWSM v3.1: A recent open-source "Whisper-style" speech model trained on public data: https://arxiv.org/abs/2401.16658