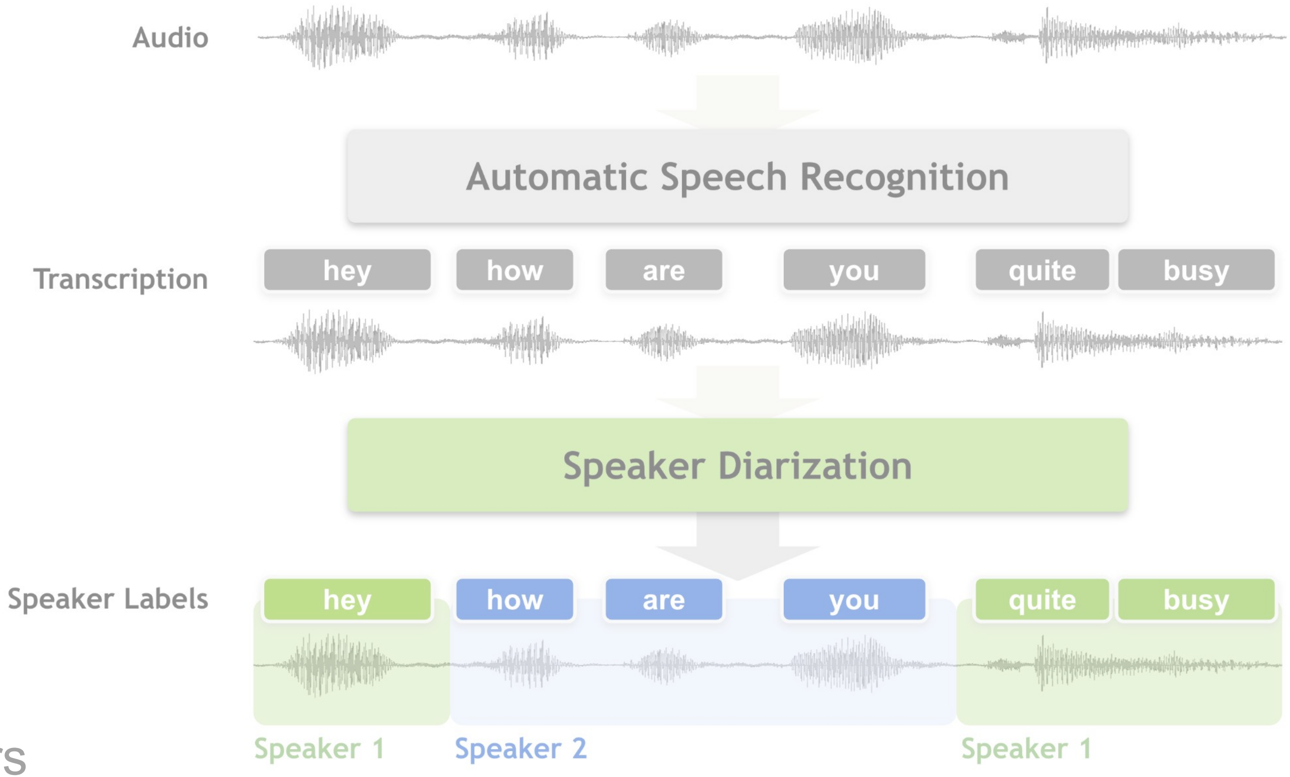
Speech-to-text



Melvin Wevers
ASH Digital History Group

Why use text2speech as historians?

- **Efficient Transcription**: Quickly converts spoken interviews, lectures, and seminars into textual format, saving time and effort.
- **Archive Accessibility**: Makes vast oral history archives searchable, aiding research and data retrieval.
- **Multilingual Support**: Facilitates transcription and analysis of sources in multiple languages and dialects.
- **Data Preservation**: Helps in digitizing and preserving deteriorating analog recordings for future generations.
- **Contextual Analysis**: With refined transcripts, historians can employ textual analysis tools to discern patterns, themes, or sentiments.

- **Two tasks**:
  - **Diarization**: distinguishing and separating different speakers in an audio recording, “who spoke when”
  - **speech-to-text conversion**: converting spoken language into written text

Diarization using PyAnnote

- **Voice Activity Detection (VAD)**: Filters out non-speech segments.
- **Embedding Extraction**:
  - Divides speech into overlapping chunks.
  - Extracts speaker-specific neural embeddings for each chunk.
- **Clustering**:
  - Uses embeddings to group speech chunks by speaker.
  - Employs metrics like cosine similarity for clustering.
- **Scoring & Decision-making**: Predicts the optimal number of speaker clusters.
- **Re-segmentation & Overlap Detection**: Refines speaker boundaries and detects overlapping speech.
- **Training & Fine-tuning**:
  - Offers pre-trained models.
  - Supports custom training on domain-specific data.

- Bredin, Hervé. 2023. "Pyannote. Audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe." In *Proc. Interspeech*. Vol. 2023. https://catedrartve.unizar.es/reto2022/PYA_report.pdf.
- https://pyannote.github.io/ (also available as CLI)

Speech-to-text using Whisper

- **Whisper is an automatic speech recognition (ASR) system developed by OpenAI**
- **Deep Neural Networks**: Pattern recognition in audio data.
- **Feature Extraction**: Uses Mel-frequency cepstral coefficients (MFCCs) to represent audio for analysis.
- **Sequence-to-Sequence**: Employs Transformers for handling audio sequences.
- **Attention Mechanisms**: Focuses on specific audio parts for accurate word prediction.
- **Language Models**: Refines transcripts for grammatical accuracy.

- Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. “Robust Speech Recognition via Large-Scale Weak Supervision.” arXiv. https://doi.org/10.48550/arXiv.2212.04356.
- https://github.com/openai/whisper (also available as CLI)