AVATecH audio detectors

Fraunhofer IAIS

AVATecH Workshop 21.04.2010 Nijmegen, Netherlands

Daniel Schneider, Sebastian Tschöpel

Fraunhofer
IAIS

Contents

- Introduction
- AVATecH Corpus
- Annotation scenarios
- Semi-automatic workflows for the annotation scenarios
- AVATecH audio detectors: State of the art
- Outlook
- Demonstration

Fraunhofer IAIS Speech Group

- Working on Spoken Document Retrieval since 2001
  - ASR, speech search, structual audio analysis

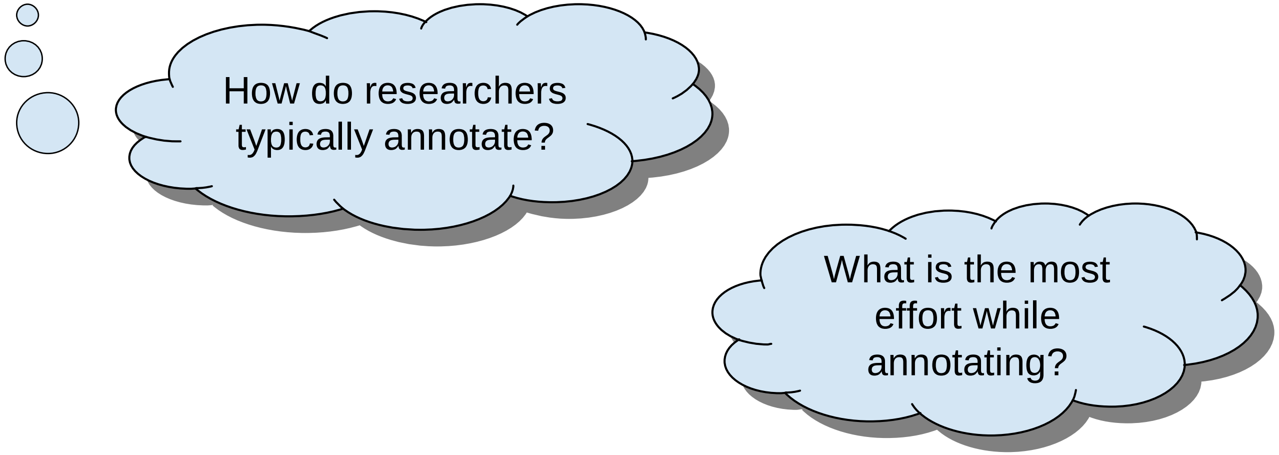
- Involved in public research projects and industry cooperations

- So far: mainly work on Broadcast data
  - Focus on language-dependent solutions for German

- But also specialized work
  - ASR on motorcycles, Animal sound discovery

Introduction

- What has been done?
  - Reviewed AVATecH corpora provided by MPI Nijmegen
  - Derived examples for *annotation scenarios*
  - Improved analysis *algorithms on difficult AVATecH data*
  - Developed concepts for detectors that *exploit user-feedback*
- Open problem: Definition of more annotation scenarios
  - how can we support your daily work?



How do researchers typically annotate?

What is the most effort while annotating?

AVATecH Corpus

- Material in various MPI corpora is
  - varying in audio quality (office experiments vs. field recordings)

    disqualifies fixed analysis models
  - varying in language

    disqualifies language-dependent approaches
  - varying in genre (interviews, monologues, ...)

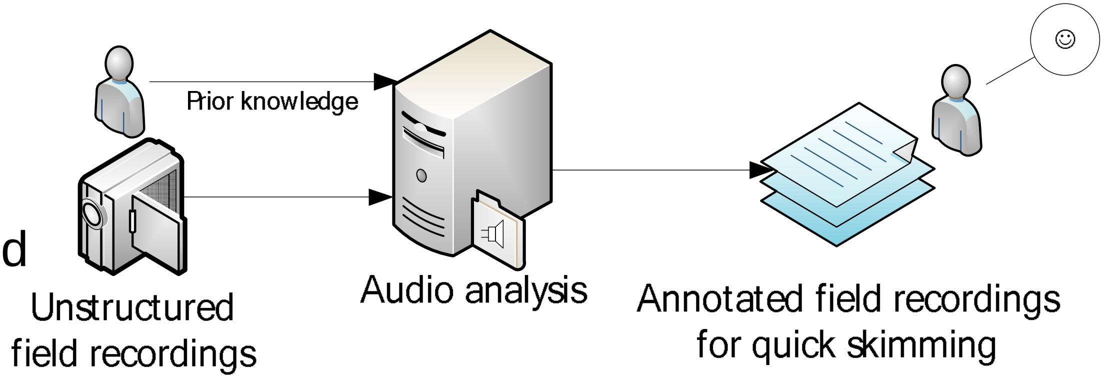
    disqualifies specialized solutions
  - not necessarily carrying information of interest in audio

- Flexible solutions needed that are able to cope
  with a large variety of annotation problems

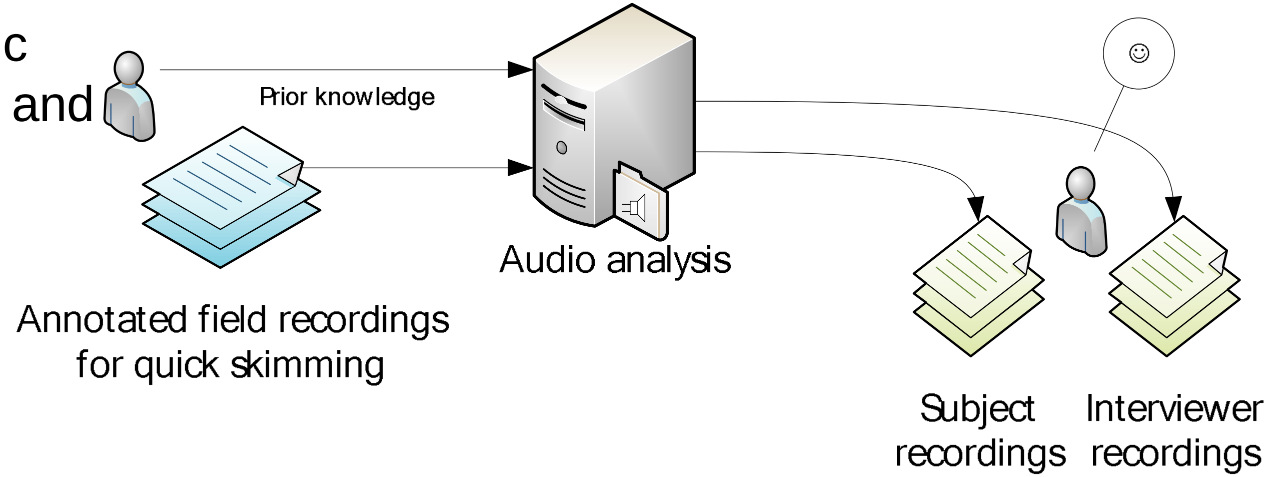
- Initially we focus on two general annotation scenarios

Initial annotation scenarios

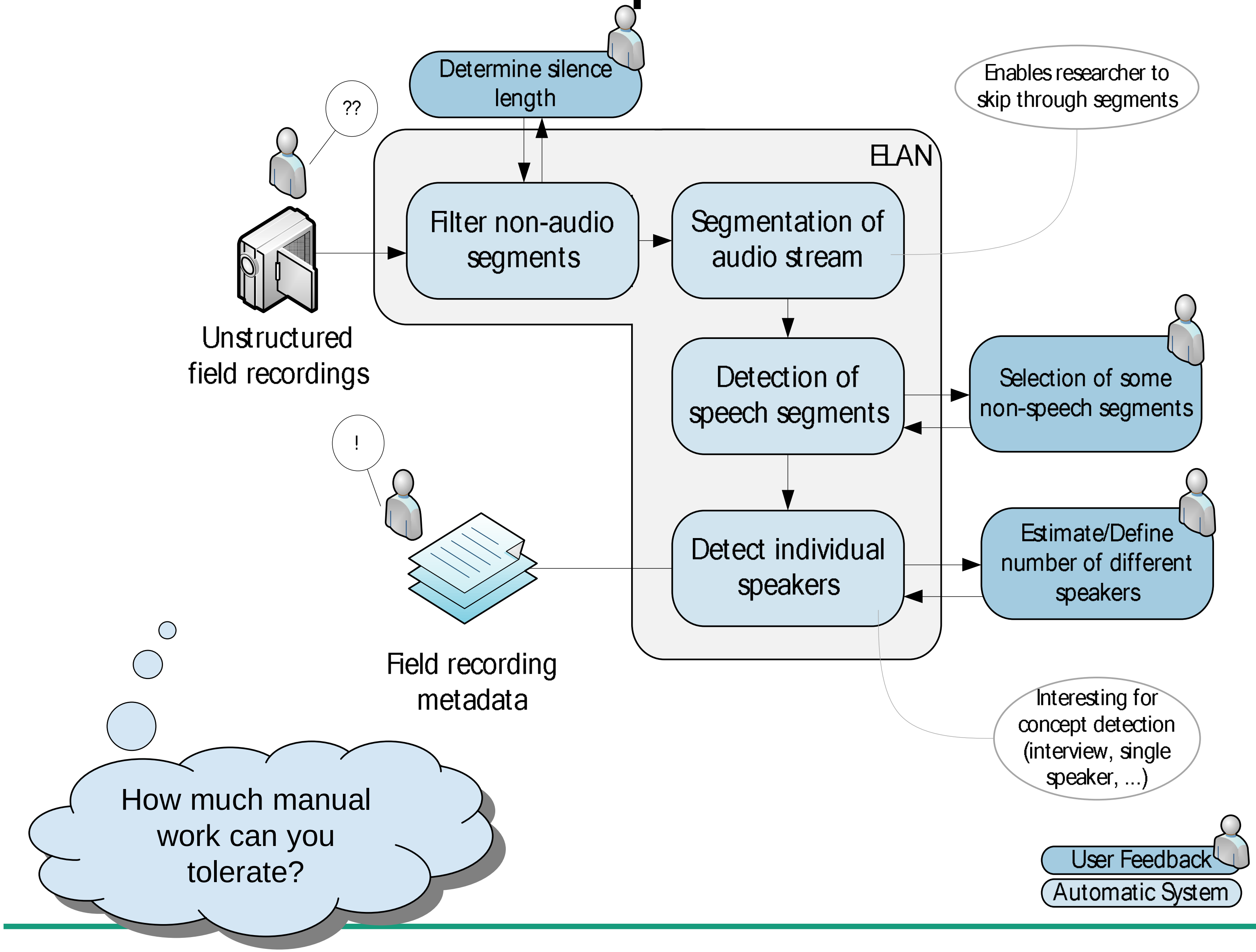
Scenario 1: Semi-automatic segmentation and labeling to support skimming of field recordings


Prior knowledge
Unstructured field recordings
Audio analysis
Annotated field recordings for quick skimming
☺

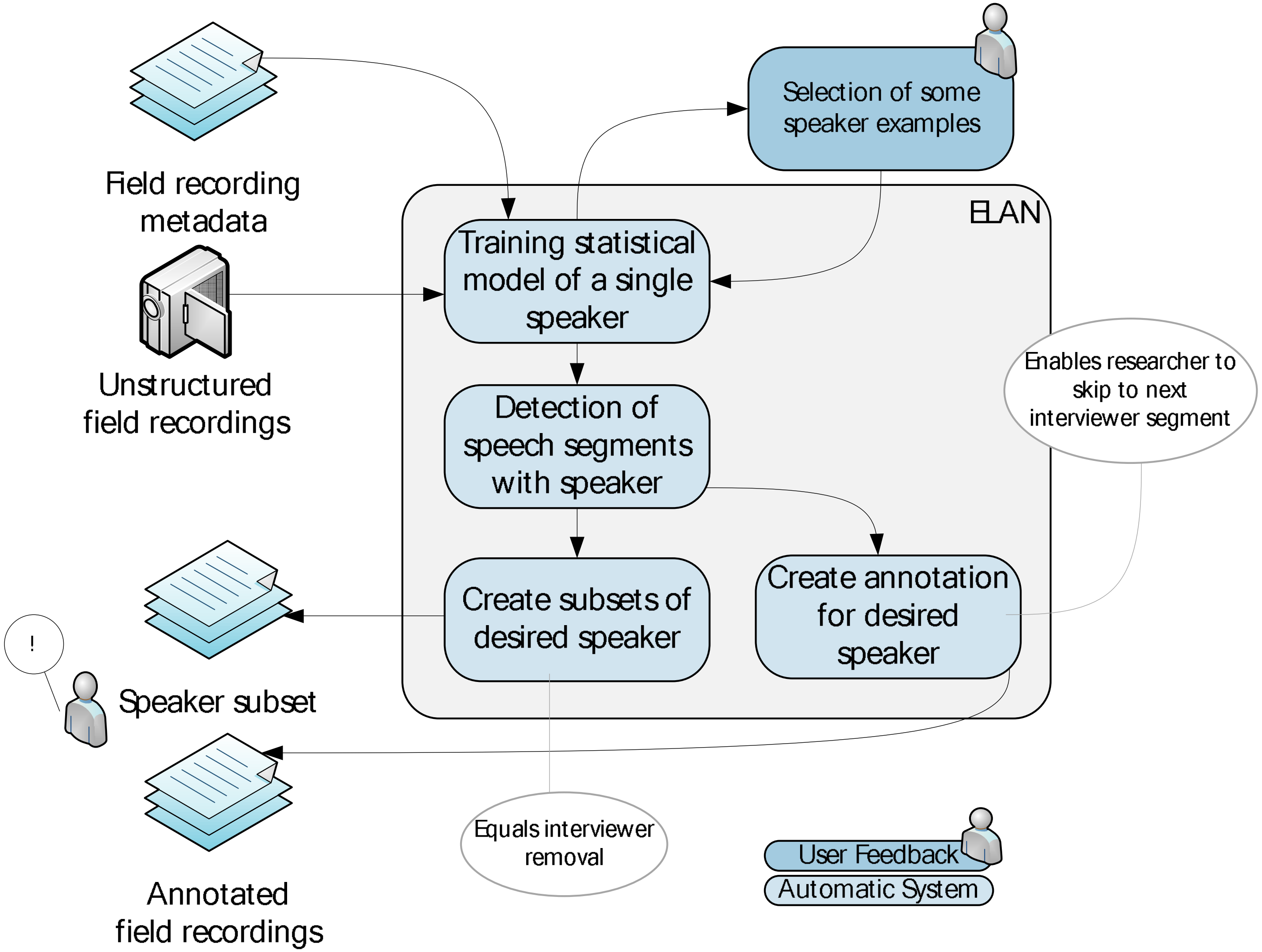
Scenario 2: Semi-automatic labeling of interviewers and subjects



Scenario 1: Workflow for pre-annotation of field recordings

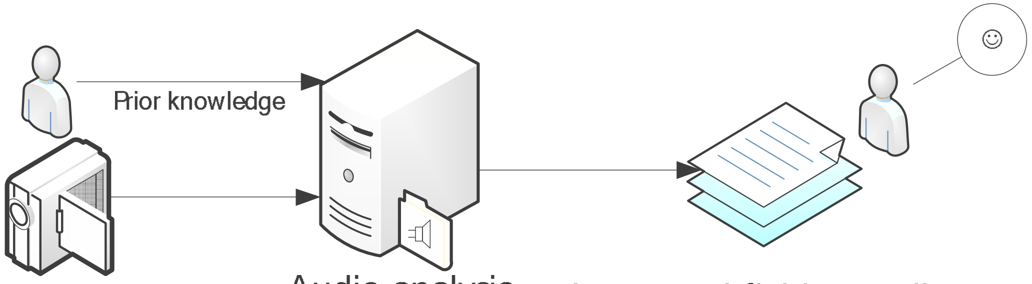

Determine silence length
??
ELAN
Enables researcher to skip through segments
Filter non-audio segments
Segmentation of audio stream
Unstructured field recordings
Detection of speech segments
Selection of some non-speech segments
!
Detect individual speakers
Estimate/Define number of different speakers
Field recording metadata
Interesting for concept detection (interview, single speaker, ...)
How much manual work can you tolerate?
User Feedback
Automatic System

Scenario 2: Workflow for interview structuring


Field recording metadata
Unstructured field recordings
Selection of some speaker examples
ELAN
Training statistical model of a single speaker
Detection of speech segments with speaker
Create subsets of desired speaker
Create annotation for desired speaker
Enables researcher to skip to next interviewer segment
Equals interviewer removal
!
Speaker subset
Annotated field recordings
User Feedback
Automatic System

Initial annotation scenarios

Scenario 1: Semi-automatic segmentation and labeling to support skimming of field recordings



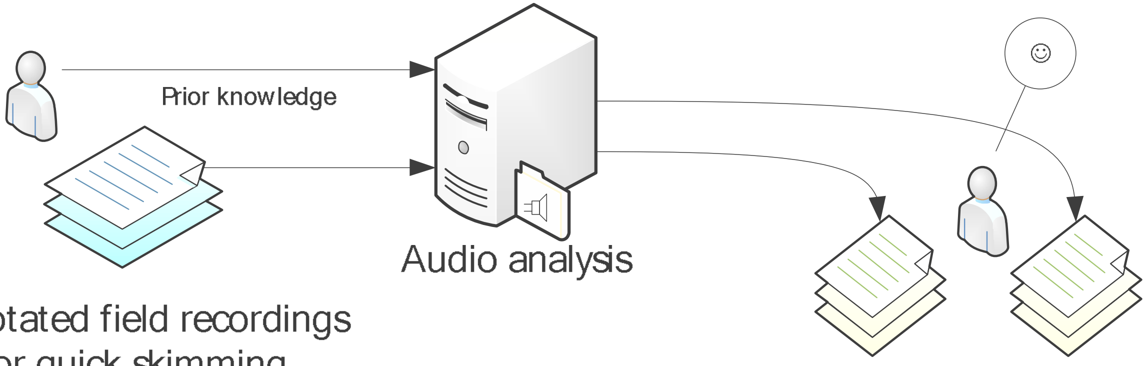
Prior knowledge

Unstructured field recordings

Audio analysis

Annotated field recordings for quick skimming

Scenario 2: Semi-automatic labeling of interviewers and subjects



Prior knowledge

Audio analysis

Annotated field recordings for quick skimming

Subject recordings

Interviewer recordings

Are these scenarios relevant for you?

Other typical annotation scenarios?

AVATecH audio detectors: State of the art

- Audio segmentation
  - *Autonomously splits audio stream into homogeneous segments*
  - Using Dynamic Programming / Bayesian Information Criterion (BIC)
  - Baseline with MFCC features
  - We investigate noise-robust features using spectral auto-correlation (SAC)
  - Essential pre-processing, works well on non-noisy data
- Speech/Non-speech detection
  - *Detects whether a segment contains speech or not*
  - Based on GMMs with MFCCs/SAC
  - Works well with in-domain training data
  - Integrate user-driven feedback mechanism for adaptation
  - Similar: Gender Detection

AVATecH audio detectors: State of the art

- Speaker clustering
  - *Joins and labels segments with the same speaker*
  - Based on Bayesian Information Criterion
  - Works well on Broadcast data, e.g. for detection of anchor person
  - Poor results on most AVATecH corpora, robustification needed
  - How to integrate user-feedback?
  - High time complexity of clustering – what about large collections?
- Speaker Identification
  - *Identifies segments with known speakers in a given corpus*
  - Using spectral and pronunciation features
  - Plan to integrate user-driven mechanisms to automatically train new speaker models

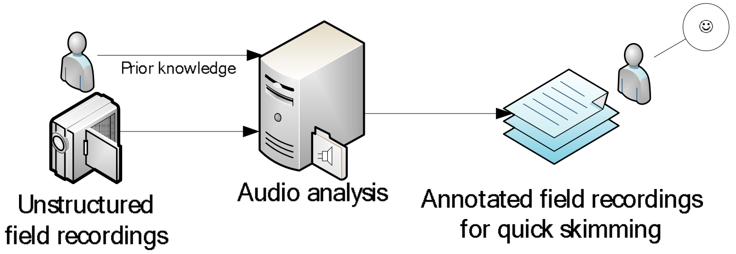
Outlook

- Language Independent Alignment
  - Approach: Top-Down method (from paragraph to word level) using different language-independent features
    - Histogram-like matching of repetitive patterns in text and audio
    - Optional anchor points available through user-feedback
  - Core difficulties: Lack of language model & noisy data

- Acoustic Query-By-Example
  - Find repeated similar audio events by marking one example
  - Approach: Fast matching in pre-computed feature index
    - Detection and discrimination of linear and noise-like spectral features
    - Sparse point of interest encoding
    - Idea from animal sound discovery

Demonstration

Scenario 1



Prior knowledge

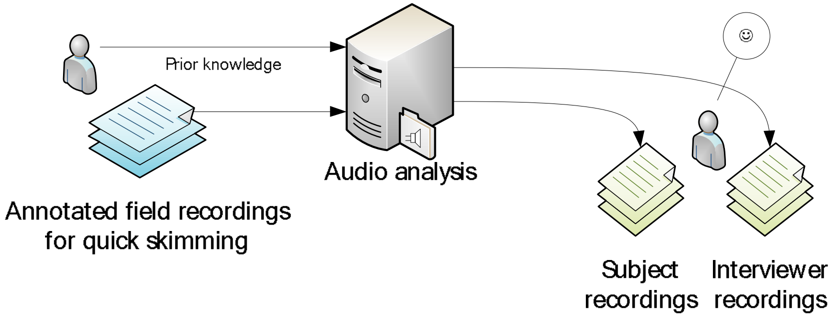
Unstructured field recordings

Audio analysis

Annotated field recordings for quick skimming

☺

Scenario 2



Prior knowledge

Annotated field recordings for quick skimming

Audio analysis

☺

Subject recordings

Interviewer recordings

Thank you for your attention!