ICSI Experience

Chuck Wooters

Senior Research Engineer

International Computer Science Institute, Berkeley, California, USA

ICSI Work On Meetings

- Began collecting data in Feb 2000
- Primary collaborations with UW, SRI; also with OGI, Columbia U, IBM; new ones with IM2, M4.
- Fundamental Goal: technology to process spoken language from “natural” meetings

Types of Meetings

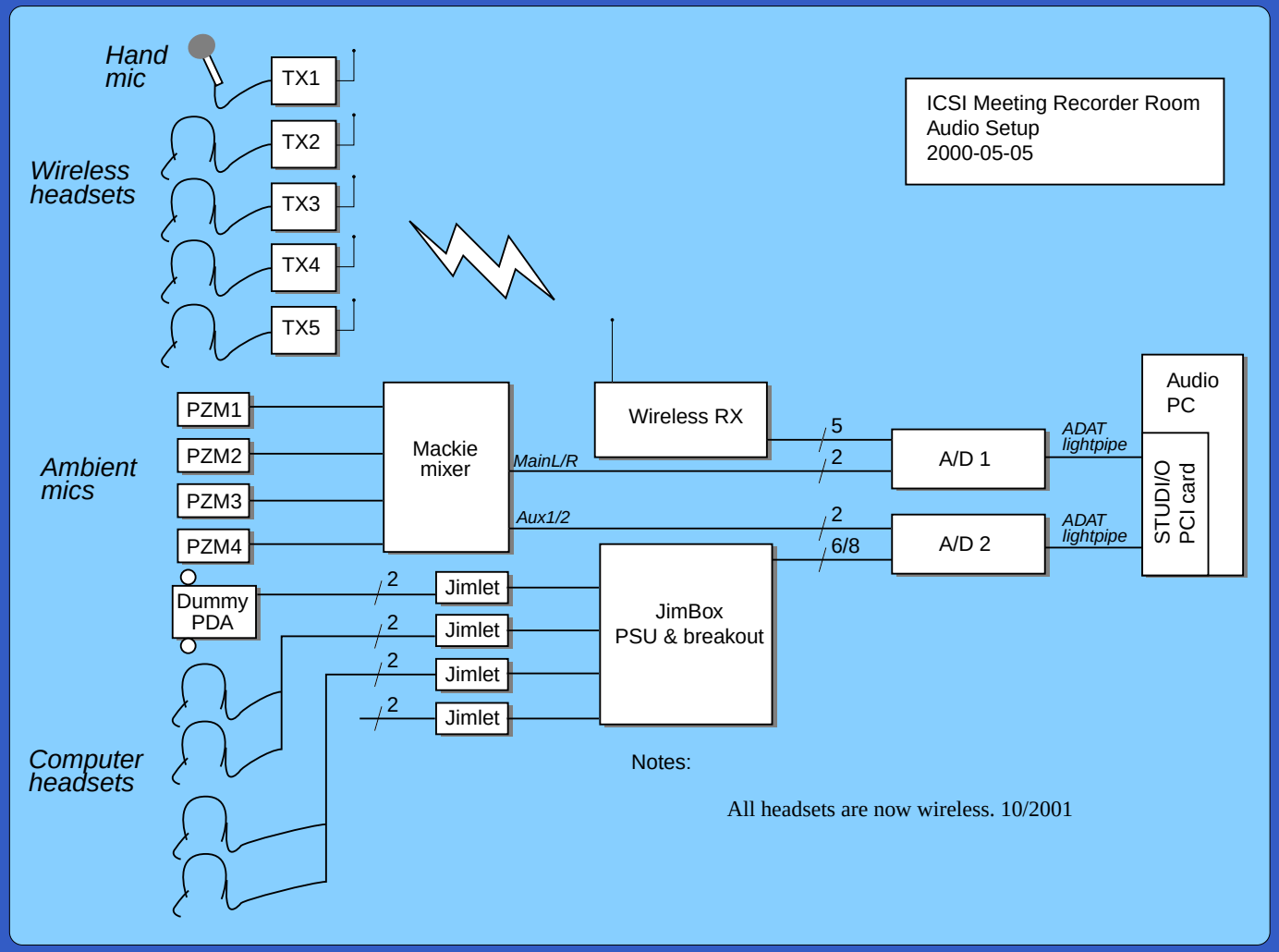
- Regular, weekly group meetings
- “Natural” data (meetings that would happen even if we weren’t recording)
- Close-talking and far-field microphones
- Digits: provide a baseline task for far-field signals
- Up to 10 speakers per meeting (averaging around 6)
- Few meeting types, but many tokens

Meeting Room

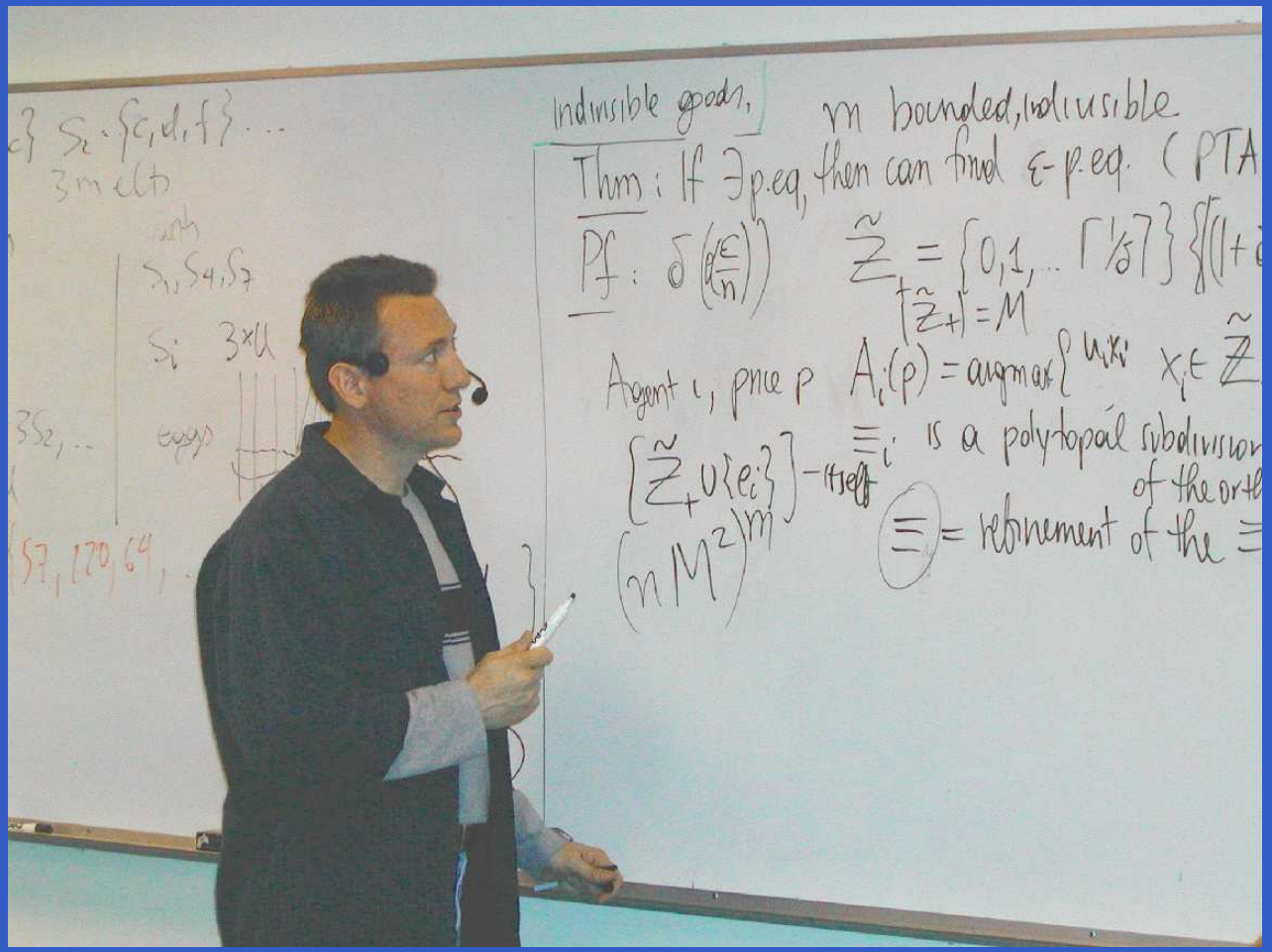
Data Collection Process

- Audio format: NIST Sphere, shortened (compressed), 16 KHz, 16 bit
- Up to 16 Channels (each in its own file):
  - 2 “PDA” mics
  - 4 PZM omni-directional (table-top) mics
  - 10 (max) close-talking (Sony® and Crown®, mostly radio)

Meeting Room Hardware

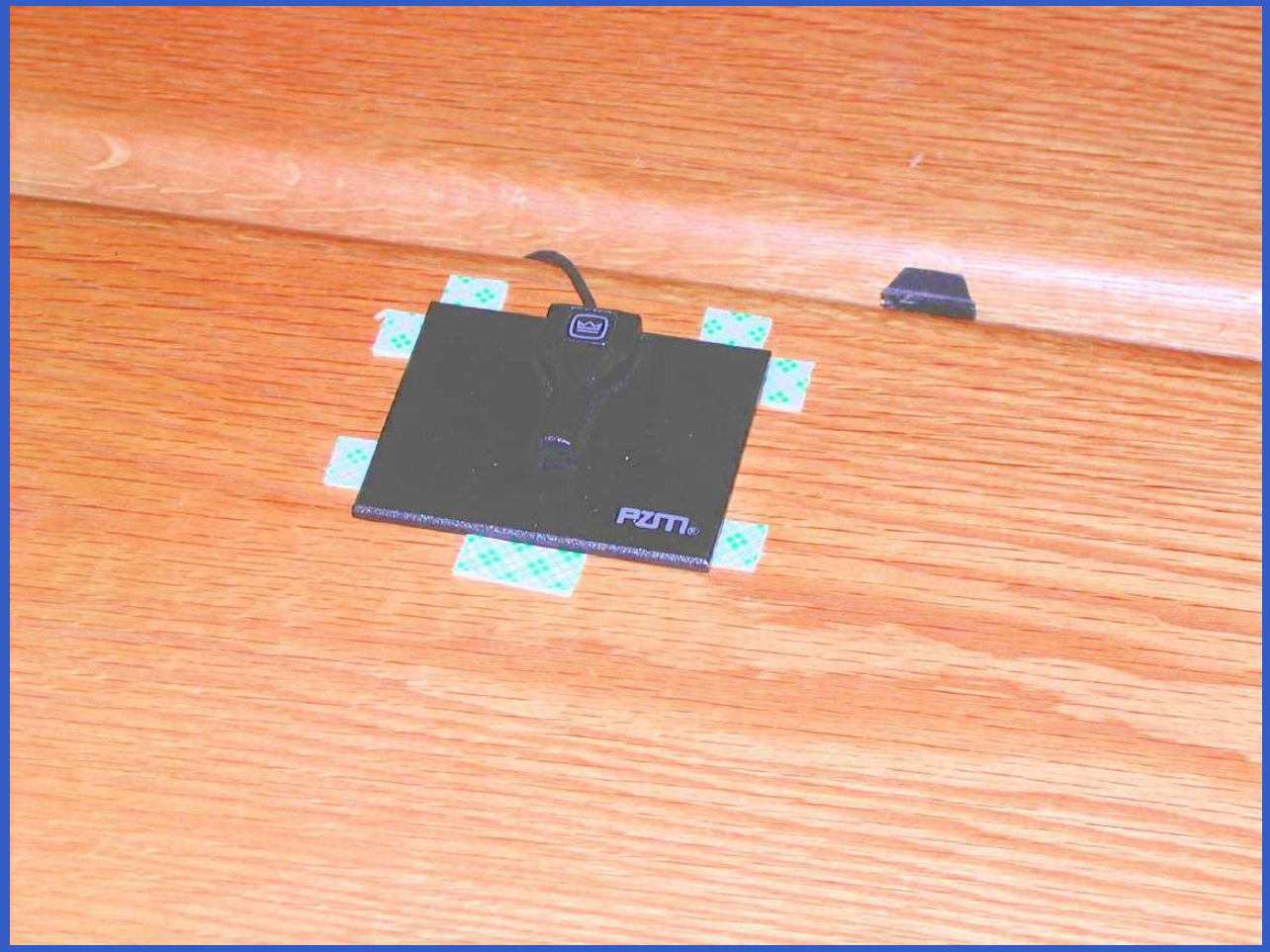


Wireless mics



PDA mics

PZM mics



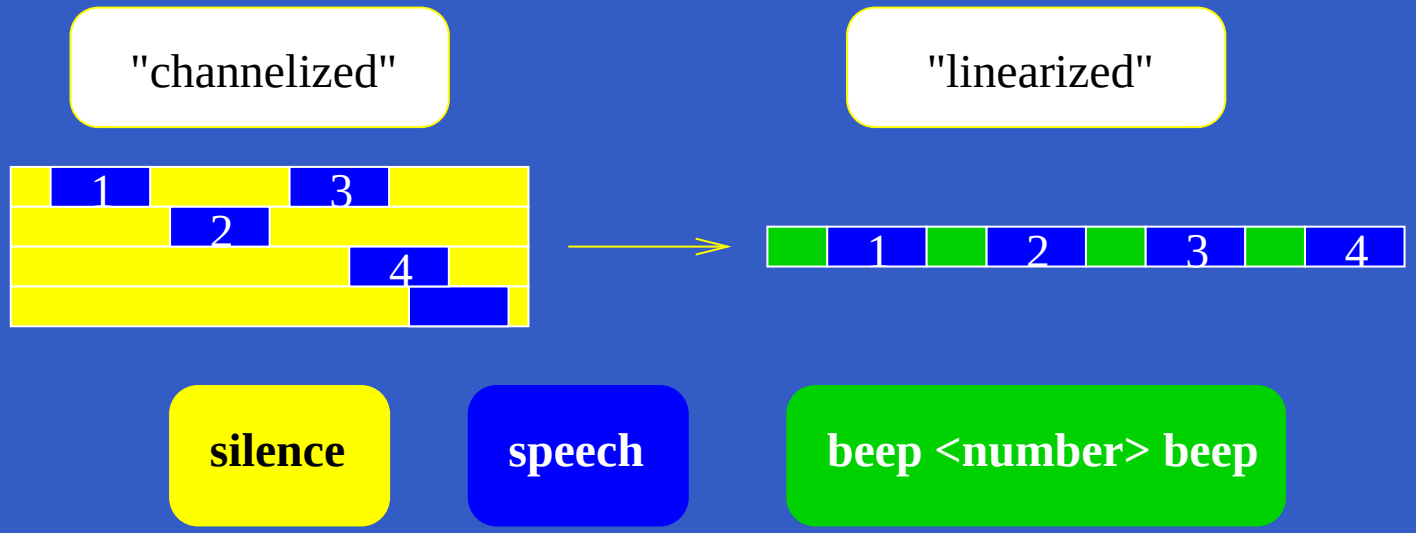
Transcription File Format

XML based on the following:

- ETCA “Transcriber” tool.
- TEI (Text Encoding Initiative), especially for time concepts.
- Annotated Transcription Graphs of Liberman, Bird et. al. — ATLAS (Architecture and Tools for Linguistic Analysis Systems).

Transcription Tools

- trans $\longrightarrow$ channeltrans
- “linearizing” transcripts (for fast first-pass transcription)

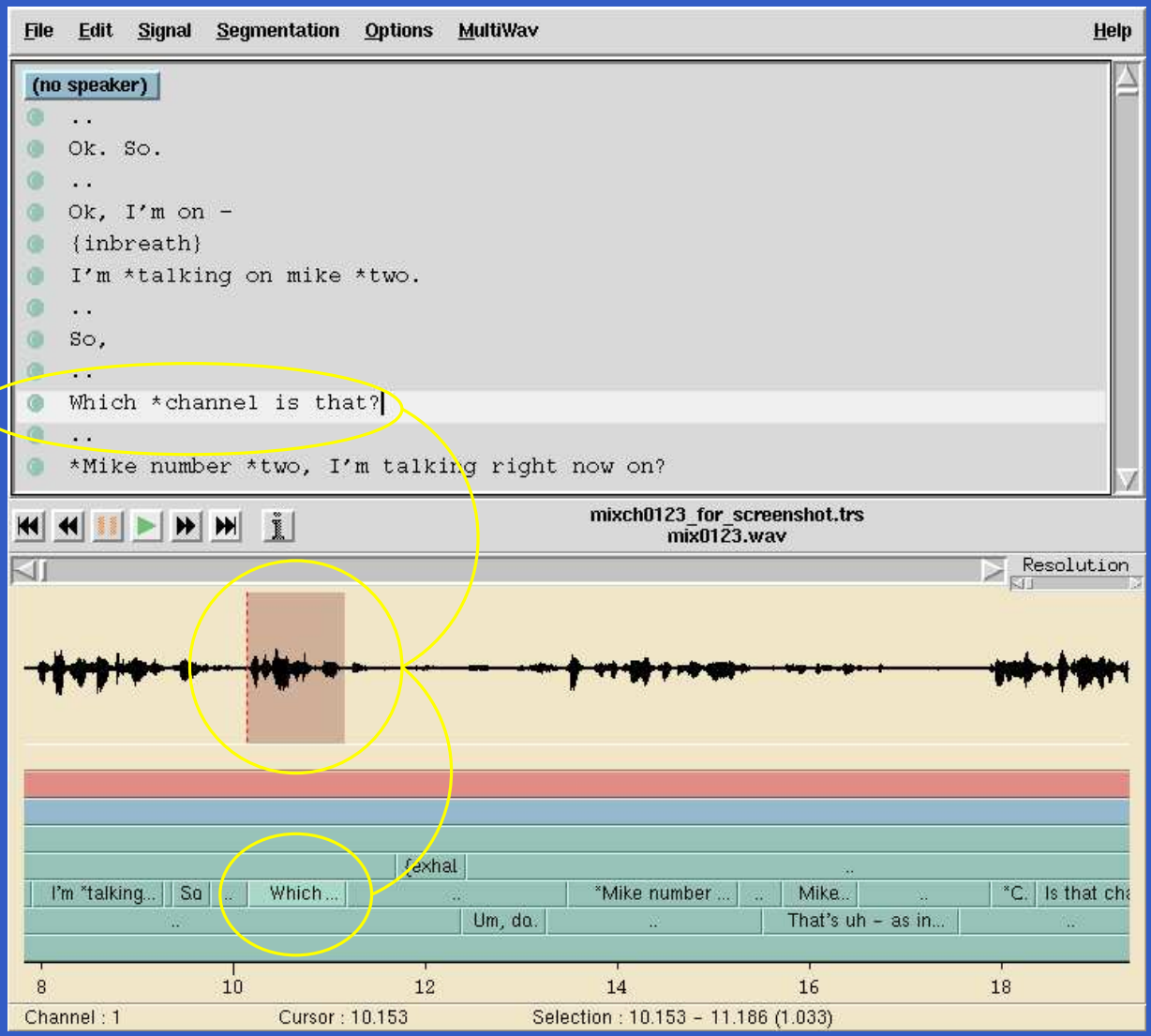


Transcription Tools (cont.)

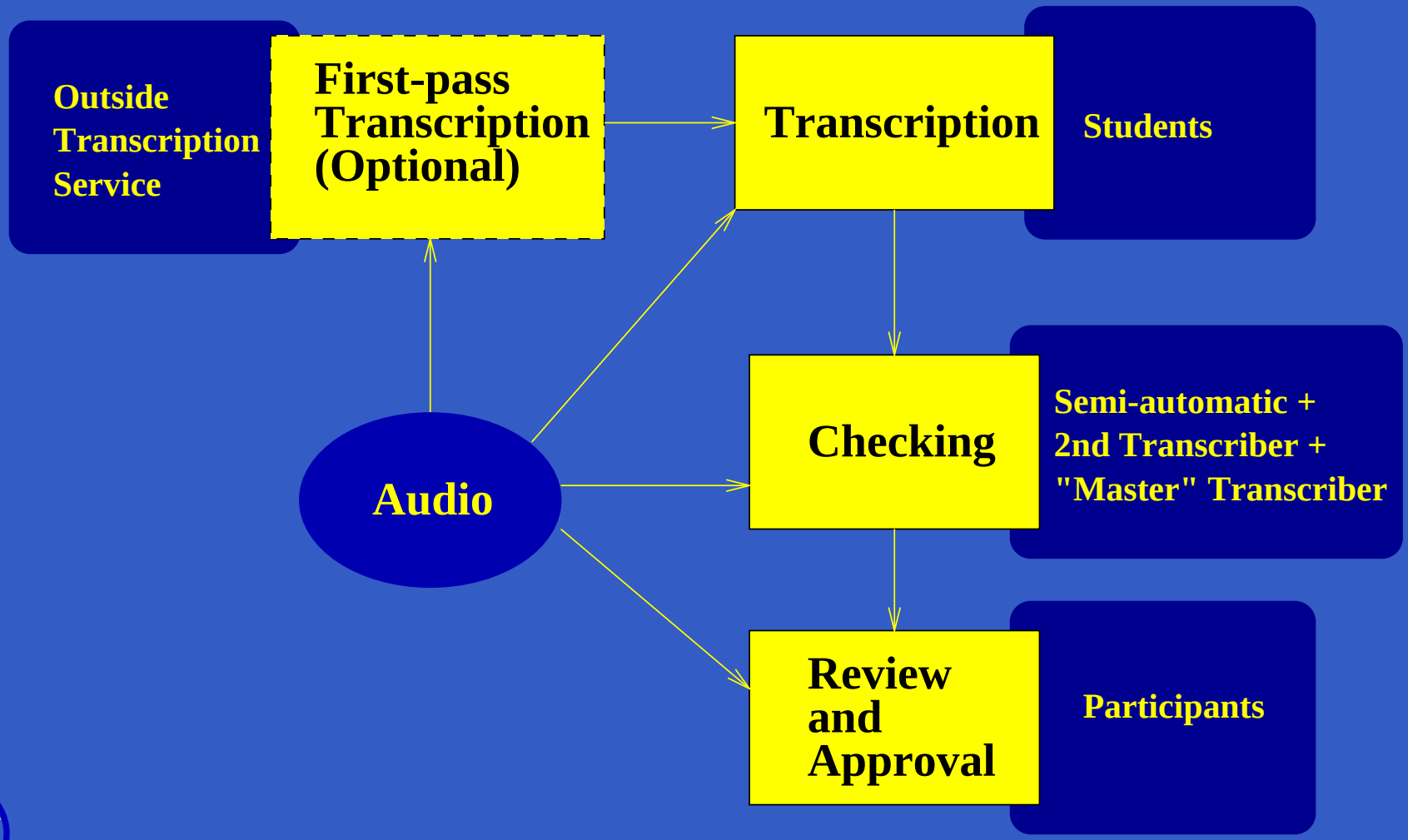
Transcription for the current channel

Current chan

All channels



Transcription Process



What do we transcribe? (Part I)

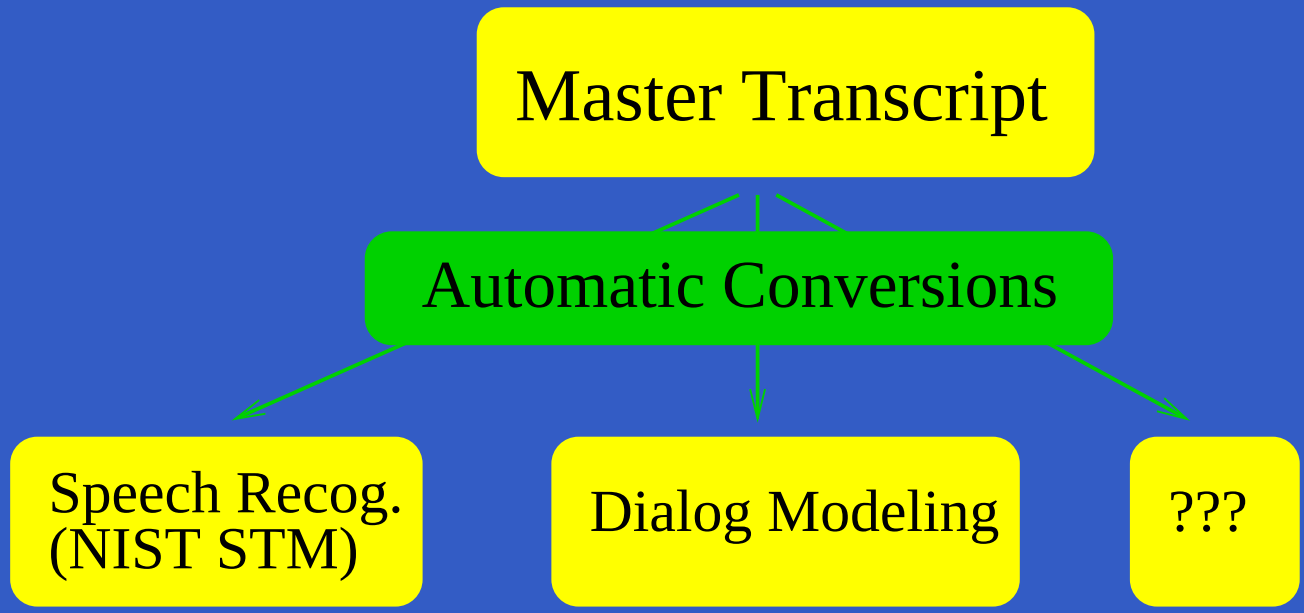
- Speakers, channels
- Words (plus abbreviations, acronyms, etc.)
- Overlaps (recoverable from time marks)
- Disfluencies (e.g. um, eh & interruptions)
- Backchannels (e.g. uh-huh)
- Non-canonical pronunciations
- False-starts

What do we transcribe? (Part II)

- Non-lexical events:
  - vocal: cough, laugh, breath, etc.
  - non-vocal: door slam, paper noise, etc.
- Acoustic uncertainty
- Qualifying information & contextual remarks
- “Bleeps”
- Utterance boundaries (via standard orthographic conventions)

Transcript “Transformations”

Master XML transcript is transformed to application specific versions.



Corpus Status

- 83 meeting-hours (87 meetings)
- 1045 total channel-hours recorded
- 547 close-talking hours (3-10 channels per meeting)
- 498 far-field hours (6 channels per meeting)
- about 60 unique speakers

Corpus Status (cont)

- Mostly transcribed, but going through final checking and approval stages
- Connection to NIST efforts: transcription standards for NIST recordings, RT-02 dev and eval data
- Planning to distribute corpus via LDC

Open Issues

- How do we distribute the data?
  - Estimating 50 Gigs of data for 100 hours
- “Bleeping” vs. discarding entire meeting
- What gets transcribed? (Can’t anticipate all desired levels of annotation nor all potential applications.)
- Legal “responsibility” of organization collecting the data.

References

- ICSI Meeting Recorder Project:
  http://www.icsi.berkeley.edu/Speech/mr/
- ETCA “Transcriber” tool:
  http://www.etca.fr/CTA/gip/Projects/Transcriber/
- TEI (Text Encoding Initiative):
  Main site: http://www.tei-c.org/
  XML for TEI Lite:
  http://www.oasis-open.org/cover/tei.html
- ATLAS — http://www.nist.gov/speech/atlas/

References (cont)

- Annotation graphs (which are a special case of ATLAS):
  http://morph.ldc.upenn.edu/AG/
- EAGLES (Expert Advisory Group on Language Engineering Standards):
  http://www.ling.lancs.ac.uk/eagles/delivera/wp4aug1.html
- MATE (Multilevel Annotation, Tools Engineering):
  http://mate.nis.sdu.dk/