Automatic Annotation of Media Field Recordings with AVATecH

Eric Auer, MPI for Psycholinguistics, 8/2010
www.mpi.nl/avatech

A joint project of MPI for Psycholinguistics Nijmegen,
Fraunhofer IAIS St. Augustin,
Fraunhofer HHI Berlin and
MPI for Social Anthropology Halle
What is AVATecH?

- Linguists collect increasing amount of recordings (cannot expect studio quality / world languages)
- Transcribe 1 hour in 35 hours, translate in 25... → Lots of audio / video material barely annotated
- Valuable material stays unused or even unfound
- Idea: Robust, easy (semi-) automatic annotation
- Assist end-user to annotate recordings faster/easier (cannot expect signal processing expertise)
Test corpus and user interfaces

- First sample: 55 audio scenes, 30 had wind noise, 15 reverb, 15 other noise, 5 humming - only 20 'easy'
- No suitable models exist for endangered languages
- Collected test corpus of 500 GB MPEG (1, 2, 4) video and 300 GB WAV audio from MPI archives
- Annotation by recognizers run in ELAN annot GUI: Open source, multi platform, flexible import/export
- AVATecH Batch eXecutor ABAX also available
- First experiences with a simple silence detector, user can inspect results immediately, can retry if needed
The AVATecH framework

Metadata

Parameters and filenames

<PARAM>
  <param name="source="/tmp/avatech/demo.wav"></param>
  <param name="output_f0_db"></param>
  <param name="vowel_stats="/tmp/avatech/demo_vowel_stats.csv"></param>
  <param name="voicing_threshold">0.45</param>
  <param name="loglevel">normal</param>
  <param name="precalc_f0_db"></param>
  <param name="max_f0">550.0</param>
  <param name="silence_threshold">0.03</param>
  <param name="vowel_stats_xml"></param>
  <param name="vowel_f0_db="/tmp/avatech/demo_vowel_f0_db.csv"></param>
  <param name="epsilon">2.0</param>
</PARAM>
Use of recognizers in ELAN

- Example: Silence detection
- Parameters
- Video
- Timeseries
- Waveform
- Annotation
- Result: Tier
Audio recognizers: Speech

- **Input:** Audio with segmentation at significant changes
- **Training material:** 2 to 3 hours of speech (once)
- **Gaussian mixture models for speech and non-speech**

- **Output:** Segments with / without speech as tiers
- **Use:** Segmentation, input for recognizers
Audio recognizers: Speaker detection

- Input: Audio, segmentation, optionally speech detection output
- Training material: 1 to 5 hours per gender for general human voice model, 5 minutes per speaker model (all GMM again)
- Output: Most likely speaker per segment
- Use: Tag interviewer versus other speakers (combine speech det.)
Audio recognizers: Speaker clustering

- **Input:** Audio, segmentation
- **Training material:** None! Algorithm first assumes each segment is a new speaker...
- **Clusters similar** (GMM, Bayesian Info. Criterion dist.), threshold
- **Output:** Tiers marking speech segments of each found speaker
- **Use:** Get basic structure of recording to annotate
Audio recognizers: More ideas

- Intonation and pitch contour pattern search
- Time alignment of words, sounds and utterances *(often only transcriptions exist yet)*
- Speaking rate detection
- more...?

- Word spotting
- Speech recognition
- Language (family) detection
- Gender detection
- Phonetic transcription
- more...?
Video recognizers: Shot detection

- Shot = segment of video of 1 camera operation
- Not just comparing pixels (movement, LQ)
- Special: fade, wipe, ...
- Motion compensation, adaptive threshold
- Measure edge changes
- Filter flashlight events
- Use: Get overview, zap
Video recognizers: Motion estimation

- Measure movement per sub-block of picture
- Histogram of directions of (big) motion vectors
- Hybrid Recursive Matching for real-time block motion estimates
- Compare by diff/diff² or normalized cross-correl
- Use: Detect camera movements, e.g. pan
- Also: Detect zoom (or cam in vehicle etc), tilt
- Distinguish cam versus content movements
- Avoid analysis of parts with camera moves
- Watch the demo
Video recognizers: Skin area tracking

- Histogram YUV colors, morphological filtering
- Pick UV range, then related Y range
- Find skin color pixels, then grow to areas
- Fit ellipses to areas
- Track area movements
- No body model yet

- Use skin: Input to other detectors (face, gesture, mimics, head, eyes...)
- Use area: Gesture / sign activity, segmentation
Video recognizers: Face detection

- Train cascade of simple features (Viola-Jones, AdaBoost) as model of what e.g. frontal face is
- Search area for matches (at different sizes, too!)
- Pool clouds of matches
- Combine other clues to reduce false positives

- Use: Count visible people for annotation, detect scene changes
- Also as input for recog., mimics, eye, head move
AVATecH results so far

• New steps in facilitating the work of annotators
• Easy to use, robust detectors integrated into the annotation work flow. Common interface and GUI
• Tool for batch processing of multiple recordings
• Helps to scan existing corpora for valuable data
• Helps to annotate more new corpora faster / better

• Questions? Ideas?