BabyExp: Constructing a huge multimodal resource to acquire commonsense knowledge like children do

Massimo Poesio[1], Marco Baroni[1], Oswald Lanz[2], Alessandro Lenci[3], Alexandros Potamianos[4], Hinrich Schütze[5], Sabine Schulte im Walde[5], Luca Surian[1]

[1]University of Trento, [2]FBK, [3]University of Pisa, [4]Tech. Univ. of Crete, [5]University of Stuttgart

1. The BabyExp project

- Current knowledge extraction methods are mostly trained on huge amounts of raw text (e.g., from the Web or the Wikipedia), although this sort of input is hopelessly impoverished compared to the rich environmental stimuli available to humans when they learn about the world
- Acquisition during childhood has three key aspects:
  1. *multimodal integration:* in human learning, non-verbal perceptual experience and visual exerience in particular crucially complements verbal information, and has a dominant role in the acquisition of particular categories and aspects of our knowledge
  2. *incrementality:* human children are exposed to increasingly more varied stimuli as time and their learning capacities increase
  3. *full immersion in a "noisy" environment:* children learn how to carve knowledge not living in a controlled laboratory, in which stimuli are presented to them in a piecemeal and regular way. Instead, they are constantly immersed in an environment full of "noise", from which they learn how to distill the relevant pieces of information
- **BabyExp** is a radically new kind of corpus, based on continuous audio and video recordings of the full indoor waking hours of a single child in an English-speaking environment
- The audio and video streams will be automatically transcribed using state-of-the-art speech recognition and person and object recognition and attention tracking techniques
- The resulting textually encoded corpus will capture utterances heard by the child as well as trajectories and various visual properties of persons and objects surrounding the child, and that the child is paying attention to
- The BabyExp project is structured into the following main components:
  1. Data collection in the child's house
  2. Audio stream transcription
  3. Video stream transcription
  4. Corpus construction from the transcriptions
  5. Commonsense knowledge extraction from the corpus

2. Data Collection and Pilot Studies

Data collection:

- Data collection started in September 2008 and it will end in August 2011 (the child was born in August 2008, and data have been collected since his second month of life)
- Data collection takes place in the PI's apartment, with the collaboration of the PI's wife (the mother of the child)
- Child room, living room and kitchen area are equipped with non-invasive cameras mounted at the 4 corners of the ceiling with environmental microphones attached to one camera
- Recorded data are periodically transferred from local server to a University of Trento server cluster, after the parents monitored them at high speed to filter out sensitive data
- As of May 2010, about 1.5 terabytes of raw data have been collected

Pilot study 1:

- Recordings from a home environment contain a large variety of (often overlapping) signals
- Goal: categorize into child speech, child-directed speech, adult-adult conversations, TV/Radio, noise or other sources
- Most important cues: short-time energy, followed by short-time spectral envelope and fundamental frequency, cf. Section 3
- Setting: position of the microphone is fixed, the distance between the speaker(s) and the microphone is variable and this significantly affects the short-time energy; but position of the speakers can be determined (with relative accuracy) from the multi-camera data

Pilot study 2:

- Continuous transcription of the visual information the child has access to will be acquired from video recordings, cf. Section 4
- Cues: (i) tracking the spatial position and head orientation of the baby, (ii) the position of the adults the baby interacts with, and (iii) detection and localization of objects of interest to common sense acquisition
- Visual focus of attention of the baby can then be logged in a post-processing step, by intersecting the viewing cone of the baby with the trajectories of the adults/objects collected in its surrounding environment
- Challenges: low resolution (facial or object features not visible in images), uneven lighting conditions (same color appears brighter near a window), and high ambiguity in the visual characterization of the baby (no hairs, facial features, complex shapes and poses)

3. Pilot study 1: Automated speech analysis

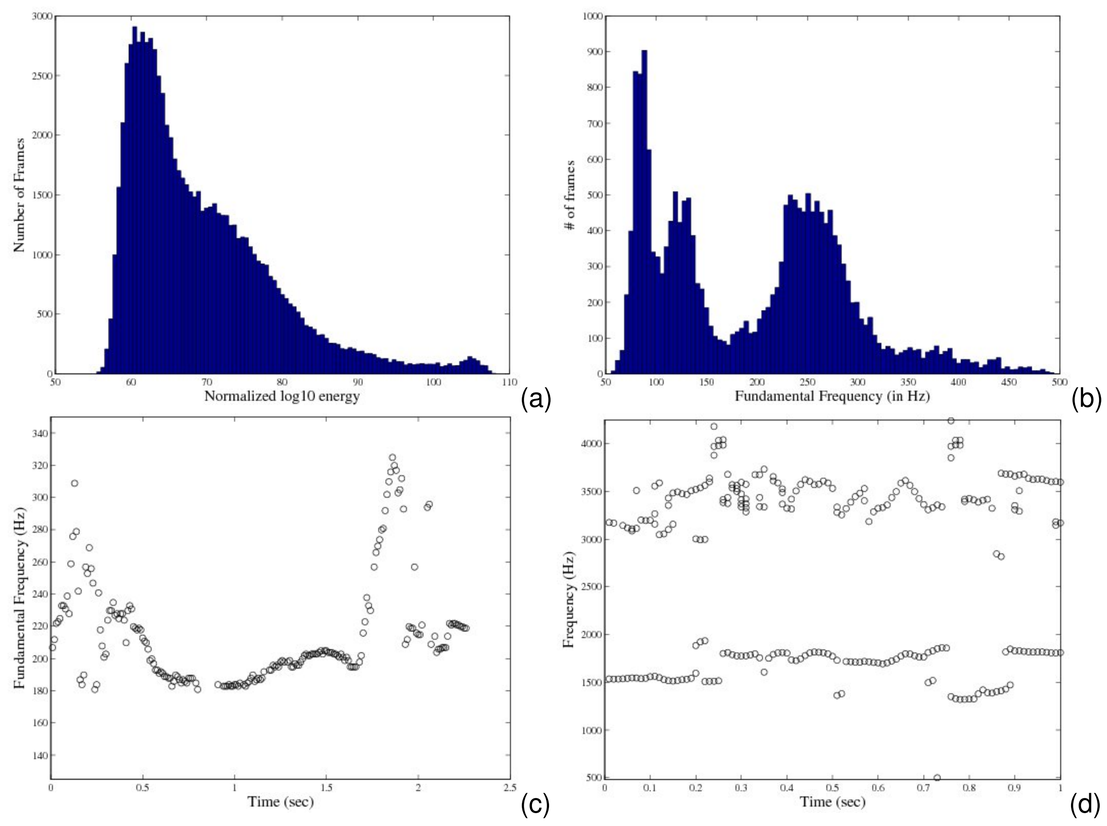

Figure 1: *(a) Short-term energy histogram and (b) fundamental frequency histogram for a typical session. (c) Fundamental frequency contour (raw estimates) for the sentences "Oh. Look at you" (in motherese). (d) Formant frequency raw estimates for an example of the infant crying.*

4. Pilot study 2: Automated video transcription

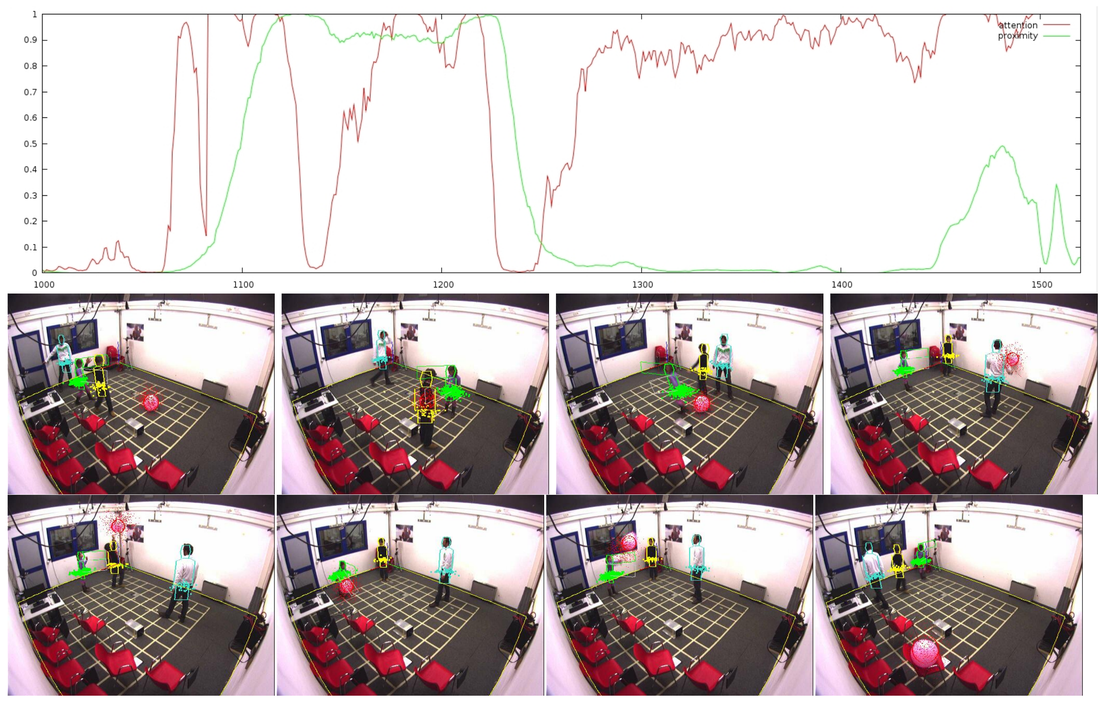

Figure 2: *Automated video transcription of pilot study sequence: visual focus of attention and proximity of the study subject towards the attention object over time (520 frames ≈ 35sec), and raw output of SmarTrack on frames 1085, 1160, 1230, 1370, 1535, 1785, 1860, 2240. The real time location and head orientation estimates of the study target are overlaid in green, and the locations of the attention object (the pink ball) and the actors are shown in red, yellow and blue.*

5. Commonsense knowledge extraction

- The main goal of the BabyExp consortium is the development of algorithms for commonsense knowledge extraction that exploit the innovative properties of the BabyExp corpus
- Focus on:
  - Cross-situational learning of entity-word association
  - Extraction of discrete feature-based representations of image schemas from video stream
  - Extraction of action structures associated to verbs from combined video-audio streams
  - Abstract concept acquisition