Transliteration of Language and Labeling of Emotion and Gestures in SmartKom

Silke Steininger,
Institute of Phonetics and Speech Communication,
Schellingstr. 3, 80799 Munich, Germany

Abstract

The requirements for the transliteration of spontaneous speech data and the labeling of video data in the project SmartKom will be explained. For this we give an overview of the goals of the project. The planned structure of the labeling system of the audio and video data is detailed. Since the system for the video data is still in development the focus of the contribution will be on the specific problems and requirements for labeling of multimodal data with regard to the SmartKom project.

The SmartKom Project[1]

The goal of the SmartKom project (started in January 2000) is the development of an intelligent computer−user interface that allows a computer novice to communicate naturally with an adaptive and self−explanatory machine. New possibilities of the interaction between human and machine are investigated: The system does not only allow input in the form of natural speech but also in the form of gestures. Additionally the facial expression is analyzed. The output of the system is presented with a graphic user interface and with synthesized language. The user interface consists of a computer screen that is projected onto a graph tablet.

To explore how users interact with a machine that has far greater communication skills than we are used to at the moment, data is collected in so−called Wizard−of−Oz experiments: The subjects have to solve certain tasks with the system (like planning a trip to the cinema, programming a VCR or navigating in a foreign town). They are made believe that the system they interact with is already fully functional. Actually many functions are only simulated by two "wizards" that control the system from another room.

The different functionalities of the system are developed by different partners. The Institute of Phonetics and Speech Communication in Munich is responsible for the collection of the multimodal data and the evaluation of the system.

Collection Of Multimodal Data

In the first phase of the project the collected data is needed for three different main purposes:
1. The training of speech, gesture and emotion recognizers.
2. The development of user−, language−, dialogue−models etc. and of a speech synthesis module.
3. The general evaluation of the behavior of the subjects in the interaction with the machine.

In each session of a Wizard−of−Oz experiment the spontaneous speech, the facial expression and the gestures of the subjects are recorded.

For the **audio** recordings we use:

- a microphone array (4)
- a directional microphone and
- (alternating) a headset or a clip−on−microphone.

Video:

- For the *facial expression* a digital camera captures the face of the subjects.
- For the gestures a second digital camera captures a *side view of the subject* (full height)
- and an infrared camera (from a *gesture recognizer*: SIVIT/Siemens) captures the hand gestures (2−dimensional).

Other:

- The coordinates of pointing gestures on the work space are recorded (with Sivit)
- as well as the inputs of a pen on the graph tablet.

Labeling Of The Data[2]

1. Speech

The recorded spontaneous speech (the dialogue between user and machine) is labeled on the word level using a broad orthographic transliteration system. It is similar to the system used in the project Verbmobil[3]. Complementary to the orthography a list of conventions is used to code such things as corrections, repetitions, reductions, hesitations as well as technical artifacts and superimposition of the speech of the two dialogue partners (the user and the machine).

a) The transliterated categories

- lexical units (e.g. words, classified words,

[1] http://smartkom.dfki.de/index.html

[2] For a detailed description of the transliteration system see: Nicole Beringer, Daniela Oppermann, Susanne Burger: Transliteration spontansprachlicher Daten − Lexikon der Transliterationskonventionen − SmartKom (Version 1). SmartKom Technisches Dokument Nr. 2, Februar 2000.

[3] http://www.phonetik.uni−muenchen.de/Bas/BasKorporaeng.html; http://verbmobil.dfki.de/overview−us.html

compounds)
- syntactical−semantical structure (e.g. sentences, repetitions, false starts)
- non−verbal articulatory productions (e.g. hesitations, breathing, laughing)
- noises, technical artifacts
- pauses
- acoustic superimpositions
- comments (peculiarities of the grammar; pronunciation)
- prosody (phrase boundaries, accentuation)

b) Technical requirements

The code is processed automatically, therefore it is required that it uses

- a consistent file structure
- informative and consistent turn names
- consistent transliterations
- non−ambiguous symbols
- ASCII−character set
- transliteration conventions that can be parsed

c) Requirements in respect of content

To be useful for the training of the speech recognizers and the evaluation of the dialogues, the code has to make sure that

- all perceivable dialogue elements can be described with the code
- syntactical−semantic markers can be placed
- speaker and noise superimpositions can be identified
- specific word categories (names, numbers) and faulty expressions can be identified
- the text files are as readable as possible

d) Limits of the transliteration system

The annotation is broad, i.e. not phonologically or phonetically. Furthermore no temporal alignment with the audio signals is done.
Noises and non−speech sounds are not described in detail.

e) Differences to Verbmobil

In contrast to Verbmobil a simpler system is used. The annotations for noises were reduced even further to fasten the processing time.
On the other hand some prosodic markers were introduced: phrase boundaries, intonation at phrase boundary, accent.

2. Gestures

The labeling system for the gestures is still in development. Most existing systems are far too specific for the purposes in SmartKom (for example methods for transcribing American Sign Language − see SignStream[4] or the "Facial Action Coding System"[5] of Ekman). Instead of coding the precise morphological shape we will try to use a simplified, practice−oriented system. Two broad categories are labeled − head gestures and hand gestures.

The hand gestures are defined functionally/intentionally (not morphologically). The definition of a unit is no small problem when labeling gestures. We will try to define the unit with regard to the intention of the user, i.e. with regard to his (assumed) discrete goal.

Each hand gesture−unit will be coded with its

- function (e.g. "pointing", "no", "back" etc.)
- reference (e.g. "Button X", "Region Y", "nothing")
- broad morphological form (e.g. "one hand finger pointing", "one hand circle", "two hand crossing")
- beginning/end
- reference word in the audio channel (e.g. "this", "here", "no")

The head gestures are coded with regard to three broad morphological categories:

- head rotation
- head incline forward/backward
- head incline sideward

The major problems we face are:

- How should a **unit** be defined? Beginning and end of a gesture cannot be perceived easily.
- **Meaning**: Morphological similar gestures can have different meanings.
- **Description**: Complex gestures can only be described roughly.
- **Reference**: The reference word or the reference location cannot always be determined easily.
- **Categories**: At the beginning of the project it is not known which categories will emerge as useful.

So on the one hand with gestures we seem to face a simpler situation than with speech: There are no turns and probably no continuous stream of gestures, only solitary events. The pool of possible gestures will be much smaller and we use only broad categories.
On the other hand an orthographic transliteration has fewer problems in finding reasonable units. We plan to define a unit intentionally, but the question is: How reliably can the intention of the subject be determined? Can functional gestures consistently be distinguished from mere manipulators?
The development of a simple, but meaningful description system for the broad morphological category is another challenge.

Our goal therefore is to develop a relatively simple, relatively fast, consistent description system for

[4] http://web.bu.edu/asllrp/SignStream/

[5] Ekman, P., & Friesen, W. V., Facial Action Coding System (FACS): A technique for the measurement of facial action. Palo Alto, Ca.: Comsulting Psychologists Press, 1978.

functional gestures (in the context of a human−machine dialogue situation) that is useful for analyzing video data and the the training of gestures recognizers.

3. Facial Expression

The labeling system for the facial expression is still in development as well. At the moment it is planned to label emotional facial expressions in six categories:

- anger/irritation
- boredom/lack of interest
- joy/gratification (being successful)
- surprise/amazement
- neutral/anything else
- face partly not visible

The impressions will be rated as weak or strong. We will test if it is useful to integrate information from the audio channel: The emotional expression of the voice will be labeled also in respect to the categories above. A label therefore will look like this: "anger, strong, source: face+voice, beginning/end" or "surprise, weak, source: face, beginning/end".

In respect to the emotions our two main problems are:

- Objectivity: The judgment of emotions varies subjectively. Emotions like "boredom" are difficult to identify, "anger" and "surprise" can be mixed quite easily.
- For practical reasons we cannot use a precise morphological coding system like the "Facial Action Coding System"[6] of Ekman that allows relatively objective categories.
- The goal will be to develop a definition system that is easy to understand and a good training method for the labeler.
- Validity: The facial expression does only in part convey the underlying emotion or its strength. This problem is not too bothering: The collected data is needed for an emotion recognizer that should be able to make educated guesses about the mood of the dialogue partner. It is sufficient that its performance is similar to that of a human dialogue partner − who is as well not always able to judge the underlying emotion of a facial expression.

4. Representation Format

For the transliteration of speech text files are generated with a modified version of xemacs. It allows a consistency check and helps with the setting of markers and structuring the turns.

With respect to the video data only a few considerations about the tool we want to use were made in this early stage, but SignStream or MediaTagger[7] are examples of the direction we want to take.

Quicktime will be used to align the different audio and video layers temporally. Other considerations about the format do not exist as yet.

Conclusion

The discussed problems are only examples from a long list of problems that have to be solved. They were chosen to demonstrate that the requirements for the labeling in SmartKom are mostly defined by the fact that the research is applied research: Time and money play an important role, the data collection serves different interests (and different research groups with different needs) and the goal is not mainly to gather knowledge but to develop a running system.

Therefore the labeling system has to be flexible enough to incorporate many needs, has to be easily understandable for the takers of the data and has to allow a fast and reliable coding of the interesting categories. As always: At lot of further research and theory is needed.

[6] Ekman, P., & Friesen, W. V., Facial Action Coding System (FACS): A technique for the measurement of facial action. Palo Alto, Ca.: Comsulting Psychologists Press, 1978.

[7] http://www.mpi.nl/world/tg/CAVA/mt/MTandDB.html