

Multimodal Action
Professor Benjamin Weiss
Quality and Usability Lab
Technische Universität Berlin
Introduction

(Refer Slide Time: 00:17)



Welcome back to multimodal interaction. The topic of this week is multimodal action.

(Refer Slide Time: 00:26)



Actually I will present this week aspects of multimodal action or interaction between humans. And this means I will focus on non-verbal aspects or non-verbal signals.

The basis for my presentation, the content is actually one book and this is Non-verbal communication in Human Interaction by Knapp and Hall. So they have several versions of this, several editions and here you have a screenshot of one of the books that I actually use.

I will start with an introduction what actually non-verbal and verbal signals and interaction means. And after that I will rush a little bit through different aspects of non-verbal interaction and these you can see already in the outline are, for example related to gaze, posture, vocal aspects, emotions and so forth.

(Refer Slide Time: 01:21)

Verbal and nonverbal information

Conversation (verbal and nonverbal interaction) is multimodal



Have a look at this little post card from the beginning of the last century. Two people engaged in telephone conversation, but as you can see they are not only talking verbally and acoustically that means with words and on the acoustic channel but they also exhibit a lot of other signals that you can perceive visually but that is of course not transported by the telephone channel.

For example, there is a posture or some gestures or even facial expressions. All these kinds of information these signals we actually produce and perceive when we engage in a conversation with other humans. These information are important. We sometimes consciously or subconsciously perceive them. And it is a little bit awkward if they are missing. They all contribute to our conversation.

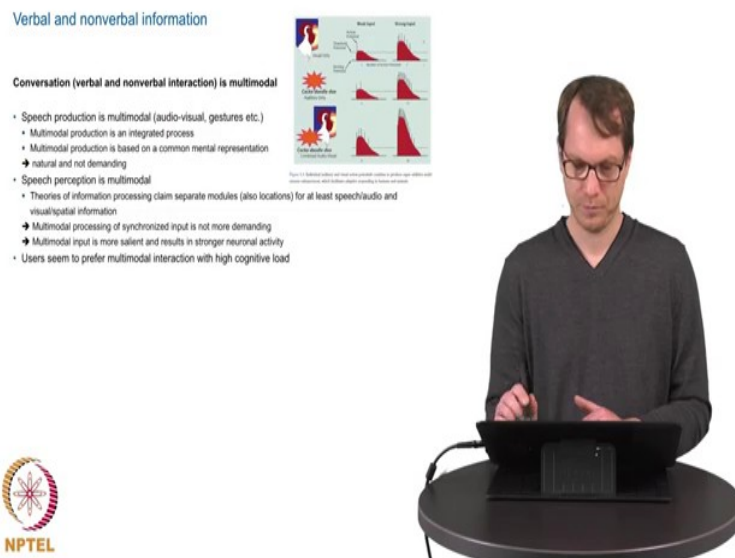
And this means they all contribute to the verbal information that we are sending and receiving. So with verbal, I am actually talking here about linguistic information, for example words and sentences and so on. And non-verbal are all the other important parts. This can also be acoustic as you can see a little bit later.

(Refer Slide Time: 02:38)

Verbal and nonverbal information

Conversation (verbal and nonverbal interaction) is multimodal

- Speech production is multimodal (audio-visual, gestures etc.)
- Multimodal production is an integrated process
- Multimodal production is based on a common mental representation
 - natural and not demanding
- Speech perception is multimodal
- Theories of information processing claim separate modules (also locations) for at least speech/audio and visual/spatial information
 - Multimodal processing of synchronized input is not more demanding
 - Multimodal input is more salient and results in stronger neuronal activity
 - Users seem to prefer multimodal interaction with high cognitive load



Just as a small recap from the last weeks, so conversation between humans is multimodal intrinsically so this means speech production is multimodal as you have seen earlier so if we produce sounds we also produce visual information of these sounds and so forth.

But also speech perception is also multimodal. That means we use these information covered by the eyes and so on in order to make senses from, or what is the actual intention of the user, the message is.

And having this multimodal interaction is not actually a problem for us. It is naturally. We are hardwired to do so.

As a repetition I presented earlier the Wickens cognitive resources which show that perceiving multimodal information is actually not more difficult than perceiving unimodal ones.

And you would also be reminded of also the small rooster example that I have also displayed here if we perceive multimodal information in this case, the shouts of this animal and also the visual information of that then the saliency in our processing is higher and even the neuronal reaction of the stimuli are even higher than just presenting one modality.

It is important to note that multimodal processing is synchronized and as I reported, as I repeat from the very first session if we produce or are engaged in conversation or interaction even with machines and cognitive load is increasing then we tend to prefer multimodal production.

(Refer Slide Time: 04:33)

Verbal and nonverbal information

Verbal and nonverbal human interaction

- Nonverbal information can ...
 - Substitute (emotions, relationships, nod).
 - Amplify (nod and "yes").
 - Contradict (e.g. for ironic effect), or
 - Modify ("a big fish")
- ... verbal information (Scherer 1979).
- Intentionally or automatically produced (Chartrand & Bargh, 1999).
- Conscious or unconscious



So how is the relationship between verbal and non-verbal signals? So we can of course produce gestures, gazes and so on in isolation but typically we produce them together with some verbal message, with some word, words and sentences utterances. And that means there is certain relationship between the verbal and the non-verbal signals.

For example here is a kind of approach by Scherer. Non-verbal information can be combined with verbal information by substitution. So for example, instead of uttering a certain word like yes or no, we use head movements. But they can also amplify by redundancy. That means if I say yes and really vividly nod with my head, then it will be a kind of amplification of what I am actually telling.

Of course for some effects, for example irony, we can also contradict, I can say no and again nod with my head. But I can also say something like I was fishing last weekend. I catch really, really a big fish and if this is contradiction in the message that we are sending this will also have a certain effect.

But the most typical one will be a modification. And this means we have a co-joined production of information, one on the verbal channel and other one on the non-verbal one.

And this modification can be just a repetition, redundancy, or really certain modification of the content, and that means if I say last weekend I was fishing and I was catching a really big or large fish and it is difference between having such a signal or such a signal.

I have to note that there are actually two ways or two characteristics of non-verbal information. And one is that it can be produced intentionally or unintentionally. For example, some facial expressions expressing my emotion can be used by purpose or unintentionally.

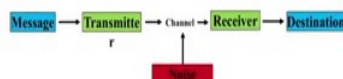
And then of course they can be unconscious. That means if I am producing unintentionally some gestures, some gaze directions I may be not aware that I am actually doing this. Sometimes playing for example with a pen or something of this may be good example for unintentional and unconscious non-verbal signals.

(Refer Slide Time: 07:23)

Verbal and nonverbal information

Conversation is multimodal

• Information transmission model (Shannon & Weaver, 1949)



Have a look at this little and quite simple communication model. You may know it already. It is a actually quite strong model and used for a lot of engineering purposes. It just means that we have a certain message; this can be verbal or non-verbal. Typically we are talking about verbal or symbolic information here. But they do not have to be this way.

And this is encoded to some, some kind of code. This is transmitted over a channel which may be affected by noise or some other effects and then decoded again to finally reach the destination, the perceiver. This of course does not differentiate between verbal and non-verbal information.

Therefore if we really want to work on and analyze non-verbal information we need a little bit more fitting model.

(Refer Slide Time: 08:20)

Verbal and nonverbal information

Conversation is multimodal

- Speech is not simple transmission of information, but social interaction
- Speech can only be interpreted in context ("underspecified") → mimic, gesture, body "language" accompany spoken language with
 - Relevant complementary information (you need them for interpretation)
 - Expected redundant information (missing them might irritate you)
- 4 Aspects of a message (Schulz v. Thun, 1981)
 - Factual information (semantic)
 - Self-revealing information
 - Information about the relationship
 - Appeal/goal
 - 4 „mouths“, 4 „ears“



And here is one. The so called 4 sides or 4 aspects of a message and the main idea is that if I as a human producing a message to communicate with other people, my dialog partners or interlocutors then there are certain aspects that I can focus on or not focus on. But usually they are all present at some point or at some degree.

For example, the most obvious one would be the factual information. This is something that can be expressed in the verbal content. This means the linguistic meaning of a sentence or

utterance. But we also have other information that we are expressing while we engage in a conversation.

One of these is the expression of my self-giving information, revealing information for my own self. This could be, for example by voice telling you my social or regional background that you can perceive, or my gender. Or with my clothing I could be expressing my own assignment of my cultural, social group. Also emotions are expressed by facial expressions, by voice or even by posture.

So for example if I would be sad I might have a different posture than if I am really happy and engaged right now.

Another one is the relationship aspect. And this means if you are, for example observing two people engaged in conversation you might already guess different roles if there is a kind of hierarchy between the two people. Or whether these two people, the couple is close to each other or do not know each other that well; for example, if they are familiar with each other.

And there is also the last aspect, the appeal or the goal of my message. So what do I really want to achieve with this? Sometimes it is also called intentions especially when I am talking about human computer interaction. But it may not be the same. So I like appeal, appeal as a good term here.

The standard example for this one is if I tell somebody in words, I am cold, it actually is factual information but my goal, my appeal is to the other, please close the window. So some of these information are verbal, mostly the factual ones.

But also there are other information. I could of course verbally tell a person how I feel. But typically this is all already transported and -produced by non-verbal information.

A typical example is two people in a car at red light, stopping. And one person tells the driver when the light which is green, the lights are green. So there is a bunch of different interpretation what aspect is actually in focus here.

Of course they are factual information. The light turned on green but there is also an appeal that could be interpreted or produced by the speaker.

For example the speaker could mean, drive, I am in a hurry. Or it could reflect on relationship so that I am, as a speaker might feel superior over the real driver and might be in the position of feeling it to be in the position to tell the driver when to start driving again.

So there are lot of aspects that might be covered by such a simple phrase. As we can see there is actually some kind of symbol here which reflects the ear and one which should illustrate a mouth. Therefore this model or approach is sometimes called also 4-mouths 4-ears model.

The whole idea behind this is if you reconsider the situation at the traffic light that I just mentioned, the person who speaks, who produces this kind of utterance, this kind of non-verbal information with the verbal information might have a different, might have certain focus on the relationship or on the factual information.

But this does not mean that the driver who is listening to this sentence is focusing on the same or hearing on the same aspect. So there can be misunderstandings.

For example if the speaker is just in a hurry and wants the driver to drive as quickly as he or she can, whereas the driver, him or herself might be little bit preoccupied or expecting to gain some critique here.

So this could be the source of misunderstanding and actually a fight. So why is this important for multimodal interaction with machines? I stress this because even when we are speaking with speech or voice or language with the verbal information, humans always produce non-verbal information.

And of course having a multimodal system, just to remind the multimodal system is a system that can process such kind of information even on a symbolic level and is able to refer to context as well. That means in such kind of hopefully intelligent system that can make sense of the user intention and the context should also be aware of at least some of these social non-verbal signals.

(Refer Slide Time: 14:09)

Verbal and nonverbal information

Conversation is multimodal

- For persuading others, not only what you say is important, but also:
 - How you say it
 - How you look like
 - How you act
- Famous examples:
 - Quintilianus, *Institutio Oratoria*.
 - Mehrabian & Wiener, 1967: The infamous formula for perceived attitude
- Some patterns are already synthesized (for HCI, but also basic research)



So as a summary, if a human person produces a sentence of verbal information then it is not only relevant how it is perceived depending on the pure linguistic content. But it is also dependent on how we say it, how we look like and how we act.

There are two famous examples that I want to mention here. There are numerous others as well. So for example, already the ancient Greeks knew about this and presented this as part of being with good lecturer or being good talker to the audience.

And there is also this one nice early study by Mehrabian and Wiener in 1967 which is often cited in popular textbooks which show that the perception of a certain word is not only dependent on the actual content but also on who is delivering this visually and vocally.

So I will not write down this infamous formula of so and so percentage of visual information or vocal information, linguistic verbal information that together comprise the whole effect on the listeners because they do not work in real life.

This is very certain, very small study I am talking about here when you will find them in popular information of course, in popular media, please do not use these except percentages

But of course the effect on the person, especially the social effect is of course highly dependent on the visual and verb, visual and vocal information and not necessarily only on the content the linguistic meaning of the utterance or sentence.

So as a short example that these information posture, mimics, facial expressions, hand gestures, gaze are actually used and produced in human computer interaction in this example of animat, embodied conversation agent as a short video which is rather old but has nicely some of these non-verbal information.

And actually please pay attention to the video. There is already some textual information telling you which to, which aspect to focus on.

(Refer Slide Time: 16:46)



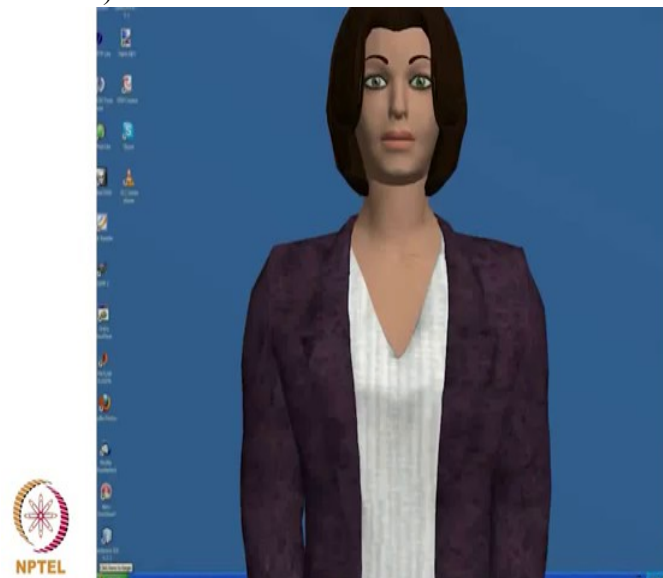
16:44 start video

Event 1: Verbal condition

(Refer Slide Time: 16:48)



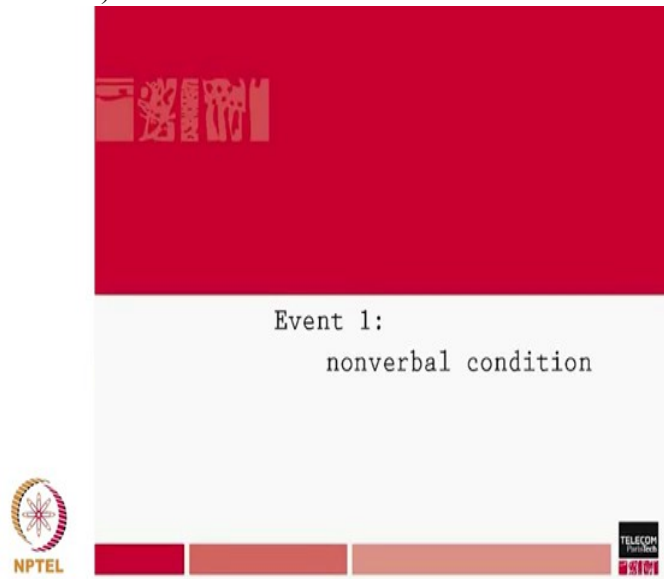
(Refer Slide Time: 16:50)



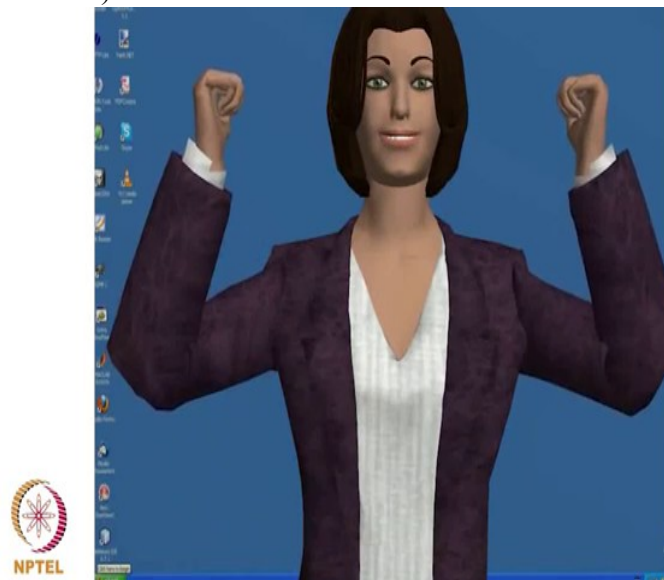
17:05

Event 1: Non-verbal condition

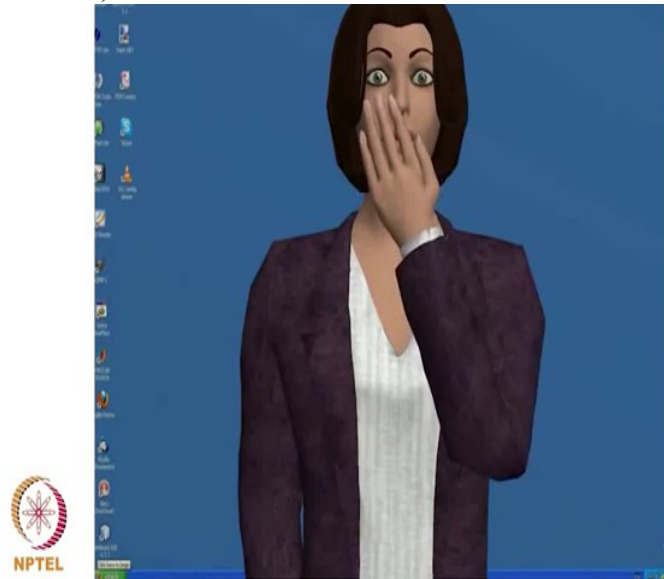
(Refer Slide Time: 17:05)



(Refer Slide Time: 17:10)



(Refer Slide Time: 17:15)



17:19

Event 1: multimodal condition

(Refer Slide Time: 17:19)



(Refer Slide Time: 17:23)



(Refer Slide Time: 17:32)



17:37 end video