



Multimodal Interaction
Multimedia and Multimodality
Professor Benjamin Weiss
Quality and Usability Lab
Technische Universität Berlin
Characteristics of Multimodal Systems

(Refer Slide Time: 00:17)

Multimedia & multimodality

Outline:

- Medium vs. modality
- Multimedia systems vs. multimodal systems
- Modality relations
- *Characteristics of multimodal systems*





The last topic of this week are characteristics of the multimodal systems and I am talking of positive characteristics here

(Refer Slide Time: 00:24)

Characteristics of multimodal systems

Making interactive systems better

1. Flexibility
2. Robustness
3. Naturalness
4. Persona Effect
5. Uncanny Valley
6. Social Facilitation / Inhibition
7. Equilibrium Theory



because we want to build better systems. There are seven characteristics and I will start with the first three.

(Refer Slide Time: 00:33)

Characteristics of multimodal systems

Flexibility:

- Provide appropriate input and output modalities for each piece of information
- Offer flexibility for different user groups, environments etc.
- → balancing user deficits, contextual limitations



Let us start with flexibility, so offering different input modes and that is the definition of modalities, as you remind, is of course nice because users can choose from a bunch of different input modes or their combinations to choose actually the best fitting modality or a combination for each piece of information they want to convey to the system.

And with multimedia systems likewise, of course having different output channels means that the system can choose the best fitting combination, or the best fitting medium in order to convey information to the user.

But there is also another kind of flexibility. This means we can cope actually for different user groups or different limits or limitations in the interaction.

If you consider for example the situation of a user driving a car, or being in the kitchen cooking, so in both situations having no hands free to use a touch screen, voice or other modalities are really nice and offer flexibilities to still use the system.

But if you consider users who have problems to articulate well, or who have trembling hands for example- having different options for input modes which means modalities, is also good to cope for these limitations and deficits.

Or if you consider environmental noise or bad lighting conditions or the camera or microphone of course having different input modes always offers flexibility to human-computer interaction to come up with the best results.

(Refer Slide Time: 02:20)

Characteristics of multimodal systems

Robustness:

- User input for multimodal systems (not multimedia systems!) can be recognized better than with simpler systems due to
 - synergies in the fusion module (→ better speech recognition with face and gesture recognition improves performance)
 - user choice of expected most robust modality
 - user habit of speaking more simple in multimodal interactions



The second characteristic is robustness. If we provide different input modes or different modalities, this means we have to integrate into the system the different kinds and sources of information. That means we use a fusion module to interpret each signal, each channel combinedly.

Having different sources of information also means that our classifiers can be more robust, because we simply have more information available and if on one modality the system is unsure then we have redundant information, then of course additional sources of information can improve the classifying success.

Also, users typically or naturally choose the best fitting modality, because they learn which works well with the system what not and they cope with different problems of the environment, for example environmental noise.

Also users have the habit of speaking more simple and actually more fluently if they are allowed to use other modalities like gesturing, because this is more natural to them.

(Refer Slide Time: 03:33)

Characteristics of multimodal systems

Naturalness:

- In contrast to Multimedia Systems, users can interact *naturally*
 - "Use" affect, emotions
 - Use accompanying "natural" hand gestures
 - "natural" body language (e.g. body direction) is processed
- Users seem to prefer multimodal interaction.
 - especially, when the cognitive load increases (time pressure, task difficulty)
 - one reason might be the usage of different cognitive resources for separate modalities → Processing of information is faster and less demanding (Wickens, 1999)



Last characteristic which works for multimodal systems is naturalness. So, using multimodal signals is actually a natural thing for humans. We all interact with each other in a multimodal way, that means that we use facial expressions, we use our gaze direction, we use our posture or gestures like pointing gestures, in order to interact with each other.

That means using these naturally, using effect that may be interpreted correctly by the system or using hand gestures which can also be interpreted by the system leads to a more natural interaction which is nice for the users.

Also, users seem to prefer multimodal interaction. Especially if the cognitive load is increased users tend to use multimodal interactions not only with a system but also with each other.

For example, if you observe people talking on the phone where only the acoustic channel is available, you will see that users, humans cannot really refrain from using multimodal signals like facial expressions or gestures. So a multimodal system is actually the more natural way to interact with the system.

Why is that so? So this is not only because we learnt our whole life to do this by interacting with other people, but also this seems to be quite hardwired.

So, I will talk in one of the next videos about the so-called Wicken's Model which says that there are different cognitive resources which are strongly related to the different modalities. That means in order to keep our cognitive load low, we just use different modalities. But I will talk about this a little bit later.

(Refer Slide Time: 05:31)

Characteristics of multimodal systems

Naturalness:

- Human multimodality is coordinated ...
- ... but not necessarily simultaneous (speech & pen, speech & gesture)
- E.g. Oviatt finds simultaneous (about 70%) versus sequential users (30%)
- But: multimodality can have negative consequences
 - modality choice as cognitive task
 - interference of modalities
 - How to build a fusion module? (Sequential users produces time lags between modalities up to some seconds)

Why provide a multimodal system, when this results in higher users' expectations?



Human interaction is coordinated so if we speak through our head movement, through our hand movements, this is all temporally and content wise and semantically coordinated. But it is not naturally in the same time, or temporally simultaneous.

That means, for example, Sharon Oviatt, the real pioneer in multimodal interaction, she found out that there are two kinds, or general two kinds of user groups. So, there are users which use multimodal inputs to the system in a sequential way and there are those users who do that in parallel, simultaneously.

So this is a nice flexibility for the users to act as they would like to but on the other hand it is a real challenge for the system designer because the fusion model has to cope up with different time windows to integrate these information which of course belong together and have to be classified and interpreted combinedly. So, this is one of the negatives consequences multimodal systems can have.

Another one is choosing the right combination of modalities or the right modality itself maybe already a cognitive task especially if you do not know the system that well. If you are

not fully or naturally interacting with the system then it might be difficult to actually decide which modality to use. And there might be interferences of different modalities.

So, a provocative question would be why to actually start to build multimodal systems when this actually result in higher expectations and maybe in less good quality for the user if we provide actually this kind of multimodal input instead of much simpler systems.

(Refer Slide Time: 07:32)

Characteristics of multimodal systems
Embodied systems

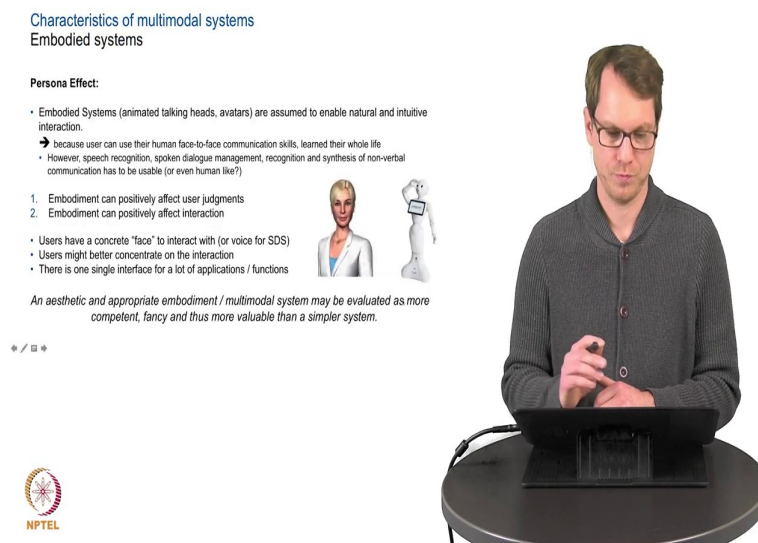
Persona Effect:

- Embodied Systems (animated talking heads, avatars) are assumed to enable natural and intuitive interaction.
- ➔ because user can use their human face-to-face communication skills, learned their whole life
- However, speech recognition, spoken dialogue management, recognition and synthesis of non-verbal communication has to be usable (or even human like?)

1. Embodiment can positively affect user judgments
2. Embodiment can positively affect interaction

- Users have a concrete "face" to interact with (or voice for SDS)
- Users might better concentrate on the interaction
- There is one single interface for a lot of applications / functions

An aesthetic and appropriate embodiment / multimodal system may be evaluated as more competent, fancy and thus more valuable than a simpler system.



The next four characteristics hold only for embodied systems and with embodied systems I am talking about human-like system interfaces. This could be like either real robots, social robots for example which look like human or nearly like human, or these could be like virtual humans on the screen.

Here's the so-called Persona Effect. These are all social effects that I am now presenting, all four. The first one is called as the Persona Effect and actually just states that if users interact with an embodied system which has the same capabilities as a non-embodied system then the users tend to like the system which is embodied more.

The second result or the second effect of this Persona Effect is that the actual interaction can be more efficient or more effective. This means less errors are done and the whole task maybe solved a little bit quicker.

This does not hold true for every kind of situation or every kind of task but these two aspects of the Persona Effect, the better perceived quality and a more efficient and a better interaction, has been observed several times.

There are several reasons why this might be the case. One reason is that we actually have one concrete face or voice in the case of a non-visual interface the user can concentrate on.

Also having just one single interface, for example for an intelligent environment might be helpful and supportive for this Persona Effect because it is easier to concentrate and deal with just one interface than to talk to different devices or talk just to the air if you want to interact and control a virtual environment.

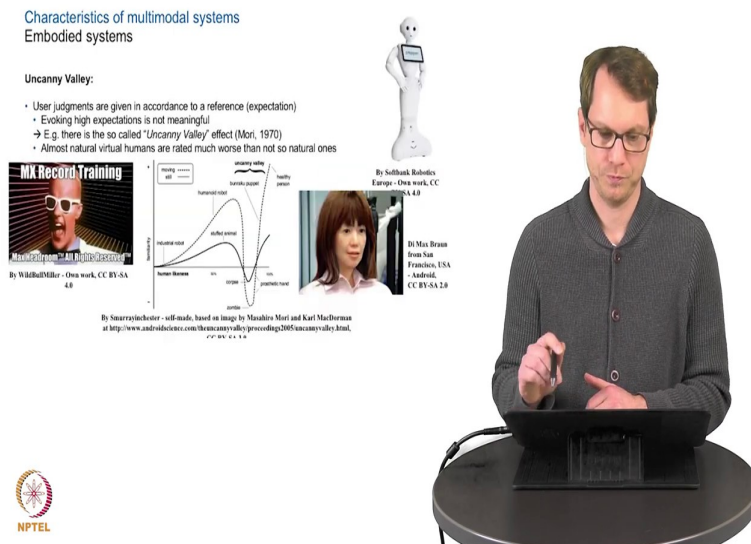
With embodied interfaces it is important to notice is that these have, especially because of the human likeliness, some aesthetic dimension. This means an aesthetic and appropriate embodiment in a multimodal system may be evaluated as more competent and even more fancy and more valuable than comparable systems which are not embodied but actually provide the same capabilities or functions.

(Refer Slide Time: 10:12)

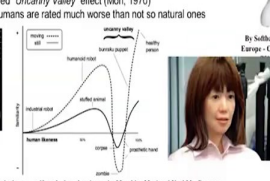
Characteristics of multimodal systems
Embodied systems

Uncanny Valley:

- User judgments are given in accordance to a reference (expectation)
- Evoking high expectations is not meaningful
→ E.g. there is the so called "Uncanny Valley" effect (Mori, 1970)
- Almost natural virtual humans are rated much worse than not so natural ones




MK Record Training
Master Program of All Online Courses
By WildBallMiller - Own work, CC BY-SA 4.0



By SmerreIncheator - self-made, based on image by Masahiro Mori and Karl MacDorman at <http://www.andriod.com/forums/viewtopic.php?p=105>

By Softbank Robotics Europe - Own work, CC BY-SA 4.0

By Max Braun from San Francisco, USA - Android, CC BY-SA 2.0



The second effect for embodied interfaces is the so-called “Uncanny Valley” effect. You must have heard of this already. You can see in the graphic down on the slides that, what the actual Uncanny Valley means.

So, what we have here is the familiarity which means high value of familiarity means it is a pleasant interface whereas a low value in this means it is a little bit awkward or even repulsive.

And if we go more and more human-like, in our interface, for example from an industrial robot to a human robot, the pleasantness, the familiarity will increase until this so-called Uncanny Valley where suddenly the familiarity of the positive affiliation drops significantly.

The reason for this might be just the shift in reference. Suddenly we do not compare the interface any more to like normal robots, or computer interfaces but to real humans and if this virtual human or this social humanoid robot is not perfect then it may be really awkward.

In order to prevent this kind of Uncanny Valley we can either just, stop right at the beginning of this Uncanny valley and you can see the white Pepper robot is one kind of an attempt to do this, but also animals or cartoon characters are a good way in order to prevent the Uncanny Valley. Or, we can just skip this Uncanny Valley and really try to make really human-like interface.

(Refer Slide Time: 12:00)

Characteristics of multimodal systems
Embodied systems

Social Inhibition and facilitation:

- In the presence of other:
 - Task performance increases for easy tasks
 - But decreases for complex tasks
- Most likely due to increased attention, e.g. faster decisions



The third Effect is the so-called Social Inhibition and Facilitation Effect and these are actually two effects which come together. This is actually not an effect known from or invented or observed from human and computer interaction but we know this from social interactions between humans.

And it means in the company, the presence of other people we tend to be better at simple tasks, but we tend to be worse in more complex tasks.

And this is assumed to be a concentration problem because we might be distracted by the social situation, have a higher cognitive load and therefore we tend to act quicker which is always nice for simple task because we know what we are doing, but for more complex tasks which require more concentration and more resources, this may then lead to actually worse results.

But we can transfer this social phenomenon, this effect to social aspects of the human-computer or human-robot interaction. And this means in for some kinds of embodied systems, like social robots or virtual people, virtual humans we may observe this social facilitations and inhibition effect.

Actually, it's a nice example to see or a method to actually evaluate whether certain computer interface has some social presence.

(Refer Slide Time: 13:43)

Characteristics of multimodal systems
Embodied systems

Equilibrium Theory (Forgas, 1999):

- Equal level of intimacy
- Balanced on modalities
- E.g.: closer distance, less gaze

◀ / □ ▶



The same holds for the last effect I want to present today. This is the Equilibrium theory and it is also observed for social human interaction. The whole idea is that we have this signal of intimacy, of the level of intimacy, this means how is the relationship between people.

If we are more intimate, for example we share more gazes we tend to be more close to each other and so on. The theory says that we balance the signals on the different modalities in such a way that they all are fitting to the relationship and level of intimacy that we have.

So for example, if we are more strangers, we tend to be more far away from each other and do not disturb the personal space of the other person. But consider the example of an elevator where it is really crowded.

You cannot avoid to disturb or come into the personal space of another person. In order to compensate for this, because the social signal of the space difference and distance is not appropriate, we tend to balance, our social signals on other modalities, for example, our gaze. We tend to avoid our gaze in order to compensate for the inappropriate distance that we have.

Again, using this kind of theory or this kind of social effect that we can observe, we can actually try to evaluate whether the embodied system like in a virtual environment, is working in such a way that it provokes some social presence.

Another example, instead of the elevator, would be people, even virtual people, in the virtual space, would go behind my back, which usually people tend to find really uncomfortable.

So these social effects which we know from social psychology are a nice way to test whether your interface or your virtual environment is working in such a way that it should be mainly producing virtual characters or virtual people or social robots that are actually perceived as having social presence.

(Refer Slide Time: 15:58)

Multimedia & Multimodality References

- R. A. Bolt, "Put-that-there": Voice and gesture at the graphics interface", Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques 14(3), 262—270, 1980.
- C. Benoit, J. Martin, C. Pelachaud, L. Schomaker, B. Suhm: "Audio-visual and Multimodal Speech Systems", in D. Gibbon, I. Mertins, R.K. Moore (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Kluwer, 2000.
- J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. Young, "Four easy pieces for assessing the usability of multimodal interaction: The CARE properties", in *Proc. Interact.*, K. Nordby, P. Helmersen, D. Gilmore, and S. Arnesen, (Eds), London: Chapman & Hall, 115—120, 1995.
- M. Dohen, "Speech through the Ear, the Eye, the Mouth and the Hand", in *Multimodal Signals: Cognitive and Algorithmic Issues: COST Action 2102, LNCS 5398*, Berlin: Springer, 24—39, 2009.
- B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal Interfaces: A survey of principles, Models and Frameworks", in *Human Machine Interaction, LNCS 5440*, Berlin: Springer, 3—26, 2009.
- M. Mori, "The uncanny valley (in Japanese)", in *Energy 7 (4)*, 33—35, 1970.
- L. Nigay, J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion", in *Proc Interact & CHI*, 172—78, 1993.
- S. Oviatt, "Multimodal Interfaces", in *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears (Eds.), Mahwah: Lawrence Erlbaum and Associates, 405—429, 2012.
- Oviatt, S., "Ten myths of multimodal interaction", *Communications of the ACM* 42(11), 74—81, 1999.
- C. Wickens, "Multiple resources and performance prediction", *Theoretical Issues in Ergonomics Science* 3, 159—177, 1999.



So, this was the first week about multimodality and multimedia and here are some references.

(Refer Slide Time: 16:05)

Multimedia & Multimodality References

- R. A. Bolt, "Put-that-there": Voice and gesture at the graphics interface", Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques 14(3), 262—270, 1980.
- C. Benoit, J. Martin, C. Pelachaud, L. Schomaker, B. Suhm: "Audio-visual and Multimodal Speech Systems", in D. Gibbon, I. Mertins, R.K. Moore (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Kluwer, 2000.
- J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, and R. Young, "Four easy pieces for assessing the usability of multimodal interaction: The CARE properties", in *Proc. Interact.*, K. Nordby, P. Helmersen, D. Gilmore, and S. Arnesen, (Eds), London: Chapman & Hall, 115—120, 1995.
- M. Dohen, "Speech through the Ear, the Eye, the Mouth and the Hand", in *Multimodal Signals: Cognitive and Algorithmic Issues: COST Action 2102, LNCS 5398*, Berlin: Springer, 24—39, 2009.
- B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal Interfaces: A survey of principles, Models and Frameworks", in *Human Machine Interaction, LNCS 5440*, Berlin: Springer, 3—26, 2009.
- M. Mori, "The uncanny valley (in Japanese)", in *Energy 7 (4)*, 33—35, 1970.
- L. Nigay, J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion", in *Proc Interact & CHI*, 172—78, 1993.
- S. Oviatt, "Multimodal Interfaces", in *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears (Eds.), Mahwah: Lawrence Erlbaum and Associates, 405—429, 2012.
- Oviatt, S., "Ten myths of multimodal interaction", *Communications of the ACM* 42(11), 74—81, 1999.
- C. Wickens, "Multiple resources and performance prediction", *Theoretical Issues in Ergonomics Science* 3, 159—177, 1999.

