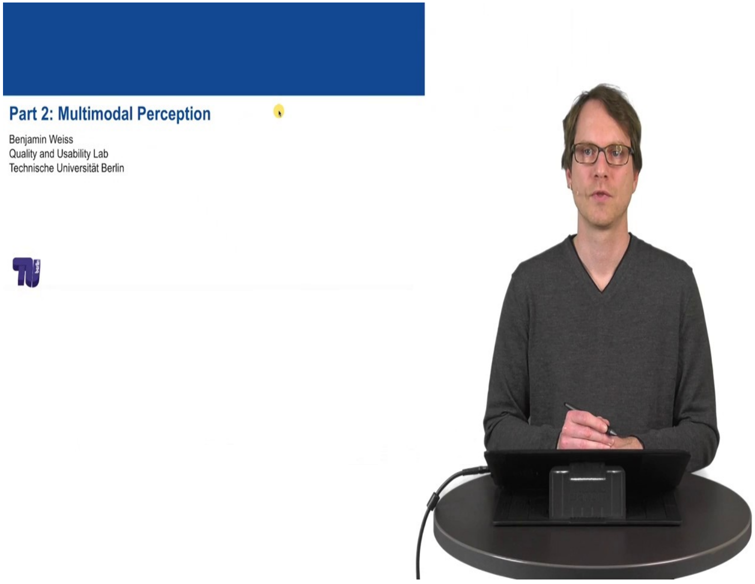
Multimodal Perception
Professor Benjamin Weiss
Quality and Usability Lab
Technische Universitat Berlin
Part 2

(Refer Slide Time: 00:17)

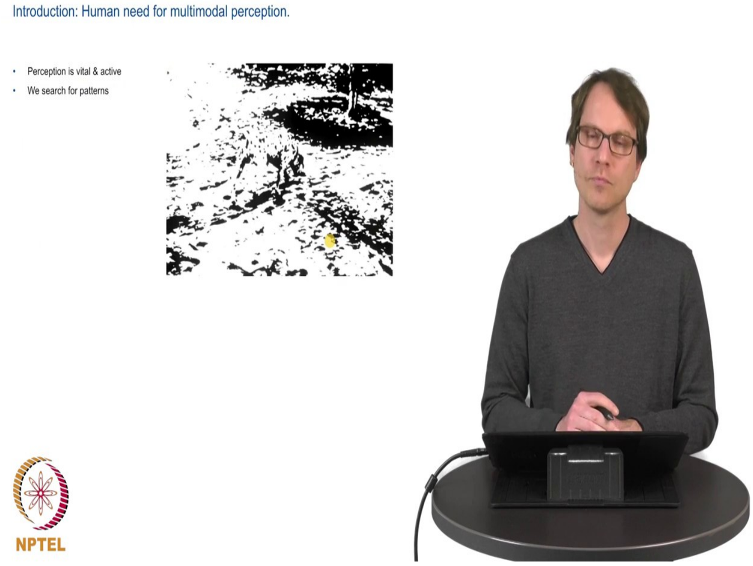


Part 2: Multimodal Perception

Benjamin Weiss
Quality and Usability Lab
Technische Universität Berlin

This week's topic is multimodal perception.

(Refer Slide Time: 00:23)



Introduction: Human need for multimodal perception.

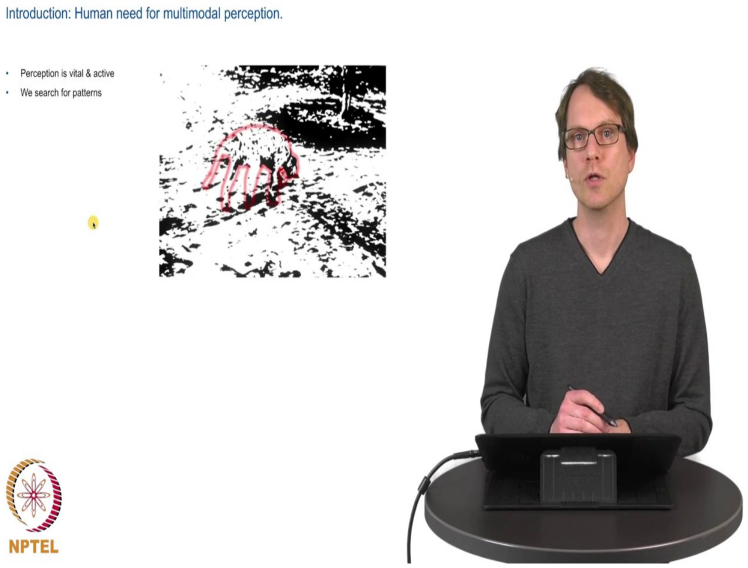
- Perception is vital & active
- We search for patterns

So humans rely on multimodal input. Just because we want to deal with the world, that means we want to encounter risks and to see which kind of objects or locations or events are actually happening.

And as these objects, events usually produce a lot of signals we rely on any kind of sensory input that we might encounter. So this is an active process and we actively search for patterns in this data in order to find out which kind of object or location or event is happening right now.

Have a look at this graphic. There is a lot of noise in there but you might, might find the actual pattern that we have hidden here. It is a dog. Have a look here. But maybe you will need some more help.

(Refer Slide Time: 01:12)

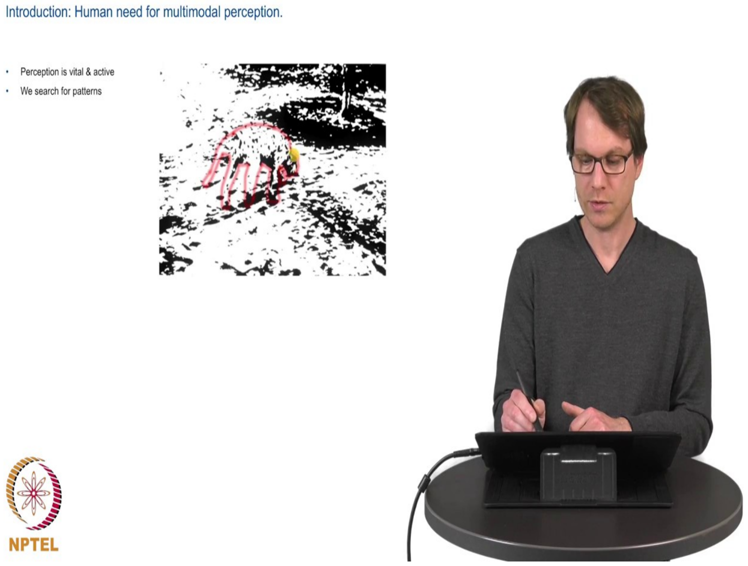


Introduction: Human need for multimodal perception.

- Perception is vital & active
- We search for patterns

So if you see these red borders there is head

(Refer Slide Time: 01:16)

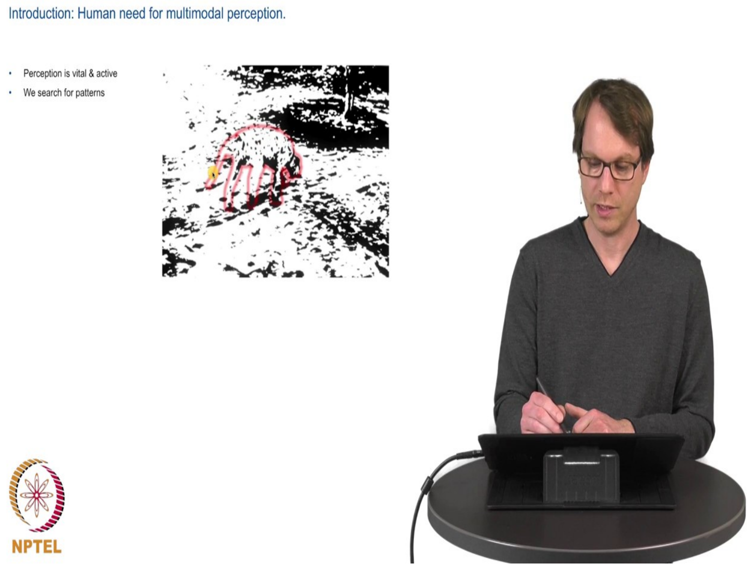


Introduction: Human need for multimodal perception.

- Perception is vital & active
- We search for patterns

of the dog and the tail.

(Refer Slide Time: 01:18)

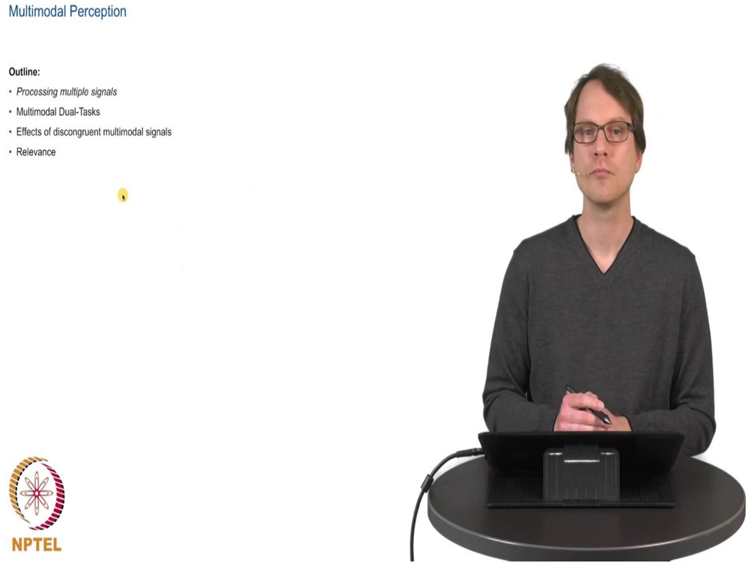


Introduction: Human need for multimodal perception.

- Perception is vital & active
- We search for patterns

So as you can see we actively search for patterns and if we know what we are looking for, it gets more easy for us.

(Refer Slide Time: 01:27)



Multimodal Perception

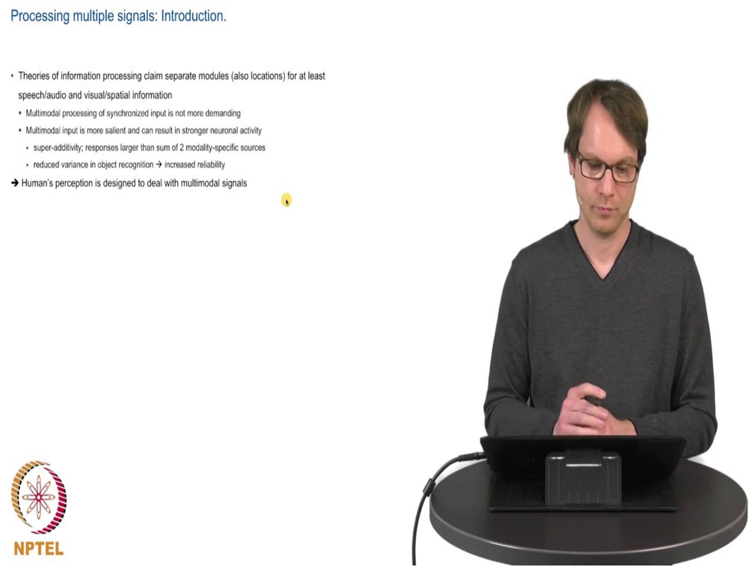
Outline:

- *Processing multiple signals*
- Multimodal Dual-Tasks
- Effects of discongruent multimodal signals
- Relevance

This is the topic of today and I have 4 subtopics. I will start with explaining the processing basics of multimodal signals.

Then explain a little bit on the two dual tasks and how we proceed and process multimodal input for dual tasks. I will then explain how effects of discongruency, of discongruent information we deal with. And I will then complete with relevance for HCI.

(Refer Slide Time: 01:59)



Processing multiple signals: Introduction.

- 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 can result in stronger neuronal activity
    - super-additivity; responses larger than sum of 2 modality-specific sources
    - reduced variance in object recognition → increased reliability

→ Human's perception is designed to deal with multimodal signals

So cognitive theories claim that we have several different processing modules and resources that we can use. And at least for the audio, speech-related information and spatial-visual information we know that these can actually be processed quite reliable and efficient in parallel.

So as I already told in the first week, this kind of processing of multimodal input from the same source is not more demanding than the unimodal processing. Actually if we encounter more information or, from, information from several channels, this input can be more reliable and more salient for us.

So this means it helps us to focus our attention to it and process this kind of information more easily, for example to recognize certain objects. I will talk about the super-additivity aspect in a second. The main result is that there is reduced variance due to this increased neural response that we have. And this helps us to actually identify objects more easily.

As a conclusion, human perception is really designed for dealing with multimodal input signals.

(Refer Slide Time: 03:32)

Processing multiple signals: Introduction.

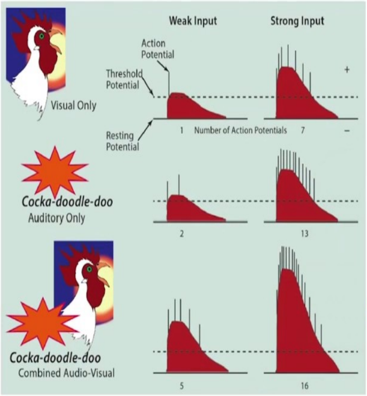


Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multisensory enhancement, which facilitates adaptive responding in humans and animals.

Oviatt & Cohen, 2015; p. 31

What you can see here is the humans neural activity as response to an animal stimulus.

Here we have a rooster’s cry shouting the Cocka-doodle-doo, once in visual-only stimulation on a display, once only the auditory information and then in the last row it is the audio-visual, the multimodal stimulation.

And you can see the weak input

(Refer Slide Time: 03:57)

Processing multiple signals: Introduction.

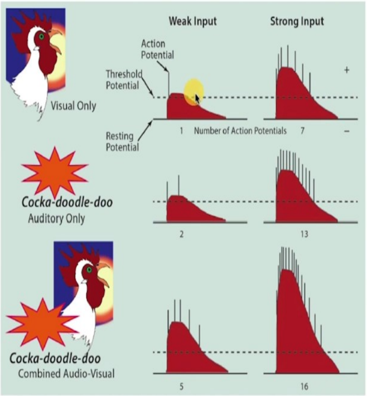
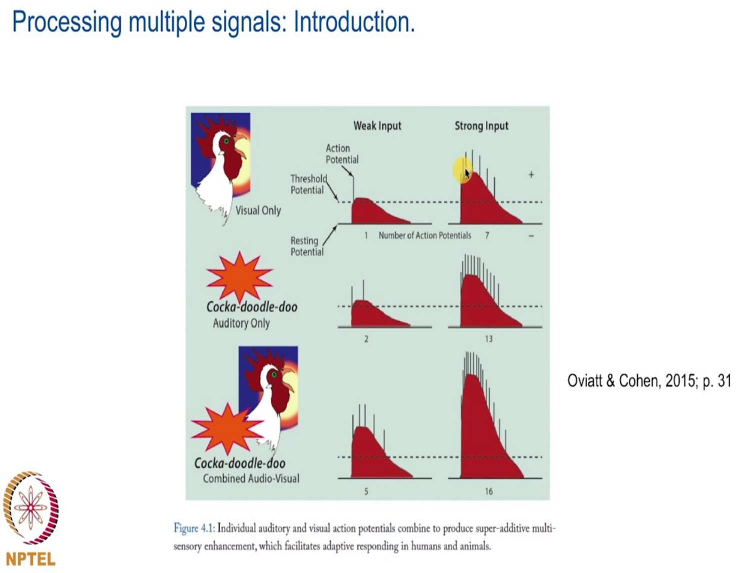


Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multisensory enhancement, which facilitates adaptive responding in humans and animals.

Oviatt & Cohen, 2015; p. 31

if it is not that strong in intensity resulting in just reaching the threshold without any neural spikes absent from these kind of cells. And

(Refer Slide Time: 04:09)



Processing multiple signals: Introduction.

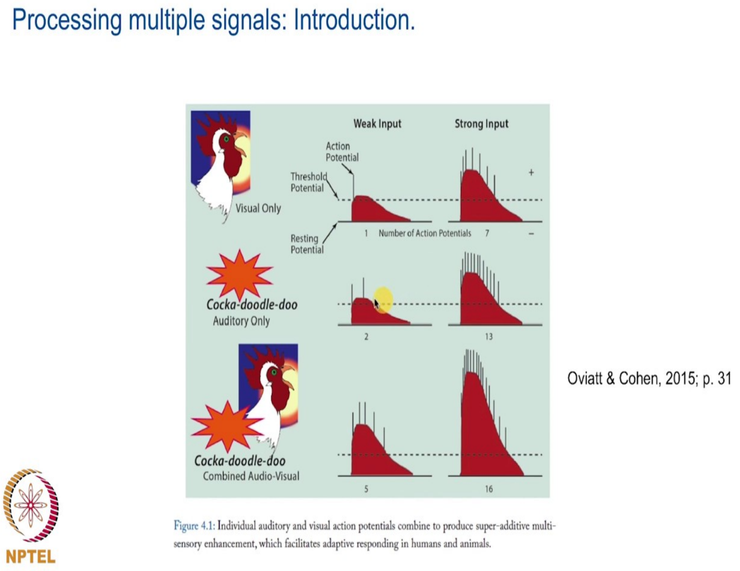
Oviatt & Cohen, 2015; p. 31

Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multi-sensory enhancement, which facilitates adaptive responding in humans and animals.

with strong impulse of strong stimulation we have these spikes here, the neural spikes which are triggered.

The same we have for the

(Refer Slide Time: 04:19)



Processing multiple signals: Introduction.

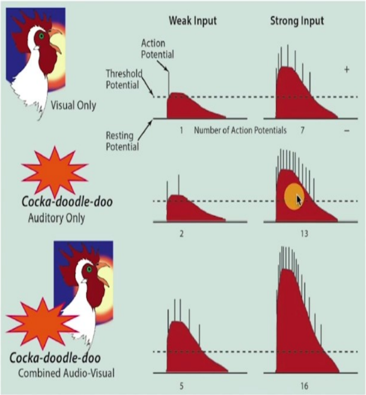
Oviatt & Cohen, 2015; p. 31

Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multi-sensory enhancement, which facilitates adaptive responding in humans and animals.

auditory modality, so also here there is a weak input with just a few spikes

(Refer Slide Time: 04:25)

Processing multiple signals: Introduction.



Oviatt & Cohen, 2015; p. 31

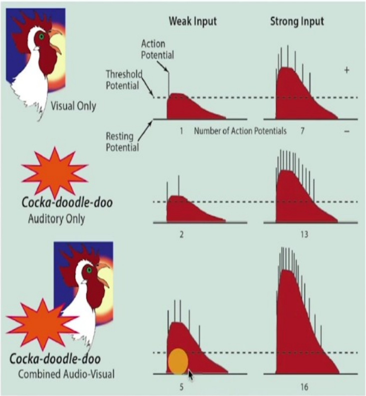
Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multisensory enhancement, which facilitates adaptive responding in humans and animals.

and a strong input with a lot of spikes.

And if we now compare the multimodal response

(Refer Slide Time: 04:31)

Processing multiple signals: Introduction.

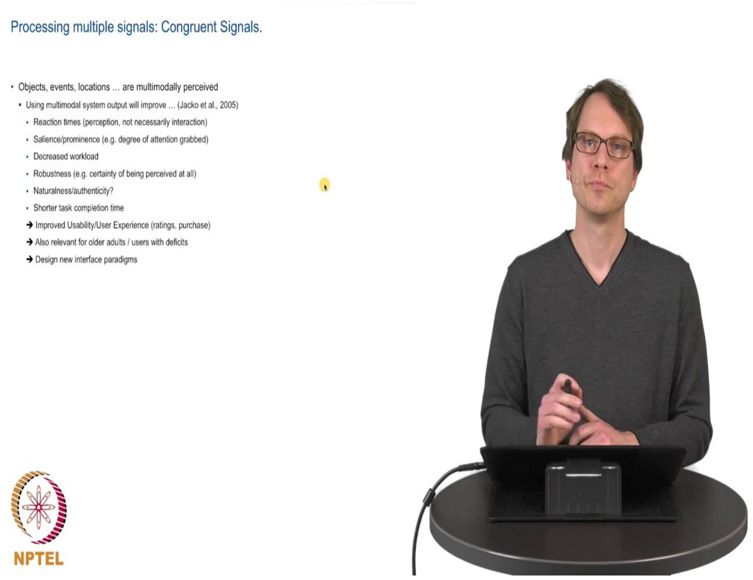


Oviatt & Cohen, 2015; p. 31

Figure 4.1: Individual auditory and visual action potentials combine to produce super-additive multisensory enhancement, which facilitates adaptive responding in humans and animals.

of the humans compared to the addition to the combined input of these two, we see that there are actually cases in which the multimodal input stimulates the human stronger than the sum of the single individual inputs. This is called super-additivity.

(Refer Slide Time: 04:54)



Processing multiple signals: Congruent Signals.

- Objects, events, locations ... are multimodally perceived
  - Using multimodal system output will improve ... (Jacko et al., 2005)
    - Reaction times (perception, not necessarily interaction)
    - Salience/prominence (e.g. degree of attention grabbed)
    - Decreased workload
    - Robustness (e.g. certainty of being perceived at all)
    - Naturalness/authenticity?
    - Shorter task completion time
    - ➔ Improved Usability/User Experience (ratings, purchase)
    - ➔ Also relevant for older adults / users with deficits
    - ➔ Design new interface paradigms

So as we are perceiving objects, locations and events in a multimodal way, there are a lot of relevance here for multimodal in HCI. And according to Jacko et al, we have multiple benefits.

If we produce multimodal output and be aware according to the first lecture of the first week that we had, that is this is actually multimedia output. So there are decreased reaction times of the users. There is increased salience and prominence.

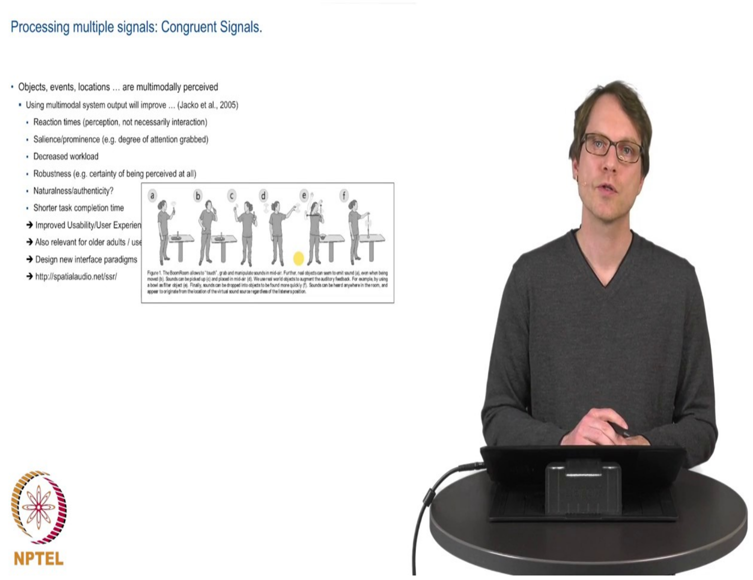
This means that user can more easily identify objects and also the attention of the users can be grabbed more easily if you provide multimedia output.

There is decreased workload, more robustness as I said, and may be even more naturalness or authenticity if you really want to synthesize certain experiences. There is also shorter completion time of certain tasks observed.

So as a conclusion there might be improved usability and user experience by multimedia system output and this actually means multimedia perception of users. It is also especially relevant for certain target groups such as older adults or people with some deficits.

And it is also used to design new interface paradigms.

(Refer Slide Time: 06:27)



Processing multiple signals: Congruent Signals.

- Objects, events, locations … are multimodally perceived
  - Using multimodal system output will improve … (Jacko et al., 2005)
    - Reaction times (perception, not necessarily interaction)
    - Salience/prominence (e.g. degree of attention grabbed)
    - Decreased workload
    - Robustness (e.g. certainty of being perceived at all)
    - Naturalness/authenticity?
    - Shorter task completion time
    - ➔ Improved Usability/User Experien
    - ➔ Also relevant for older adults / use
    - ➔ Design new interface paradigms
    - ➔ http://spatialaudio.net/ssr/

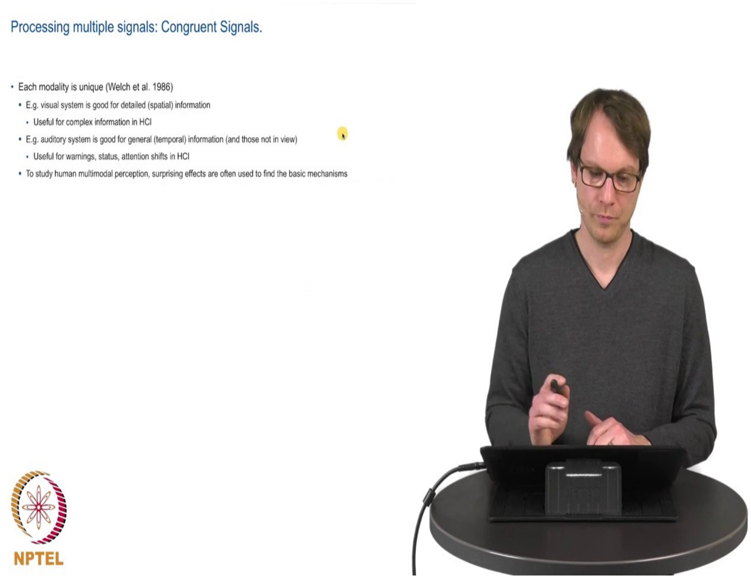
Figure 1. The BoomRoom allows to "touch", grab and manipulate sounds in mid-air. Further, real objects can seem to emit sound (a), even when being moved (b). Sounds can be picked up (c) and placed in mid-air (d). We use real world objects to augment the auditory feedback. For example, by using a bowl as filter object (e). Finally, sounds can be dropped into objects to be found more quickly (f). Sounds can be heard anywhere in the room, and appear to originate from the location of the virtual sound source regardless of the listeners position.

For example have a look at the website of our colleagues here. They developed a multimodal system in certain virtual environment. This means they have a 3D audio synthesis system and also a 3D virtual environment with the visual information.

And what they basically do is they not only use a certain 3D display in order to synthesize or create virtual objects but they try to find out whether it is better, improves the user experience if we also have located three dimensional sound sources at the same spot and the same location in space as the visual information of a certain object.

And you can create these kind of virtual objects in an auditory and visual way in order to interact with the system. You can also enhance the real object that you use as a tangible interface with the system by virtual sound sources. And they built up the system and they found out that indeed, the user experience is better than without the auditory information.

(Refer Slide Time: 07:42)



Processing multiple signals: Congruent Signals.

- Each modality is unique (Welch et al. 1986)
  - E.g. visual system is good for detailed (spatial) information
    - Useful for complex information in HCI
  - E.g. auditory system is good for general (temporal) information (and those not in view)
    - Useful for warnings, status, attention shifts in HCI
  - To study human multimodal perception, surprising effects are often used to find the basic mechanisms

So each modality is special at least to us humans. This means that for example, that the visual information are specially good for spatial information and also for details. So this means we want to use them for complex information and spatial information.

And the auditory system in us humans is specially detailed and good for temporal information. And of course we can hear from everywhere, at least if the sound sources reaches our ears. So this is specially good for using it for warnings, grabbing attention and so on.

In order to study human perception especially the way the different signals are integrated and processed in the human, we typically also rely on fake data, which means that we try to synthesize information on different channels say the auditory channel and the visual channel and they are not congruent. They do not match that good.

But I will talk about this in the next one, in the next session.