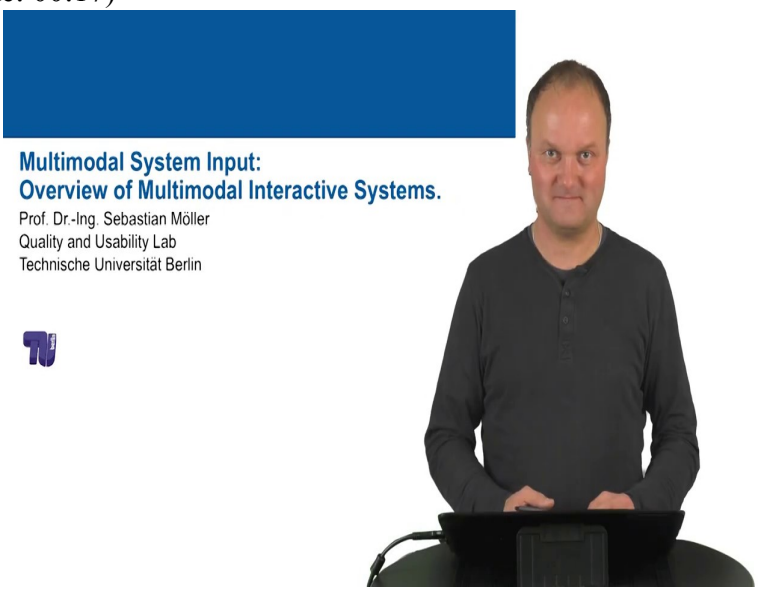
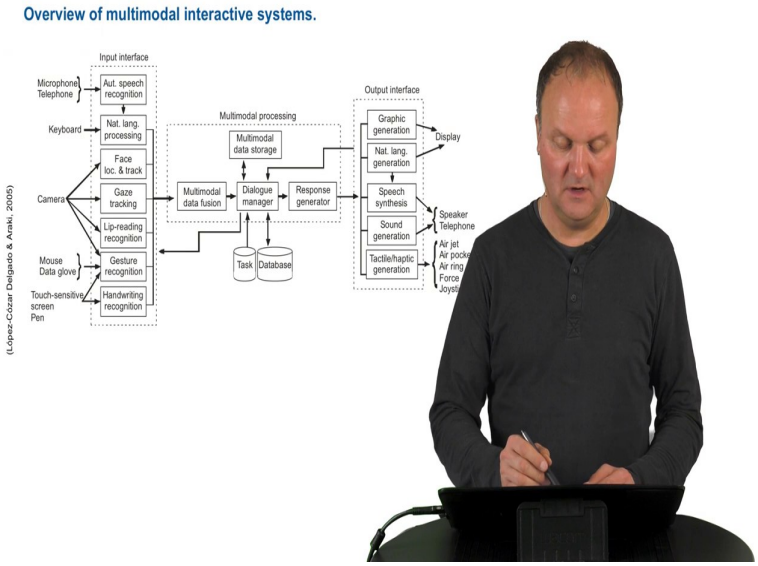
Multimodal System Input
Professor Doctor Sebastian Moller
Quality and Usability Lab
Technische Universitat Berlin
Overview of Multimodal Interactive Systems

(Refer Slide Time: 00:17)



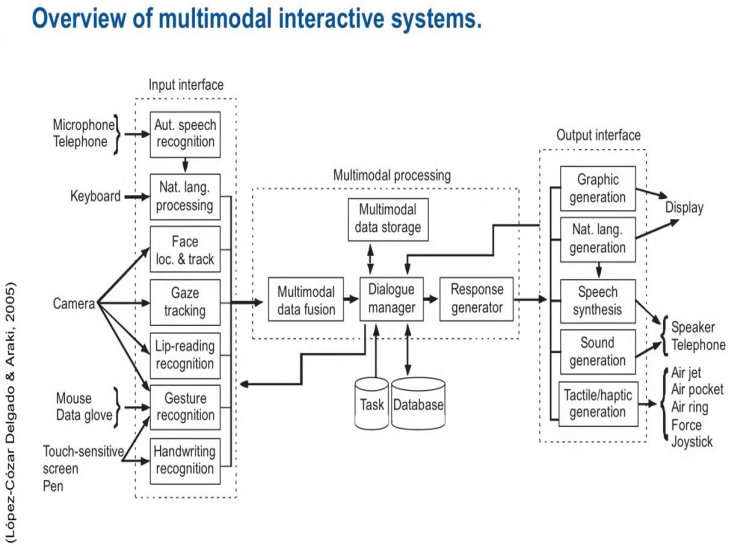
In this film I would present you with an overview of a multimodal interactive system and what we consider to be multimodal system that is what our modality is, what are interaction modalities and how they can be categorized? Let us start with an overview

(Refer Slide Time: 00:37)



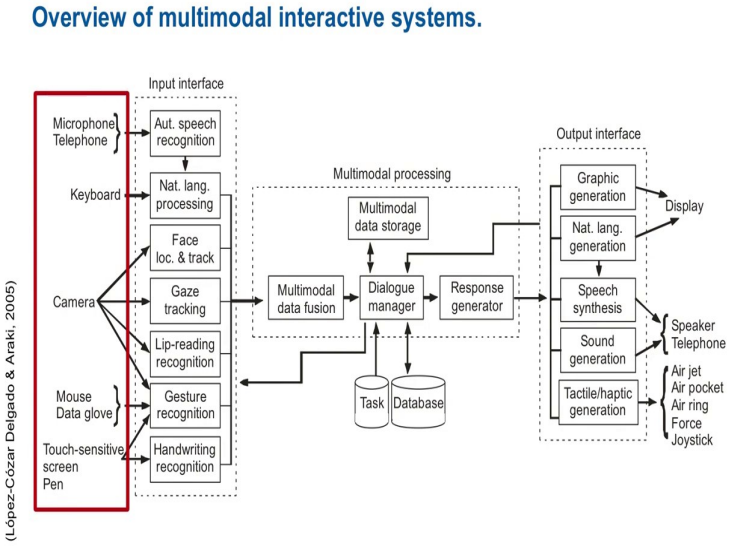
of such a system.

(Refer Slide Time: 00:39)



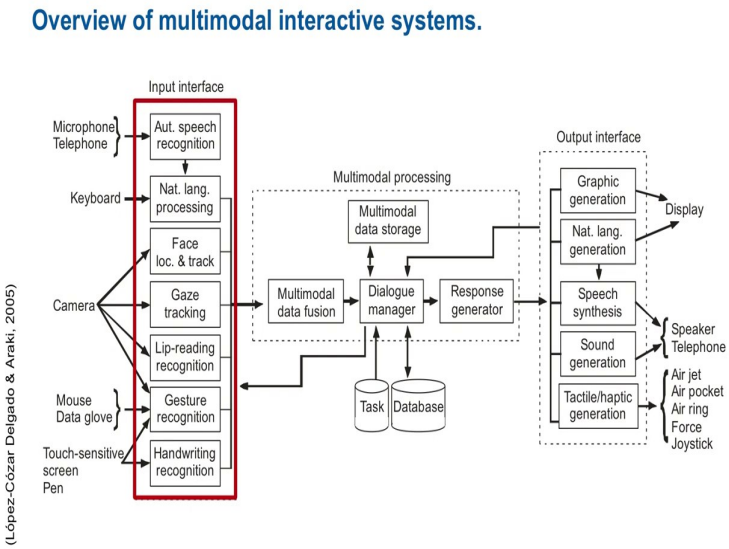
On the left hand side you see different devices which can

(Refer Slide Time: 00:43)



be used to access information through different physical channels and to serve as an input to the input interface. The input interface

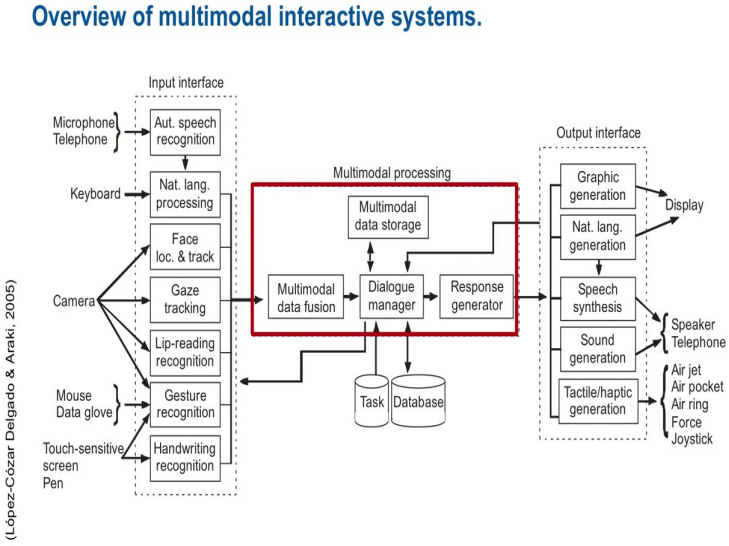
(Refer Slide Time: 00:53)



is actually the mechanism behind these devices which makes use of that information and translates it into semantically meaningful pieces of information which can be processed by an interactive system.

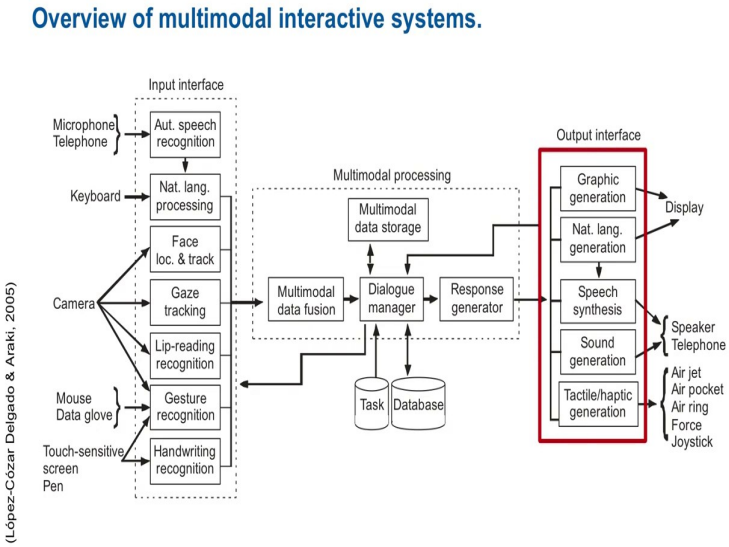
In the middle of the picture, you see the multimodal processing units

(Refer Slide Time: 01:09)



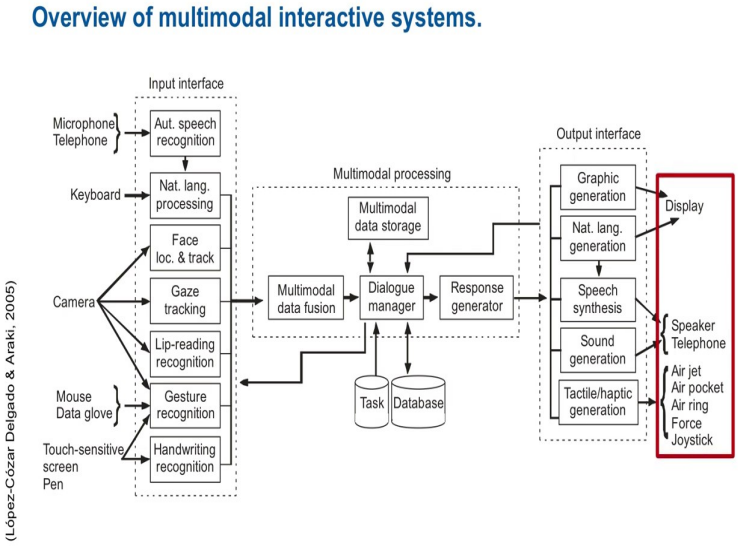
and on the right hand side you see different output interfaces. They are

(Refer Slide Time: 01:14)



again being connected to different devices giving access to the physical

(Refer Slide Time: 01:18)

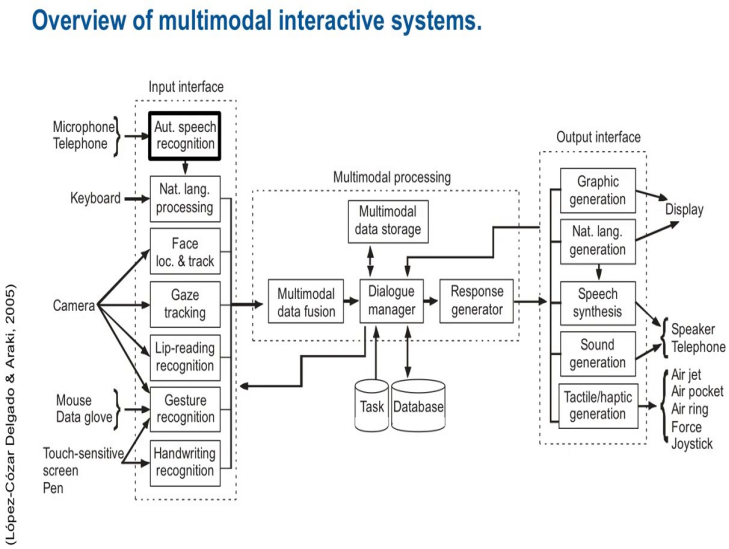


channel. And on the other side of this physical channel you have to imagine the human user of such a multimodal interactive system.

In this chapter we would like to give you an overview of different input interfaces, interfaces to access information from different physical input channels.

On the top layer here you see automatic speech recognition

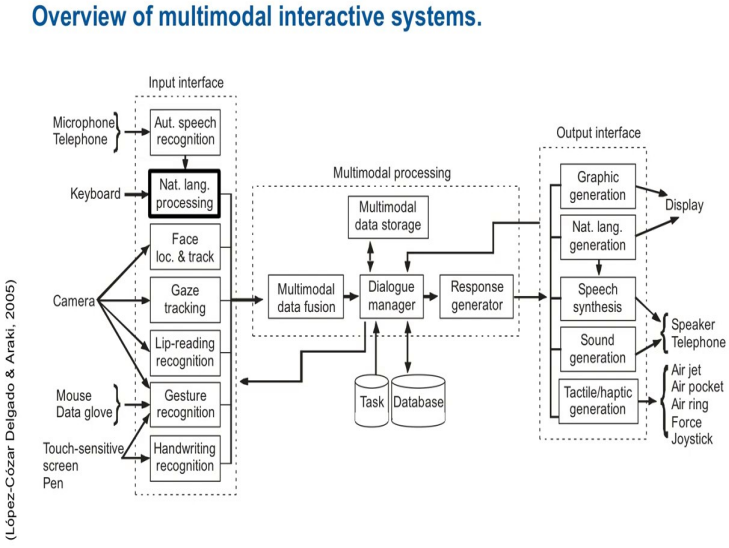
(Refer Slide Time: 01:42)



which makes use of a microphone signal perhaps integrated into telephone and then provides a transcription of that microphone signal into written text.

And this written text needs to be interpreted and this is the task of natural language processing

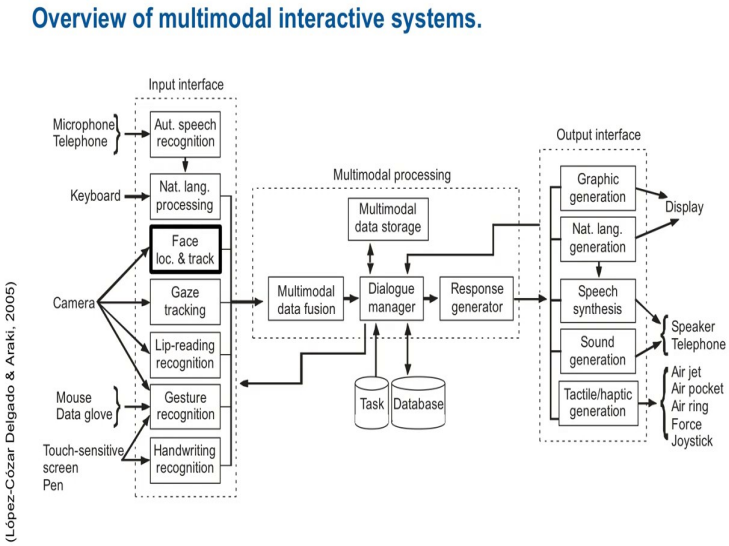
(Refer Slide Time: 01:57)



unit. That natural language processing unit can of course also operate on the basis of written text or typed text put in for example with the keyboard.

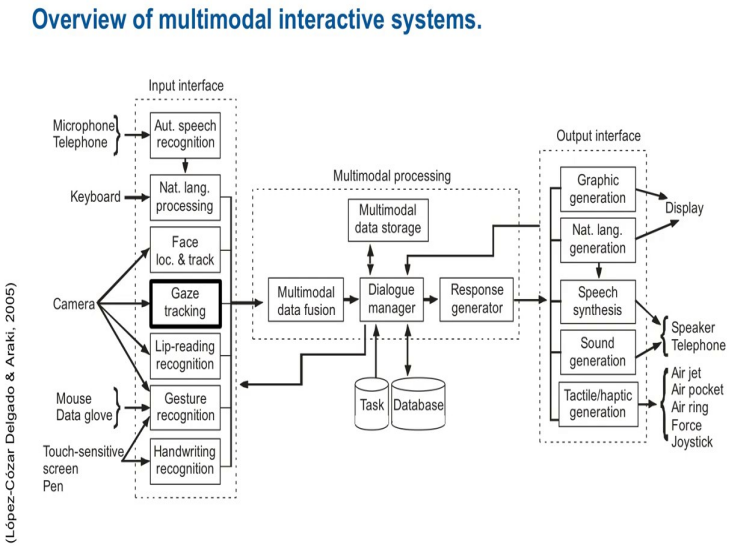
Then you have different modules which can make use of camera information, for example a module for face location and tracking

(Refer Slide Time: 02:16)



of faces. And if you have recognized the face you can also recognize where a person looks to, that is the gaze.

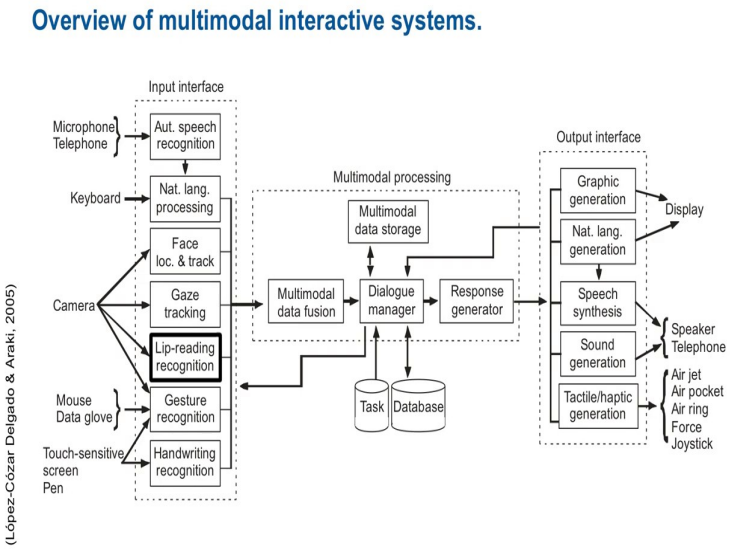
(Refer Slide Time: 02:23)



So there are mechanisms for gaze tracking.

You can even concentrate on the lips and do something which we call lip reading

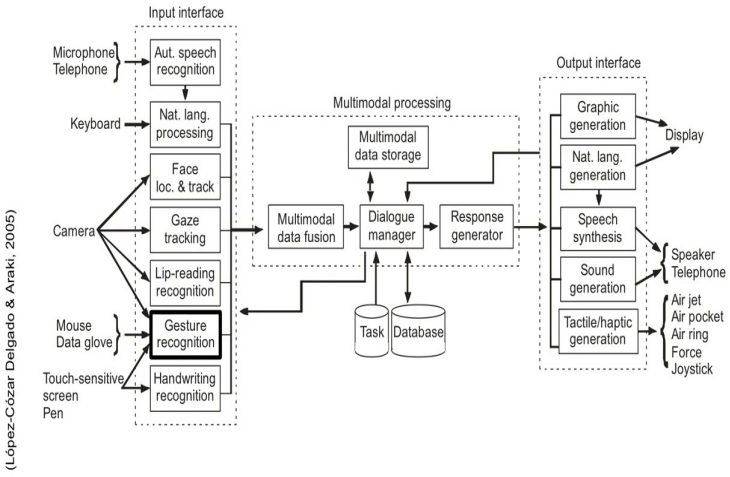
(Refer Slide Time: 02:30)



and the lip reading recognition may help the automatic speech recognition in order to make use of the auditory and the visual information as we will see later in this chapter.

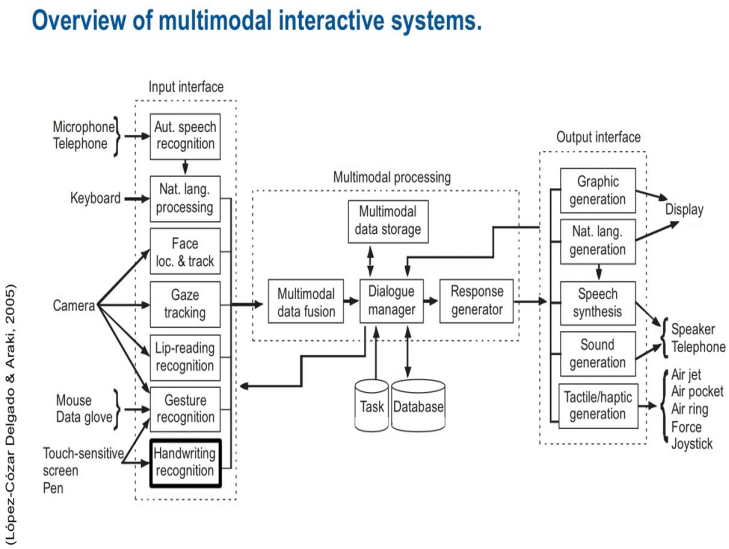
Cameras can also use to record gestures of humans

(Refer Slide Time: 02:44)



but we can also of course use other sensory information, for example coming from the mouse or data glove or touch-sensitive screen or input panel stylus. This can be used to make 3D gestures which are then recognized by the system, 2D gestures on the screen for example and even things like handwriting can be recognized

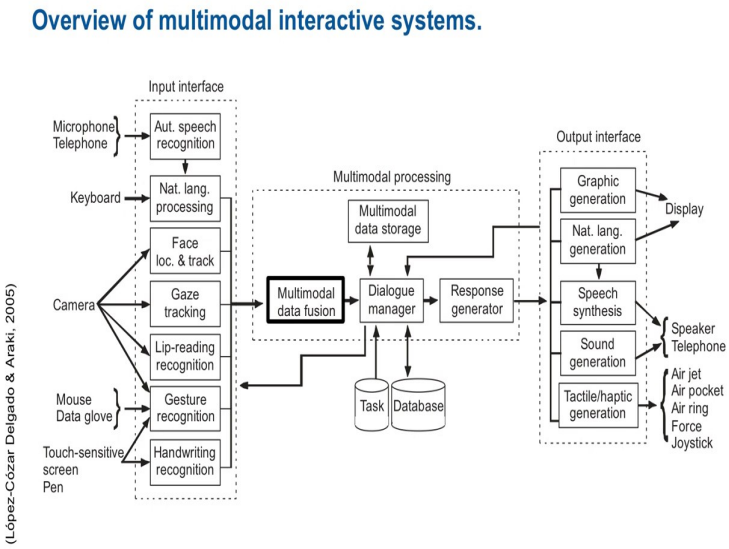
(Refer Slide Time: 03:06)



from that basis.

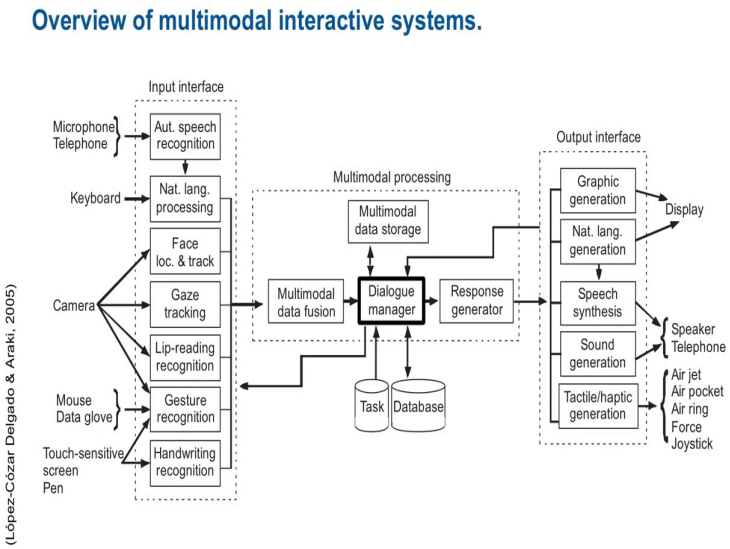
The information kinds of these different channels needs to be fused or merged in order to be usable by the interactive system and this is the task of the multimodal data

(Refer Slide Time: 03:19)



fusion component. This fusion component gives the semantic meaning of the input to a central unit which is called the

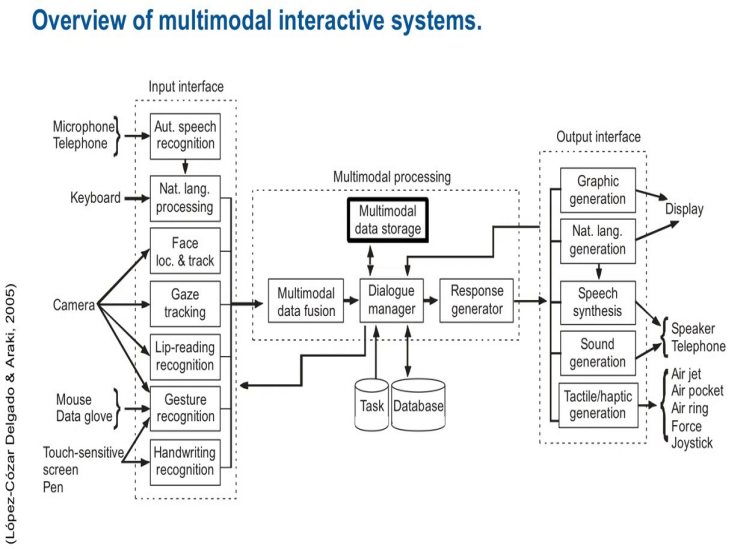
(Refer Slide Time: 03:28)



dialog manager or interaction manager. And this manager is responsible for the behavior of the system towards the user.

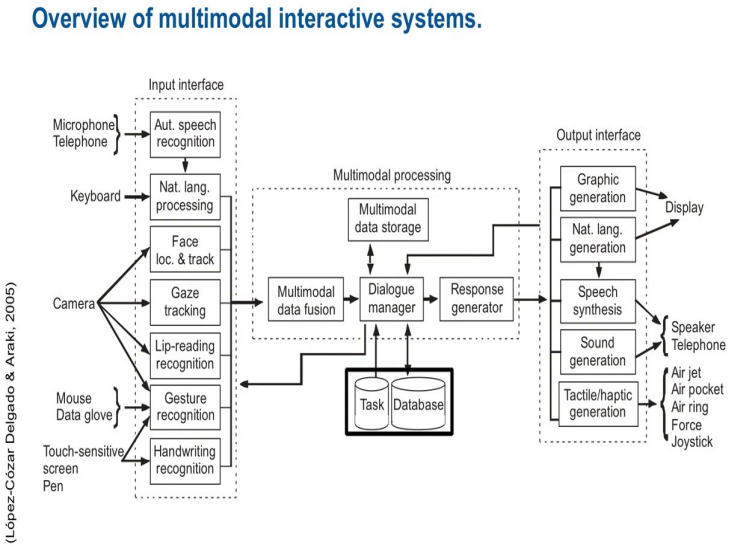
It may make use of a multimodal

(Refer Slide Time: 03:38)



data storage, remembering what type of information has been provided at which time by which modality and of course it should also have access to task

(Refer Slide Time: 03:49)

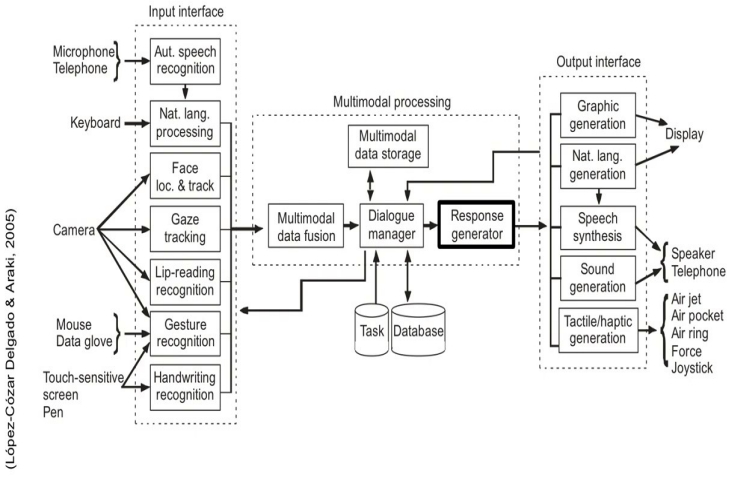


or to databases which help functionality of the system.

So for example it can be a booking system or reservation system or transaction system and this interaction manager or dialog manager needs to have access to it. Then the dialog manager has to provide output information to the user, either requesting new information or generating new interaction behavior or giving the desired response to the user.

And the response needs to be generated first on an abstract level, on a conceptual level and then distributed to the different output interfaces. This process is called the fission of

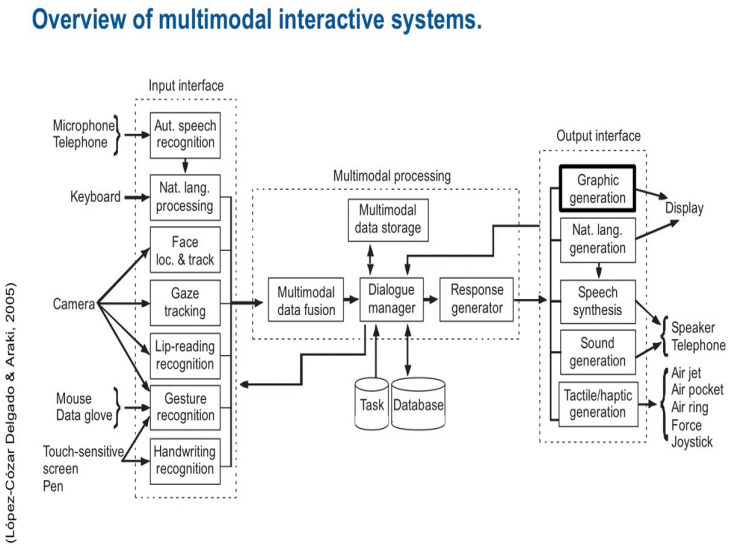
(Refer Slide Time: 04:28)



information, so the opposite of the fusion process.

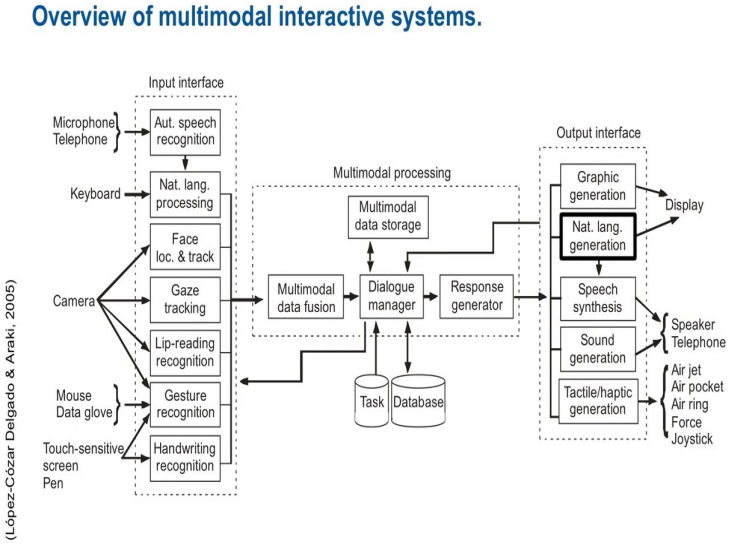
The fission distributes the information to different output interfaces which may be using graphics.

(Refer Slide Time: 04:39)



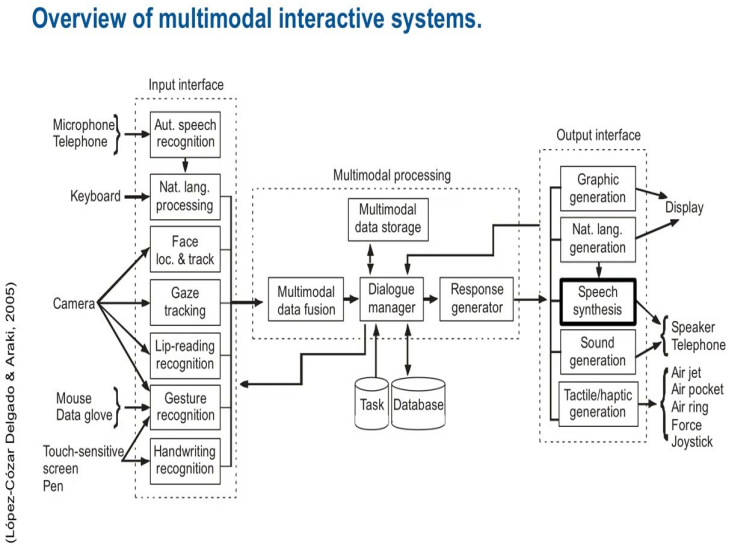
For example you have a graphic component, graphic generation component or it can be natural language

(Refer Slide Time: 04:44)



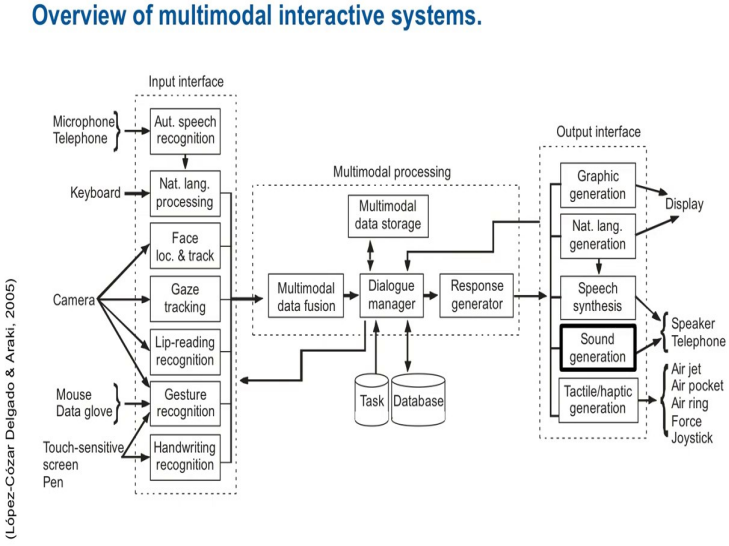
which is then either displayed to the user or transformed into an acoustic signal with the help of speech synthesis.

(Refer Slide Time: 04:51)



You can also generate other audible sounds,

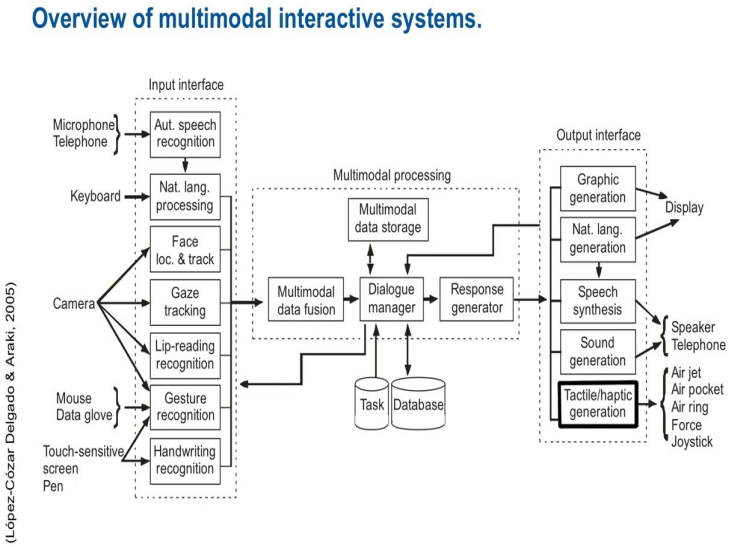
(Refer Slide Time: 04:55)



auditory icons or earcons for example which are then displayed to the user via the loudspeaker.

And you can also generate haptic signals

(Refer Slide Time: 05:05)

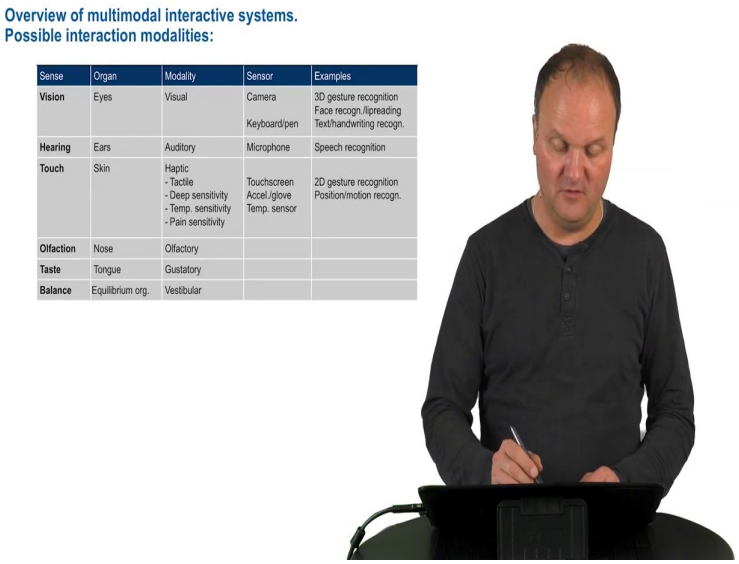


which can be perceived, for example using devices like air jets, air pockets, air rings or force feedback devices like joysticks. All this output information may then be transformed to the user transferred to the user.

The question in such a multimodal interactive system is which modality should we use for which purpose?

And in order to

(Refer Slide Time: 05:29)



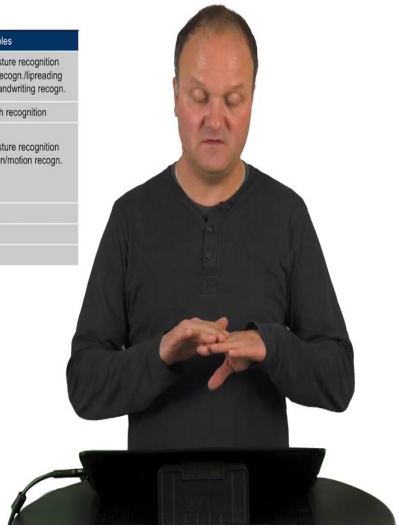
address this point and to provide you with more information on the input modalities I would first like to discuss some terms which are related to the use of different media for interaction and that is actually what we define as a modality.

In psychology the modality is usually defined on a perceptual level; that is we have a visual modality and auditory modality, a haptic modality which consists of the surface

(Refer Slide Time: 05:59)

Overview of multimodal interactive systems.
Possible interaction modalities:

Sense	Organ	Modality	Sensor	Examples
Vision	Eyes	Visual	Camera Keyboard/pen	3D gesture recognition Face recogn./lipreading Text/handwriting recogn.
Hearing	Ears	Auditory	Microphone	Speech recognition
Touch	Skin	Haptic - Tactile - Deep sensitivity - Temp. sensitivity - Pain sensitivity	Touchscreen Accel./glove Temp. sensor	2D gesture recognition Position/motion recogn.
Olfaction	Nose	Olfactory		
Taste	Tongue	Gustatory		
Balance	Equilibrium org.	Vestibular		



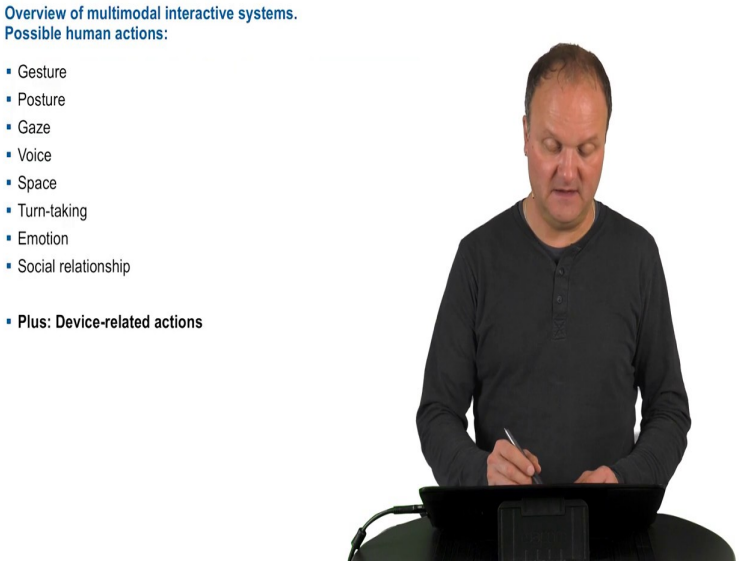
sensitivity, the tactile sensitivity, the deep sensitivity of the skin, the temperature and the pain sensitivity. And then we have the olfactory perception, the gustatory perception and the vestibular perception.

In a technical system, these perception modalities do not have necessarily one counterpart. For example for processing visual information you may use a camera but also a pen to provide information into the system using, for example 3D gestures or face recognition or lip reading or text or handwriting recognition.

You may use microphone in order to provide a signal which then can be recognized by a speech recognizer. Or you might use a touch screen for putting in 2D gestures into a system provided by 2D gesture recognition. So you see that there is no one to one mapping between the human capabilities in terms of modalities, perception modalities and the system equivalence to them.

Actually from an information theory point of view and from an informatic and media science point of view we usually talk about the modality as the use of a particular physical channel within our actions in order to provide information to the system. And this can be done in

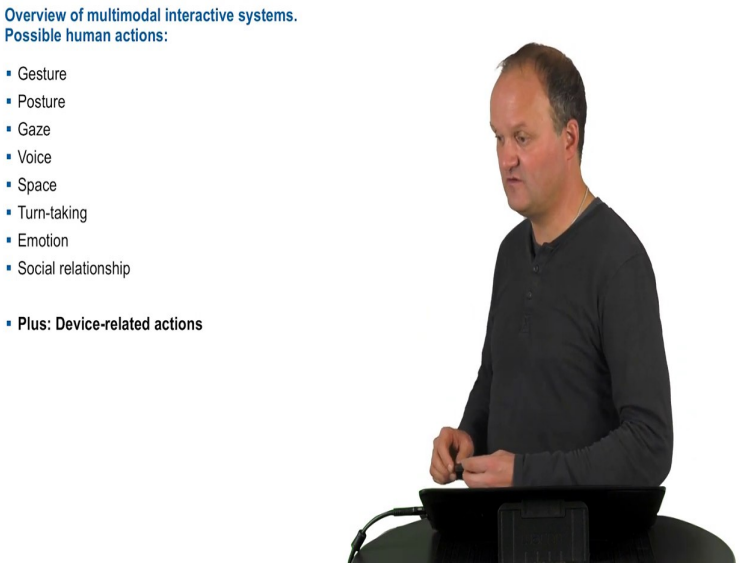
(Refer Slide Time: 07:19)



different ways.

For example by doing gestures, by exercising a certain posture

(Refer Slide Time: 07:27)

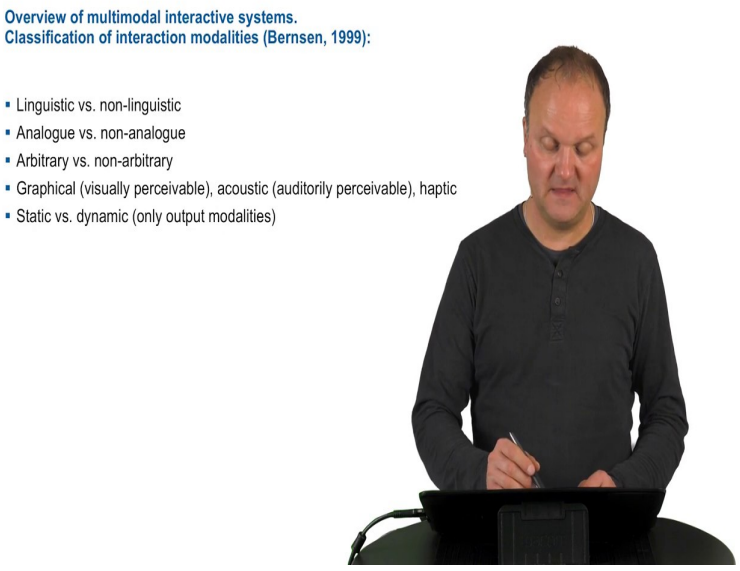


towards the systems, so I am moving my body towards the system, I can use my gaze that is where I am looking to in order to provide information to a system, I can use voice which is then recognized by the system or speech, I can make use of the available space.

I can use a different turn-taking mechanisms during interactions for example, interrupting the system, talking over the system is on, I can express emotions and I can also try to display social relationship.

These are actions which are normal human actions which we use in our human-human communication scenario and there are in addition to that, device-related actions that are actions which are displayed towards the system because there is a physical device like a keyboard which I am typing on in order to provide information to the system.

(Refer Slide Time: 08:19)



There have been different attempts to classify the interaction modalities with multimodal interactive systems; the most popular of these schemes was developed by Niels Ole Bernsen and collaborators in 1999.

He distinguishes between linguistic versus non-linguistic modalities for example speech, analog versus non-analog modalities, arbitrary versus non-arbitrary modalities, graphical versus acoustic versus haptic modalities and static versus dynamic or dynamically changing modalities.

And each of these modality classes have certain characteristics. Niels Ole Bernsen calls them modality properties which might be helpful for the system designer to decide on which interaction modalities to choose. For example an acoustic modality has the property that it is

omni-directional. So you would not like to use the acoustic input modality for provide your secret pin code to a banking ATM machine.

And these are, this is just one example of certain characteristics which result from these classification of modalities which can be helpful in order to decide which interaction modality is helpful, is useful and appropriate for a certain system.

(Refer Slide Time: 09:43)