Gestures and Speech in Cars

Uli Reissner
reissner@in.tum.de

ABSTRACT

Haptic use of infotainment (information and entertainment) systems in cars, deflect the driver from his primary task, the driving. Therefore we need new Human-Machine Interfaces (HMI), which do not require the driver's full attention, for controlling infotainment systems in cars. Gestures, speech and sounds provide an intuitive addition for existing haptical controls in automotive environments. While sounds are in common used as output, speech and gestures can be used as input modalities as well. The combination of all these devices leads us to multimodal HMIs, where different input and output devices can be used at the same time.

1 INTRODUCTION

1.1 Evolution of Human-Machine-Interfaces in Cars

An increasing number of new functionalities that enhance safety and driving performace or increase the level of comfort are included in modern cars. While first passive safety systems like airbags were integrated, later systems, that directly affect the driving process, like active cruise control were added. Additionally more and more information systems have become of general interest, for example: navigation systems, restaurants guides or telephone systems. In the beginning of the 80s, the integration of new infotainment systems, like radio or telephone, which where developed by different suppliers, results in an unmanageable amount of displays and control elements (see fig. 1). Use of different user interfaces by each supplier complicates usage. This development requires invention of other strategies, to enable handling in spite of increasing functional range. Since the middle 90s the trend is to include complete infotainment functions in one *integrated operating concept*, such as the BMW iDrive [1], were over 700 different Functions are handled in full feature mode.

1.2 Driver Distraction

According to studies [3, 4], 20 to 40 percent of car accidents are caused by driver distraction. Driver attention is affected by a lack of concentration, a talk to passengers, outside events, the use of infotainment devices and others causes.
The studies although shows, that a good design of HMIs, which are easily and intuitive to use, can really decrease the driver distraction from his main task, driving.

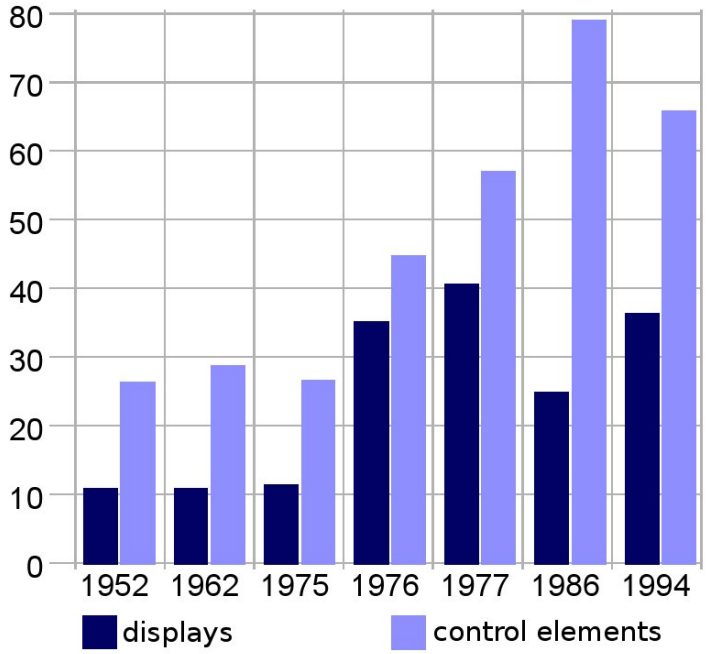

Figure 1: Increase of displays and control elements in BMW cars [2]

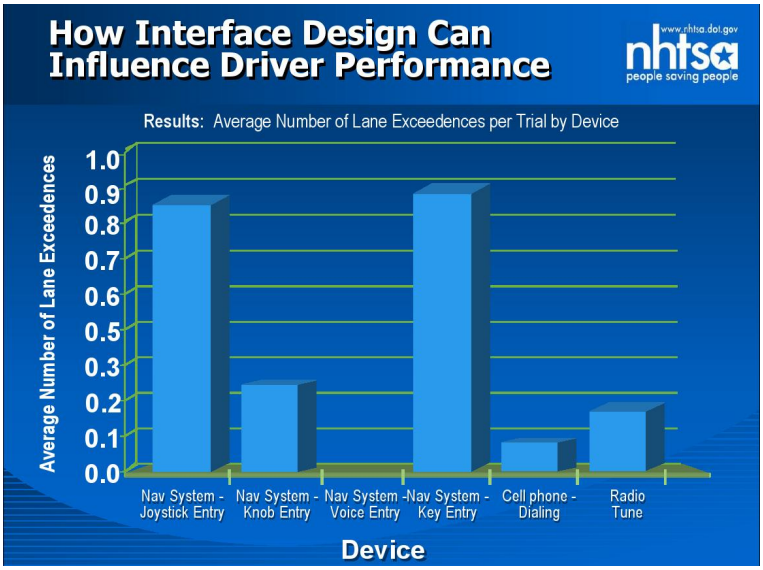

Figure 2: How interface design can influence driver performance [4]

2 SOUND AND SPEECH

2.1 Input

Speech seems to be a good alternative to standard haptical input devices in cars, but speech recognition is not the answer for all problems that appear in automotive environments. Visual attention is affected by speech input, despite the fact that speech uses a nonvisual channel. Users need a physical reference point to understand whether the system understood their spoken input or not[5]. While and shortly after speaking, drivers look at a point in the instrument cluster. In addition, speech recognition in cars has the same problems as in other VE environments [6]: poor speech recognition systems and missing strategies for segmentation. Also cars are a noisy environment which additionally complicates speech recognition.

2.2 Output

Auditory displays provide a second communication channel, reducing the load on the visual channel. Minimal interference with the visual channel do occur since some attention is drawn to the perception of the acoustic signals [7].

The auditory channel can be used to issue informational or alarm signals and for speech based communication. Alarm signals are given, for instance, when the environmental temperature is below a certain threshold or when the Active Cruise Control (ACC) works outside its operating range (too high driving speeds) and thus turns itself off. Such sounds inform the driver about a contextual change that can have influence on the car's safety. Sounds can be directionally encoded [8] to provide spatial cues. In this respect, 3D sound displays have the potential to support drivers in locating objects in the environment, thereby warning drivers about imminent dangers of collision. For example, an older version of BMWs parking assistant provided warning sounds from corners of the car close to nearby obstacles. The use of auditory warnings is promising because they immediately capture a drivers attention. On the other hand humans have problems distinguishing between many different pitches. Therefore i suggest to use sounds just in situations of an alert, where immediate attention of the user is required. An auditory hint can inform about the direction of the danger and further information could be provided in another modality.

Speech based communication (input and output) is most often used for tertiary interaction, e.g., for menu selections or other user actions. Speech output can also be used as feedback to actions that have not used speech input. It can be helpful when large quantities of information (texts) have to be conveyed to the driver. Examples are incoming messages: SMS or electronic mail.

However, there are limitations to the use of auditory displays. Some types of information are transmitted faster and more understandably in a visual representation. Spatial information, such as distances to environmental obstacles, is generally easier to perceive visually than aurally. Due to the sequential nature of spoken words, such continuous information is easier presented in the visual channel. In addition, the use of different sound frequencies can result in inaccurate interpretations. Furthermore auditive output interferes with the driving capabilities to a certain degree. When auditive output is given just when a difficult driving situation occurs, the drivers attention can be captured by the acoustic output and the imminent danger might go unnoticed. Speech output for warnings and alerts is not recommendable. Warning sounds have to be understood immediately, but in case of speech output, the driver has to listen to the whole notification.

Sound output quickly becomes annoying since it disturbs other auditive activities like music or conversations. Furthermore, drivers may feel embarrassed, if comments on driving mistakes are overheard by fellow passengers.

3 HEAD AND HAND GESTURES

Sometimes, like when one is driving a convertible or if it is too noisy, speech input can not be used. Also if one want to continuously zoom or scroll in a map, speech input is not the right way. Here, gesture input can be a good alternative to existing input devices. Gestures correspond to a movement of individual limbs of the body and are used to communicate information. Moreover the recognition of gestures can easily be executed by humans as an unconscious process. Human beings can identify certain movements as gestures although they neither know the semantic nor the specific form of the gestures.
Pilot study [9] proved, that gesture operations are advantageous and solve following questions.

- How *intuitive* is gesture?
  Intuitive use of gestures is the primary aspect for later use in cars. Some functions, like lift the telephone receiver or adjust the volume, are intuitive to use. Others, like activating the air conditioning system with one gesture, are not suitable.
- How good is the user acceptance?
  Before the users self tested gesture input, most had a positive tenor, but some could not imagine to use gestures to controll devices. After the study almost all users were thinking positive of gesture input.
- How gesture operation can be influenced by design of Human Machine Interface (Visualization)?
  The study shows, that a good visualization of the input device, can help using gestures.

3.1 Types of gestures

First, is has to be clarified, how the term *gesture* is used in this work. Therefore first a schematic overview is given, in which the modality gesture has to be divided into different categories. An important aspect for this work, is the communication purpose, which is used to outline gestures from random body movements. Therefore gestures differ as follows[2]:

- *primary gestures*: Gestures, which are only used in a communicative aim.
- *secondary gestures*: Acts, which communicates casually information, but this is not their main purpose.

For gestures as input modality, only primary gestures are used in the following, which can be be categorized as follows.

- *full body gestures*:
  complete body is used as communication device (e.g. pantomime).
- *partial body gestures*:
  the commonly used gesture result from head and hand gestures.

- *dynamic gestures*:
  body movement, where the main information aspect, which should be communicated, is in the motion sequence(e.g. head nodding)
- *static gestures*:
  here, no body movement is necessary, only the shaping of a body part is essential. The manual alphabet mainly use static gestures(see fig. 3).

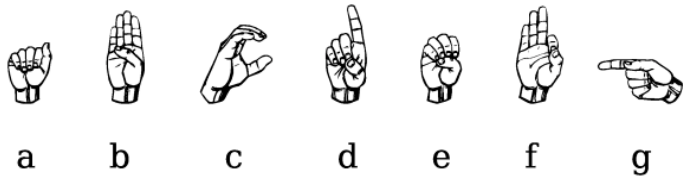

Figure 3: manual alphabet of American Sign Language [10]

According to preferably natural modes of communication, dynamic partial body gestures, namely head and hand gestures, are used as input for human machine dialog.Static gestures should not be used, because they are a more artificial mode of communication. In addition almost no static gestures were used in user studies[4] to gesture interaction. Dynamic gestures can be divided into two different groups:

- *discrete dynamic gestures*:
  closed motion sequence, which communicate a specific semantic content and causes one specified system reaction
- *continuous dynamic gestures*:
  the system status is changed directly while the gesture take place. In contrast to discrete dynamic gestures, information content is in moving direction and also in their amplitude. Thus usage of continuous gestures allow stepless manipulation of control variables, like sound level or position of a graphic object.

With regard to a technical recognition system, gestures can be identified on the basis of a corresponding movement trajectory that is characterized by selected attributes like symmetry and temporal seclusion.

3.2 Application Scenarios

In general, gestures facilitate a natural way to operate selected in-car devices. Thus gestures can increase both comfort and driving safety since the eyes can focus on the road. The recognition of head gestures mostly concentrates on detecting shaking and nodding to communicate approval or rejection. Thus head gestures expose their greatest potential as an intuitive alternative in any kind of yes/no decision of system initiated questions or option dialogs. As an example, incoming calls can be accepted or denied, new messages can be read, answered or deleted and help can be activated. Hand gestures provide a seamless way to skip between individual cd-tracks or radio stations and to enable or disable audio sources. In addition, they can be used for shortcut functions, enabling the user to switch between different submenus of the infotainment system faster and more intuitive compared to standard button interactions.

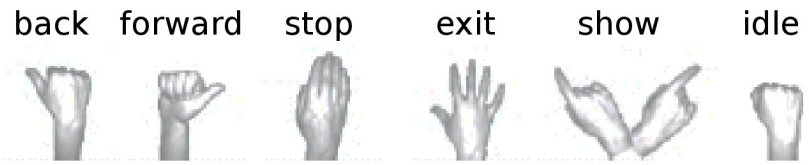

Figure 4: static gestures and dedicated functions [11]

3.3 Different Implementations

Many research groups have contributed significant work in the field of gesture recognition. With regard to an automotive environment, Akyol [11] has developed a system called iGest, that can be used to control traffic information and email functions. Totally, 16 dynamic and six static gestures (see fig.4) can be differentiated. The images are captured by an infrared camera that is attached to a active infrared lightning module. Due to the complex classification algorithms, only static gestures can be evaluated in real-time. Concentrating on head gestures, Morimoto [12] has developed a system that can track movements in the facial plane by evaluating the temporal sequence image rotations. The parameters are processed by a dynamic vector quantization scheme to form the abstract input symbols of a discrete Hidden Markov Model which can differentiate between four different gestures (yes, no, maybe and hello). Based on the IBM PupilCam technology, Davis [13] proposed a real-time approach for detecting user acknowledgements. Motion parameters are evaluated in a finite state machine which incorporates individual timing parameters. In an alternative approach, Tang [14] identifies relevant features in the optical flow and uses them as input for a neural network to classify the gestures. As an advantage the system is robust with regard to different background conditions.
In 3.3.2 a system from Geiger [2], in which a field of infrared sensors is used to locate the hand and the head, is introduced.
In 3.3.1 a system form Althoff [15] which use a near infrared imaging approach is introduced.

3.3.1 Gesture Recognition with Cameras

Figure 5: Entropy motion segmentation: (a) IR camera image, (b) difference image, (c) entropy image, (d) binarized entropy image, (e) result after geometrical forearm filtering[15]

As a result of a longterm research cooperation between the Technischen Universiät München and BMW Research and Technology, a robust and flexible system for the video-based analysis of dynamic hand- and head gestures has been implemented in a BMW car for demonstration. A near infrared imaging approach and motion based entropy technique has been applied

instead of conventional, mostly color based methods. As the directional information is more important than the accuracy of the segmentation, a more robust and plain motion based technique is used instead of threshold operations to detect hand gestures.

First the entropy (see fig. 5(c)) is calculated on difference pictures (see fig. 5(b)) instead of plain images as depicted in figure 5(a). To suppress meaningless movements the entropy image is binarized (see fig. 5(d)). Afterwards morphological operations remove areas of noise and clean up the remaining regions (see fig. 5(d)). Finally a forearm filtering process is applied on the region with the biggest area.

A form-based segmentation algorithm has been chosen to localize the head and to extract all relevant facial features. The initial head extraction is performed on the whole image. Further searching steps are limited to the last head position increased by an additional confidence area. Likewise the search region for the eyes is limited to the upper half of the extracted head region.

Further Hidden Markov Models are used to classify the gestures. While Hidden Markov Models (HMM) have originally been applied in the field of automatic speech recognition, they have successfully been used for other dynamical classification tasks, such as gesture recognition [16], in the last years. To reduce the amount of information provided by the image sequences, the gestures are limited to their relevant data, consisting of the trajectory, velocity and hand form of the gesture. These samples are used to train a stochastic model for every gesture. In the recognition phase an output score is calculated for each model from the active input sequence, giving the probability that the corresponding model generates the underlying gesture. The model with the highest output score represents the recognized gesture.

3.3.2 Gesture Recognition with Infrared Distance Sensors

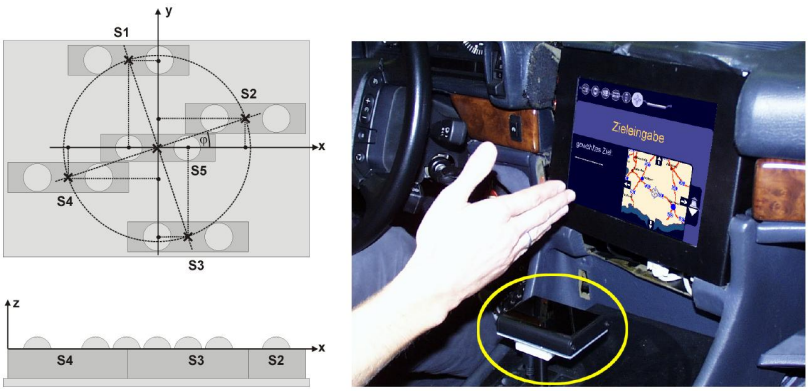

Figure 6: sensor arrangement for hand gestures(left) which is mounted on top of the gear shifter(right) [2]

Geiger [2] has presented an alternative to video-based systems. In his work he used a field of infrared distance sensors to locate hand and head.
For head gesture recognition he used 4 infrared distance sensors mounted around the head rest of the driver (see fig. 7), which is optimized for nodding and shaking.
For hand gesture recognition, 5 infrared distance sensors are mounted in-plane (x/y-level), so

that the distance measurement is in z direction (see fig. 6). The arrangement of sensors S1 to S4 is optimized for recognition horizontal hand movements. Furthermore the sensor S5 is mounted in the middle to better recognize vertical hand movements.

The structure in principle, he uses to recognize gestures correspond to classic pattern recognition. First information, about the gesture to classified, is recorded by distance measurement and a preprocessing is done. In the following segmentation start and end of the movement are detected The information, gain in this time slice, is edited and sent to the classification.

In the classification, the unknown gesture sequence is compared to all reference sequences. There it must be noticed, that the same gesture can need different time, because of different execution speed. So the Dynamic Time Warping Algorithm (DTW) [17] is used. The DTW algorithm calculate the distance between two sequences. The motion sequence is mapped to the gesture with the smallest distance. The gesture vocabulary mainly consists of directional gestures to navigate within a menu structure or map and to control a music player. Although the sensor array does not achieve the resolution of a video-based image analysis, this system is highly robust, needs only low computing power (a Intel Pentium 133) and can get along with simple sensor hardware.

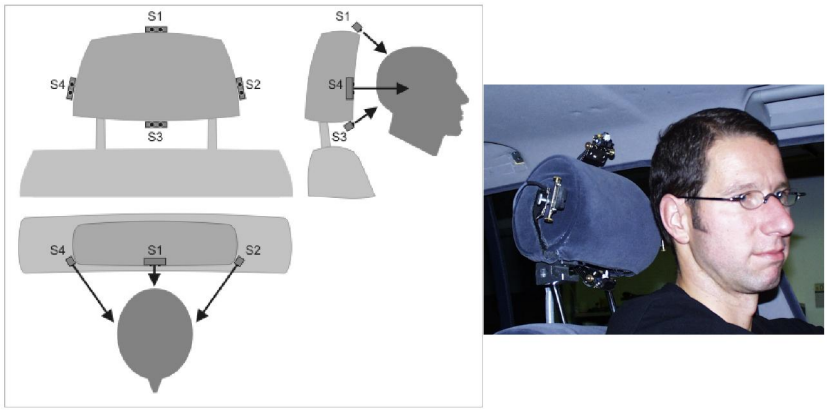

Figure 7: sensor arrangement for head gesture recognition: schematic diagram (left) and real installation with user(right) [2]

3.4 Evaluation of Gesture Recognition

Both systems (Section 3.3.1 and Section 3.3.2) achieve good recognition rates about 95 %. While the system form Geiger [2] provide good result also with a small number of training's, systems with the Hidden Markov Models based classification provides better result with a high number of training's (see fig. 8).

Additional Geiger[2] analyzed how gesture input distract the driver in comparison to haptic input. In his studies he used a drive simulator, where the steering variances between the expected value and the current value are measured.

In addition to his main task, driving, an operator presents a second task to the user, like tune radio to 93.3 MHz. While doing this, an object is presented to the user, which he has to recognize and later present to the operator (see fig. 9). The steering variances at the presentation of the second task show a good example for Cognitive Capture.

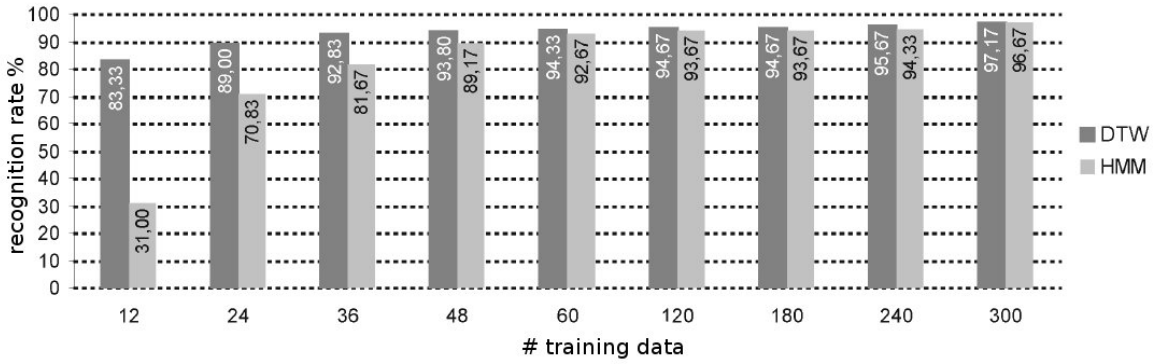

Figure 8: recognition rate for *Dynamic Time Warping algorithm (DTW 3.3.2) and Hidden Markov Models (HMM 3.3.1) [2]*

The studies shows that gesture input distract the driver less than haptic input. Use of haptic input causes longer input times, more steering faults and more faults in object recognition. Also the driver subjective feel the gesture input more comfortable and less distractive.

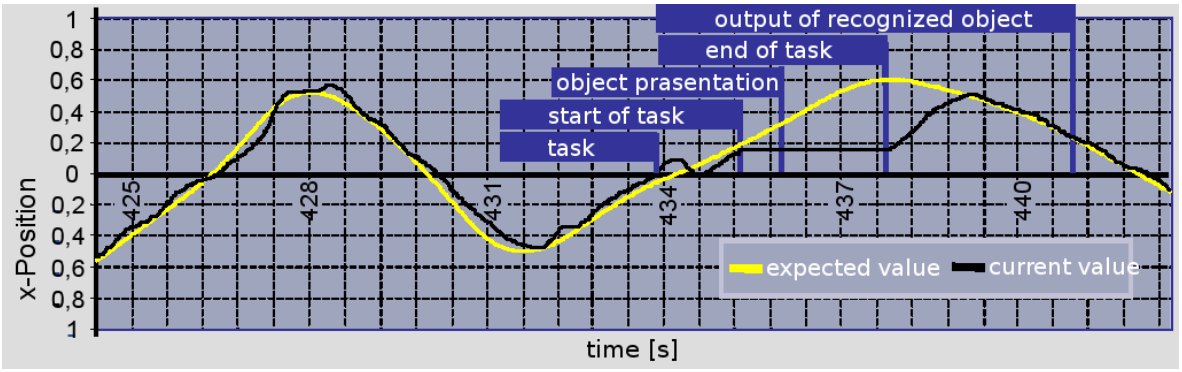

Figure 9: study schedule: while steering the user has to execute a second task, like tune radio to 93.3 MHz, and recognize an object.[2]

4 MULTIMODAL HUMAN-MACHINE INTERFACES (HMI)

Multimodal HMI intend to combine the advantages of different input and output facilities. While modern automotive infotainment systems already make use of multimedia output, like navigation systems and park assistant, input is performed with mechanical devices. This haptic operation of infotainment system compete with the use of hands to drive [18]. For this, the drivre has to avert his eyes from the traffic. To avoid this, new forms of input devices, like speech and gesture, have tob be combined with speech and sound output to multimodal HMI.

In case of multiple requirements to users Akyol [18] shows, that the best is to use different resources. According to Wickens [19] humans have independent resources, which can be used

to increase efficiency.
For information exchange Wickens differs between sense and output resources, which are again subdivided after modalities (see fig. 10). Thus a intermodal independency exists, by which concurrent speech and gesture or hear and see is possible without performance loss. Also a certain independency between sense and output processing is noticed. For example, the vision capacity is not reduced by simultaneously speech. This applies particularly, if no mental resource is used together. According to this it is possible to reduce the workload with a multimodal input system and thereby enhance operation efficiency.

An extra motive for development of multimodal HMI is the inter human communication. There, more modalities are used combined or the modality is used depending on the situation.

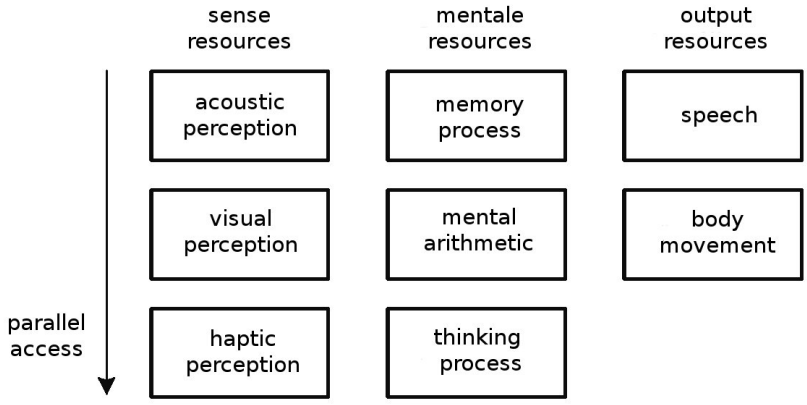

Figure 10: dependency of multiple resources after Wickens[19]

5 CONCLUSION

Gestures and speech provide an intersting alternative to classical, tactile control, as they are an important part of inter-human communication. Thus automatic recognition of gestures and speech in an automotive environment can increase both the usability of complex driver information systems and driving safety since the eyes can be kept on the road.

The strength of speech is for example, compact transmission of complex instructions. Also gesture input has specific advantages. On the one hand, it can be used in noisy environments. On the other hand, gesture control is faster and more efficient for small commands, like “next menu point” or “mute audio”. System callbacks can intuitively and easily answered with head nodding or shaking. Analog settings, like setting music volume can comfortably get done with continuous gestures. There is a weakness of speech, which generally is only suitable for absolute information.

The ability to develop future multimodal systems depends on the knowledge of natural integration patterns that typify peoples‘s combined use of different input modes. Given the complex nature of users multimodal interaction, cognitive science will play an essential role in guiding the design of robust multimodal systems [20].

References

[1] BMW iDrive. http://de.wikipedia.org/wiki/IDrive, March 2007.

[2] M. Geiger. *Berührungslose Bedienung von Infotainment-Systemen im Fahrzeug*. PhD thesis, TU München, 2003.

[3] SP McEvoy, MR Stevenson, and M. Woodward. The imapct of driver distraction on road safety: Results from a representative survey in two Australian states. *Injurence Prevention*, (12):242–247, 2006.

[4] Joseph N. Kanianthra; National Highway Traffic Safety Administration. Driver Distraction: Understanding the Problem, Identifying Solutions. International Consumer Electronic Show, 2005.

[5] U. Plangger, A.Khosravi, and G. Helas. Einfluss von sprachlicher und haptischer Bedienung auf die Fahrleistung und das Blickverhalten. Master's thesis, Technische Universität München, Chair for Ergonomics, 2005.

[6] D.A. Bowmann, E. Kruijff, J.J. LaViola, and I.Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley, 2004.

[7] Marcus Tönnis, Verena Broy, and Gudrun Klinker. A Survey of Challenges Related to the Design of 3D User Interfaces for Car Drivers. In *Proceedings of the 1st IEEE Symposium on 3D User Interfaces (3D UI)*, mar 2006.

[8] Marcus Tönnis and Gudrun Klinker. Effective Control of a Car Drivers Attention for Visual and Acoustic Guidance towards the Direction of Imminent Dangers. In *Proceedings of the 5th International Symposium on Mixed and Augmented Reality (ISMAR)*, October 2006.

[9] M. Zobl, M. Geiger, K. Bengler, and M. Lang. A Usability Study on Hand Gesture Controlled Operation of In-Car Devices. In *Proc. of the 9th Int. Conf. on Human-Computer Interaction (HCI International 2001)*, New Orleans, Louisiana, USA, 5.-10.08.2001.

[10] American Manual Alphabet. http://de.wikipedia.org/wiki/Fingeralphabet, March 2007.

[11] U.Canzler and S. Akyol. *GeKomm - Gestenbasierte Mensch-Maschine Kommunikation im Fahrzeug*. PhD thesis, Rheinisch-Westflische Technische Hochschule Aachen, 2000.

[12] C. Morimoto et al. Recognition of head gestures using hidden markov models. In *Proceedings of IEEE Int. Conf. on Pattern Recognition* , Vienna, 1996.

[13] J.Davis et al. A perceptual user interface for recognition head gesture acknowledgments. In *WS on Perceptive User Interfaces (PUI 01)*, USA, 2001.

[14] J.Tang et al. A head gesture recognition algorithm. In *Prov. of the 3rd Int. Conf. on Multimodal Interfaces*, Beijing China 2000, 2000.

[15] F. Althoff, R. Lindl, and L. Walchshaeusl. Robust Multimodal Hand- and Head Gesture Recognition for controlling Automotive Infotainment Systems. In *VDI-Tagung: Der Fahrer im 21. Jahrhundert*, Braunschweig, Germany, November 22-23 2005.

[16] H. Lee and J. Kim. An HMM-Based Threshold Model Approach for Gesture Recognition. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 21, pages 961–973, October 1999.

[17] Tobias Sielhorst, Tobias Blum, and Nassir Navab. Synchronizing 3D Movements for Quantitative Comparison and Simultaneous Visualization of Actions. In *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05)*, pages 38–47, Vienna, Austria, Oct. 2005.

[18] Suat Akyol, Lars Libuda, and Karl-Friedrich Kraiss. *Kraftfahrzeugführung*, chapter 9. Multimodale Benutzung adaptiver Kfz-Bordsysteme S.137-154. Springer-Verlag, Berlin, 2001.

[19] C.D. Wickens. *Engineering Psychology and Human Performance (2nd Edition)*. Harper Collins, New York, 1992.

[20] Sharon Oviatt. Ten myths of multimodal interaction. *Commun. ACM*, 42(11):74–81, 1999.