Patterns for Robust and Flexible Multimodal Interaction

Andreas Ratzka
Institute for Media, Information and Cultural Studies
University of Regensburg
D-93040 Regensburg, Germany
Andreas.Ratzka@sprachlit.uni-regensburg.de

Abstract

Multimodal interaction aims at more flexible, more robust, more efficient and more natural interaction than can be achieved with traditional unimodal interactive systems. In order to achieve this, the developer needs some design support in order to select appropriate modalities, to find appropriate modality combinations and to implement promising modality adaptation strategies. This paper presents a first sketch of an emerging pattern language for multimodal interaction and focusses on patterns for "flexible interaction" and patterns for "robust interaction". This work is part of a thesis project on pattern-based usability engineering for multimodal interaction.

1 Introduction

Multimodal interaction means interaction via several interaction channels such as speech, pointing device, graphics and the like. An interaction modality is defined by its interaction channel (acoustic, visual, haptic/tactile) and the interaction "language" (pointing, naming, emulating). Example modalities are pointing gestures, keyboard input, speech input, speech output, graphic output, and tactile output (e.g. vibrating steering wheels).

Although every interactive system combines at least two interaction modalities (one for input, another one for output) not everyone is necessarily multimodal. Multimodality implies the use of at least either more than one input channel or more than one ouptut channel. Although mouse-based pointing and keyboard input are different modalities, typical WIMP[1] and desktop applications are not classified as multimodal systems. Only when different *channels* (acoustic, visual, haptic/tactile) are used for either input or output, a system can be called multimodal.

According to Oviatt & Kuhn (1998) goals of multimodal interaction are

- flexibility and adaptability of the system with respect to users and context of use,
- high interaction robustness due to mutual disambiguation of input sources,
- interaction efficiency because of better integration into the work situation, and

[1] Windows, Icons, Menus, Pointing device

- the possibility to interact with the system in a natural way.

This work is related to a thesis project about usability engineering of multimodal interactive systems. One assumption of this thesis project is that, although multimodal interaction is a relatively new field with very little market penetration, there exists already a corpus of well founded research results and successful system implementations in which recurring patterns can be found (Ratzka & Wolff 2006).

The usability engineering lifecycle (Mayhew 1999) comprises the phases of requirements analysis, work reengineering, design standards, detailed design and implementation. All of these lifecycle steps need both specification formalisms and knowledge-based design support. Traditional approaches for multimodal interaction design provide implementation frameworks (Niedermaier 2003) or formalisms for specifying detailed design (Dragičević 2004, Duarte & Carriço 2006). Other work provides knowledge-based design support in the phases of requirement analysis and work reengineering (Bernsen 1999, Bürgy 2002, Obrenović et al. 2007).

The application of patterns can complement these approaches since they seem to be an appropriate tool for providing knowledge-based design support accross all lifecycle phases.

This pattern collection is based on a thorough literature review of multimodal interaction in industrial and research projects. The following questions helped to find an adequate categorisation of question-solution pairs and thus a basis for pattern mining:

- When should certain interaction modalities (speech input or pointing input, speech output or graphic output) be used?
- How should multiple interaction modalities (speech and pointing, speech and graphics etc.) be combined?
- How can modalities be adapted according to context of use (user, environment, or situation)?

2 Pattern Overview

This paper focusses on usability design patterns for robust and accessible multimodal interaction. Two pattern sub-collections are presented. The first one, *Flexible Interaction*, focuses on accessibility and cross-context usability. The abstract principle lying behind this sub-collection is giving the user the possibility to select appropriate interaction modalities according to context factors. In order to achieve this goal, this pattern group has to make use of suitable adaptation strategies which are described in the patterns *Multiple Ways of Input*, *Global Channel Configuration* and *Context Adaptation*.

The second sub-collection *Robust Interaction* focusses on avoiding and corroborating recognition errors as well as assuring communication between user and system. It contains the two general patterns *Redundant Input* and *Redundant Output* which present its abstract principle consisting in exploiting several interaction channels. The pattern *Important Message*, which has been described by Tidwell (1999, view section 5 in this paper) can be seen as concretisation of *Redundant Output*. The patterns *Multimodal N-best Selection* and *Spelling-based Hypothesis Reduction* are specializations of the pattern *Redundant Input*. They can be used in connection with the pattern *Speech-enabled Form* (Ratzka 2008, view section 5 in this paper), a specialization of the pattern *Form* (Tidwell 1999) which can be itself enriched with the patterns *Dropdown Chooser* and *Autocompletion* (Tidwell 2005). The patterns *Dropdown Chooser* (Tidwell 2005) and *Continuous Filter* (van Welie & Trætteberg 2000) are used to implement *Multimodal N-best Selection* and *Spelling-based Hypothesis Reduction* respectively.

A short outline of patterns referenced in this paper but described elsewhere can be found in section 5.

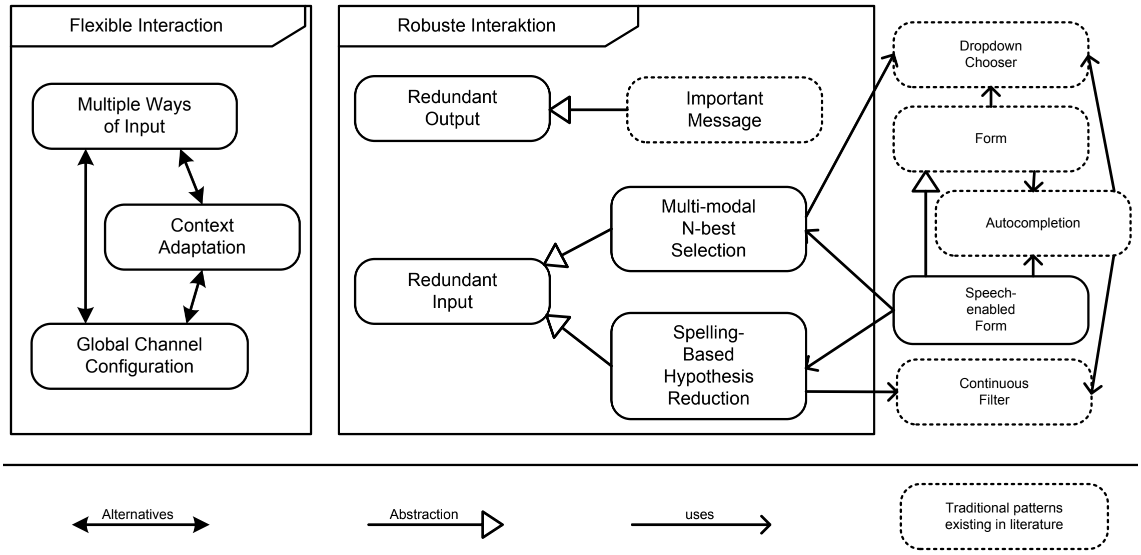

Figure 1: Pattern Map

3 Patterns for Flexible Multimodal Interaction

The patterns for flexible interaction focus on accessibility and usability across changing situations, environmental and other context factors. The three patterns described here provide each one different runtime modality adaptation strategies, i.e. ways of dynamically allocating interaction modalities:

- *Multiple Ways of Input* offers the user several alternative interaction techniques. The user is free to select the interaction modality that is most appropriate in the current context. He need not perform additional configuration steps.
- *Global Channel Configuration* provides several interaction profiles with each different configurations of input and output channels. This way, the system can be tailored to typical contexts of use. The user can switch the interaction profile in just one interaction step using an always-on-top widget or function button. Thus, the user can switch on/off audio output or speech recognition when appropriate.
- *Context Adaptation* requires the system to evaluate interaction history and environment data to adapt system behaviour accordingly. In contrast to *Multiple Ways of Input* and *Global Channel Configuration*, *Context Adaptation* provides system-initiated adaptation strategies.

These patterns try to resolve following forces:

- Typing is powerful for a lot of tasks but if the target user group includes typing-unskilled or even illiterate people other alternatives have to be used.
- Speech input is well suited for both text input and item selection. But in loud environments, speech recognition will be very error prone because the background noise masks the proper signal or is interpreted as input where there is none.
- Mechanical input via keyboards or pointing devices is widely applicable. But if the user's hands are occupied, wet, or dirty, there is no way to operate the systems.
- Graphic output and feedback is useful in a lot of situations but cannot be perceived in bad lighting conditions or by blind people. At the same time, speech output might be difficult to understand or even be overheard in loud environments.
- Speech output and input can make mobile interaction more comfortable. But confidential data must not be read out loudly in public environments.
- Environmental factors can be controlled via special installations such as specially mounted lamps, phone booths, directional speakers, earphones, or view shields, but these measures are not viable in every case such as mobile interaction.

Multiple Ways of Input

Context Context factors are not always predictable. This holds especially for mobile interaction in changing environments or public information kiosks that should be usable for quite different user groups.

Problem How can input modalities be adapted to the context of use without burdening the user with additional configuration tasks?

Forces

- The user should be able to interact with the system using preferred and task appropriate interaction styles. However, disabilities, or changing environmental conditions such as lighting or background noise may affect the usability and robustness of task-optimized interaction modalities.
- If background noise is low, speech input and output provide a valid interaction style. However, in public environments, bystanders might feel annoyed by persons conversing with an interactive system.
- Environmental factors can be controlled via special installations:
  - Specially mounted lamps help to overcome bad lighting conditions.
  - Phone booths reduce background noise and help to assure privacy.
  - Earphones and directional speakers avoid to annoy bystanders and enhance privacy.
  - View shields help to assure privacy.

  But these measures are not viable in every case such as mobile interaction.
- Users might differ according to preferences, language, reading, typing skills etc. The system can provide alternative interaction styles and adapt to the user. However, the system is not able to predict precisely which input modalities are most likely to be selected by the user in the current situation.

Solution Enable each system function to be operated via several alternative interaction modalities differing in the interaction channel.

Enable the user to select his preferred interaction channel, be it speech, typing or pointing. This modality should be active, i.e. the user should not have to switch the system configuration to use it.

The system has to be designed in a way, that the user can change interaction modalities wherever sensible, even for single interaction steps while performing a certain task.

Labels, help messages, prompts, console and speech commands should share a uniform vocabulary in order to minimise habituation and learning efforts.

Consequences

- Providing several alternative interaction styles, the system supports access for physically disabled or illiterate people.
- As the user can select from a set of complementary modalities which are each differently affected by environmental factors, the system can be used in varying environmental conditions.
  - In loud or public environments the users can simply sidestep to pointing / text input.
  - When background noise is low users can opt for speech interaction.
- Expert users can estimate whether speech input or pointing gestures are possible or desirable at the current moment.
- Interaction can be assured even though no special installations such as view shields or directional speakers are available.
- It is up to the users to select their preferred interaction style. They do not need to rely on mystic autoadaptation mechanisms. Instead, they gain self-confidence as they are controlling the system which leads to higher user satisfaction.
- First time users might not know which interaction styles are available at all. The system must provide effective help and prompting strategies that reveal alternative interaction modalities.
- The more flexible a system is, the more planning, testing and reviewing is needed during design as the number of error sources increases rapidly with system complexity.

Rationale According to user characteristics, preferences, environment and situation, different interaction modalities are preferable (Oviatt et al. 2000, Bürgy 2002, Obrenović et al. 2007).

Users can judge better than the system, which interaction modality and style is appropriate, e.g. whether background noise or bad lighting impedes the use of speech or graphics, whether surrounding people might feel annoyed because of spoken interaction or whether private information is being conveyed.

Ibrahim & Johansson (2002) have shown for their multimodal TV guide, that users prefer, when they can choose to use direct manipulation and speech either unaccompanied or combined in order to adapt to the current context of use.

Users learn to estimate and select the most context-appropriate modality. After recognition errors, users tend to switch the interaction modality (Oviatt et al. 2000, Oviatt & vanGent 1996, Oviatt et al. 1999, Oviatt 1999).

Known Uses In the SmartKom system (Reithinger et al. 2003) the user can interact either via pen or speech.

Mobile systems such as Microsoft's MiPad (Miyazaki 2002, Microsoft) and IBM's Personal Speech Assistant (Comerford et al. 2001) are good examples of systems allowing users to flexibly select input modalities. The same holds for driver assitance system such as the ones decribed in Neuss (2001) and Pieraccini et al. (2004).

Further examples of this pattern can be found in the interactive TV guide by Ibrahim & Johansson (2002), and MICASSEM (McCaffery et al. 1998).

Related Patterns This pattern is on the one hand an alternative to *Global Channel Configuration* and *Context Adaptation*. On the other hand these patterns can be combined.

Whereas *Global Channel Configuration* requires an additional interaction step to select the input / output profile of the system, *Multiple Ways of Input* offers a wider spectrum of input modalities which can be used without additional configuration steps. In contrast to *Global Channel Configuration*, this pattern is restricted to the adaptation of user input.

That's why this pattern should be used with *Context Adaptation*. This way, system output can be adapted according to the user's input behaviour.

In contrast to system-driven *Context Adaptation*, *Multiple Ways of Input* offers, similarly to *Global Channel Configuration*, user-driven system adaptation.

Global Channel Configuration

Context Interactive devices that offer several alternative and complementary interaction channels such as audio input and output, typing and graphic manipulation have to be adapted to the context of use.

Problem How can interaction (input and output) be adapted to the current context of use, while giving the user control over the system without burdening him with too much configuration tasks?

Forces

- One can keep several input channels active and leave it up to the user to select the most appropriate interaction modality for the current context. However, the system might try to interpret input from unused distorted channels, e.g. because of background noise. This might lead to *false positives*, that is the system misinterprets background noise as input.

- Redundant output via several channels can assure information perception by users. However, in public environments, bystanders might feel annoyed by persons conversing with an interactive system. In the same way, it is not desirable, that private data be read out loudly by the system.

- The system may analyse interaction behaviour, lighting conditions, background noise, movement and position changes and adapt to the user and context of use. However, the system does not find out as fast as the user does which interaction modalities are most appropriate in the current context of use or for the current user.

- Even if automatic adaption works quite well, many users will prefer being able to control the system.

Solution Provide several interaction profiles with input and output channel configurations tailored to each context of use. Enable the user to select the interaction profile with only one additional interaction step.

Display for instance always on top buttons with self explaining icons or provide physical *push-to-talk* and *mute* buttons etc. so that the user can select the interaction profile (speech input, audio output), the notification profile of a mobile phone (ringing, vibrating, mute) with one click.

Consequences

- The user can quickly react to context changes and reconfigure the system accordingly in one interaction step:
  - Input channels such as speech input can be deactivated when necessary (in loud environments) without great effort by simply clicking the necessary button.
  - In order not to disturb bystanders in public environments, the user can deactivate speech output with one click.

- Users feel better when exercising control over the system instead of being delivered to its "caprices".
- Users have to do at least one additional interaction step which might be annoying anyway.
- Users might by fault select an inappropriate interaction profile.
- The users might not be able to rembember the current configuration state of the system and that they have to reconfigure the system at each situation change.

Rationale According to user preferences, abilities, environment and situation, different interaction modalites are appropriate and have to be preferred (Oviatt et al. 2000, Bürgy 2002, Obrenović et al. 2007).

Users learn to estimate and select the most context-appropriate modality. After recognition errors, users tend to switch the interaction modality (Oviatt et al. 2000, Oviatt & vanGent 1996, Oviatt et al. 1999, Oviatt 1999).

Known Uses The multimodal map-based system *SmartKom Mobile* covers the use cases of pedestrian and automotive navigation as well as map-based queries (Malaka et al. 2004) and allows the user to switch between several interaction modes (cf. Wasinger 2006, p. 59):

- *Default*: All input and output modalities are supported.
- *Listener*: Speech+graphics are supported for output, for input only pen gestures are possible.
- *Silent*: Only graphics for output and pen gestures for input are supported.
- *Speech Only*: Only speech interaction is active.

Some mobile phones such as *Nokia E71* offer the user to select profiles such as *office* or *home*. In addition to different startup screens, these profiles can be set to an appropriate context-dependent notification mode (ringing, vibrating).

With some restrictions, desktop applications, operating system environments, and multimedia applications that provide an audio icon in the system tray for setting the system's audio characteristics can be seen as examples for this pattern.

Push-to-talk buttons in some speech based driver assistance systems are examples, too: When operated once, speech output is stopped and speech input activated. Operated once more, speech interaction can be deactivated and, lateron, reactivated.

Related Patterns This pattern is an alternative to *Multiple Ways of Input* and *Context Adaptation* but can be used in combination with them, too.

In contrast to *Multiple Ways of Input*, which offers the user to select among several alternative input modalities without having to perform additional configuration actions, this pattern requires one additional interaction step. Furthermore, it comprises, in contrast to *Multiple Ways of Input*, the adaptation of system output.

In contrast to system-driven *Context Adaptation*, *Global Channel Configuration* and *Multiple Ways of Input* are forms of user-initiated system adaptation.

Context Adaptation

Context Examples for this pattern can be found in interactive devices that support different alternative and complementary interaction channels such as audio output and input, typing and graphic manipulation.

Problem How can interaction (input and output) be adapted to the current situation, environment and user without the user having to perform additional interaction steps?

Forces

- Redundant output via several channels can assure information perception by users. However, superfluous spoken output might disrupt or slow down the user's secondary task or annoy third persons.
- One can keep several input channels active and leave it up to the user to select the most appropriate interaction modality for the current context. However, the system might try to interpret input from unused distorted channels. This might lead to *false positives*, that is the system misinterprets background noise as input.
- Alternatively, one could let the user configure system input and output according to the current context of use. But, as long as sufficient information is available to the system additional configuration steps should be avoided.
- Letting the user configure the system himself seems not to be a problem at first. But the user might not be able to rembember the current configuration state of the system and to reconfigure the system at each situation change.

Solution The system should analyse as much assured context information as available to setup system configuration autonomously.

One information source that can be used is the interaction history:

- If the user interacts via speech, it is not clear where the user looks. In this case speech output should complement display updates.
- If the user does not want to annoy surrounding people, he will avoid speech interaction.
- If the user interacts with pointing device or keyboard – might be in order to avoid annoying surrounding people – he is usually looking at the display such that speech output is superfluous.

Other aspects of system state can be exploited for adaptation, too: A driver assistance system should disable touch screen input while the car is driving – the driver should keep his hands on the steering wheel. Smartphones should disable touch screen input while the user is making a telephone call – the ear of the user should not lead to unwanted actions.

Consequences

- Information output can be restricted according to the context of use.
- False positive recognitions can be avoided via context dependent recogniser activation.
- The user does not need to reconfigure the system repeatedly where this can be done by the system itself.
- The user need not remember configuration states and reconfigure the system at each situation change.
- Automatic adaptation can fail or be unappropriate. That's why the user should always have the possibility to carry over control and perform interaction configuration himself.

Known Uses In the SmartKom system (Reithinger et al. 2003) the user can interact either via pen or speech. System feedback is adapted in a way that fits to the user's attention: The TV-guide subsystem of SmartKom presents the TV-program usually as a listing but reads out spoken feedback to spoken queries. When the user is watching TV, the system presents the program list verbally, too, because it supposes that the visual channel is occupied.

Driver assistance system such as the ones decribed in Neuss (2001) and Pieraccini et al. (2004) offer the user to interact using speech or manual input devices. System output is adapted acordingly.

Smartphones disable touch screen input during telephone calls.

Related Patterns This pattern is an alternative to *Global Channel Configuration* and *Multiple Ways of Input* but may be used complementarily to these ones. In contrast to these user-driven adaptation strategies, *Context Adaptation* implements a system-initiated adaptation strategy.

This pattern can be used as complement to *Multiple Ways of Input* in order to update system output according to user input.

In addition to *Context Adaptation*, *Global Channel Configuration* should be used, to give the user control over the system.

4 Patterns for Robust Multimodal Interaction

This subcollection comprises a set of four patterns: two abstract two concrete. Both abstract patterns *Redundant Output* and *Redundant Input* go back to the same basic principle of mutual disambiguation of redundant signals. Whereas the pattern *Redundant Output* is used to present redundant data to the user such that the user is more likely to understand or at least perceive the information conveyed by the system, the pattern *Redundant Input* fuses user input coming from several channels in order to reduce recognition and interpretation errors.

The pattern *Redundant Output* has no concretisation within this collection. Tidwell's pattern *Important Message* (Tidwell 1999, view section 5 in this paper) can be seen as a concretisation of this one, as it suggests the usage of several modalities (visual and auditive signals) to attract the user's attention and convey urgent information.

For the pattern *Redundant Input* two refining patterns are given:

- *Multimodal N-best Selection* combines several interaction modalities for input and disambiguation dialogs. The user first speaks a word or command. As speech recognition is not a sharp but rather statistical process a list of *n* best hypotheses is returned. The user can now directly select from this list e.g. via pointing. When this step has to be done via speech, too, alternatives to simply re-speaking the item – e.g. naming the row number – should be supported and prompted for in order to avoid endless loops of repeated errors due to intrinsic acoustic confusability.

- *Spelling-based Hypothesis Reduction* combines several input modalities to reduce recognition errors. Typically, the user first inputs some letters via typing in order for the system to reduce the set of possible alternatives. Then, the user speaks the desired entry which can be properly recognised by the system. Spelling can be done after spoken selection, as well. Then, the system has to recalculate the recognition hypotheses according to the updated list of available alternatives.

These pattern, although focusing on robust interaction, are closely linked to patterns for efficient multimodal interaction, especially the patterns *Voice-based Interaction Shortcut* and *Speech-enabled Form* (Ratzka 2008, view section 5 in this paper) as well as to traditional user interface patterns such as *Continuous Filter* (van Welie & Trætteberg 2000, cf. section 5), *Form* (Tidwell 1999, cf. section 5), *Autocompletion* (Tidwell 2005, cf. section 5) and *Drop-down Chooser* (Tidwell 2005, cf. section 5).

Redundant Output

Context Communication channels might be unpredictably distorted due to bad lighting conditions, background noise, technical (network) problems or disabilities such as speech, motor or perception disorders.

Public systems that should be accessible for everybody should use this pattern especially during the first interaction steps when it is not clear which interaction channels are appropriate for the current user.

Problem How to assure information output when communication channels are distorted in an unforseeable way?

Forces

- The system can be configured or adapted to output information using modalities that are less affected by channel disorders. However, in some cases several interaction channels are distorted to some degree. Examples are:
  - Visually impaired people or illiterate people that want to interact in loud environments.
  - Deaf people that want to interact in bad lighting conditions or while moving around.
- Potential channel distortions might be circumvented by selecting alternative interaction channels. However, if the potentially distorted channel were otherwise the best candidate, abandoning this channel cannot be justified.
- The system can use those modalities that are most appropriate in the current environmental context. However, when the user's attention does not fit to the situation he might miss important notifications. E.g. when the system uses purely visual output due to high background noise level but the user's visual attention is focused elsewhere the system fails to notify the user.

Solution Combine several output channels in order to make use of redundancy. Information should be output both visually and acoustically and possibly even in a haptic way (e.g. using vibration) to raise the probability that it is perceived and can be understood by the user.

Consequences

- The use of several channels raises the probability that the user is able to perceive the information conveyed to him by the system. Visually impaired people in loud environments or deaf people in bad lighting conditions can process more data when output is presented redundantly to them.
- You don't need to abandon an output channel totally, only because it might be distorted. Visually impaired people might have problems reading a text and recognise each letter. After hearing the spoken variant and knowing what the text is about the visual representation can be used as memory hook. The same might be true for dark environments or mobile scenarios, when it is difficult to fix visual attention to the text.

- It is more likely to attract the user's attention when information is output via several channels, e.g. both audio and sight, than when only one output modality is used.

Rationale Independent disturbances of different channels rarely affect the same aspects of the content.

Multi-channel feedback of written and spoken text has proven to be effective for elderly (Emery et al. 2003) and visually impaired users, especially those suffering from *AMD*[2] (Vitense et al. 2002, Jacko et al. 2003, 2004, Edwards et al. 2004).

In the context of language understanding, it has been proven that users understand language better when they can read the lips of their interlocutor simultaneously to hearing (Sumby & Pollack 1954, Neely 1956, Binnie et al. 1974, Erber 1969, 1975, Summerfield 1979, Schomaker et al. 1995). This holds especially in loud environments.

Plosives ([p], [t], [k], [b], [d], [g]) sound similar and are likely to be confused when sound quality is low. At the same time these phones have distinctive lip shapes such as open lips (in the case of [g] and [k]) vs. initially closed lips (in the case of [b] and [p]). Lip shapes may differ for some similar sounding vowels, too.

Known Uses Some interactive systems such as *PPP* (André et al. 1996), *NUMACK* (Kopp et al. 2004), *COMIC* (Foster et al. 2005) and *SmartKom* ("Smartakus" Wahlster et al. 2001) display talking heads, in order to exploit the advantages of audiovisual language understanding (Benoît et al. 1998).

Mobile phones combine visual (blinking), auditive (ringing) and haptic (vibrating) signals in order to notify the user about phone calls or incoming short text messages.

Related Patterns Tidwell's *Important Message* (Tidwell 1999) is a concretisation of this pattern. In this case, information is output multimodally in order to attract the user's attention. Beyond attention attraction and raising perception probability, the generic pattern *Redundant Output* tries to overcome comprehension problems caused by context factors such as bad light or background noise.

[2] Age-related Macular Degeneration

Redundant Input

Context Communication channels might be unpredictably distorted due to bad lighting conditions, background noise, technical (network) problems or disabilities such as speech, motor or perception disorders.

Problem How to assure input when communication channels are distorted in an unforseeable way?

Forces

- The system can be configured or adapted to recognise and interpret that modality that is less affected by channel disorders, but in some cases all available interaction channels are distorted to some degree.
  - In loud environments users with motor disabilities or illiterate people have problems to interact with the system.
  - In dark environments or in hands-free scenarios (e.g. while carrying a bag, wandering around, driving a car) people with speech disorders have problems to input data.
  - In exerted conditions, both speech input and pen gestures are problematic.
- Interaction can be alleviated if passive modalities for data input (gaze input, free gestures) or authentication (voice recognition, face recognition) are used. However, these interaction channels are error prone so that they cannot be applied directly.

Solution Combine several interaction channels in order to make use of redundancy. Input coming from several channels (visual: e.g. lip movements, auditive: e.g. speech signal) should be interpreted in combination in order to reduce liability to errors.

Consequences

- The use of several channels raises the probability that the system is able to recognise and interpret the information input by the user in the desired way.
- Even if several channels are distorted the distortion rarely affects exactly the same pieces of information. Fusion mechanisms allow for reconstructing of at least some part of the input information.
- "Imperfect", error prone interaction channels can be combined to mutually disambiguate recognition errors.

Rationale Independent disturbances of different channels rarely affect the same aspects of the content. That's why for instance audio-visual speech recognition, which combines acoustic signals and lip movement analysis, leads to better recognition performance than unimodal speech recognition (cf. Benoît et al. 1998, S. 24 f.):

Plosives ([p], [t], [k], [b], [d], [g]) sound similar and are likely to be confused when sound quality is low. At the same time these phones have distinctive lip shapes such as open lips (in the case of [g] and [k]) vs. initially closed lips (in the case of [b] and [p]). Lip shapes may differ for some similar sounding vowels, too.

Channel distortions rarely affect both the recognition of a specific phoneme in the acoustic signal and of the corresponding viseme in the visual signal in the same way. Fusion algorithms allow to combine sound pieces of information from several channels such that some distorted parts can be reconstructed.

Studies conducted by (Oviatt 1999) revealed that an appropriate recogniser architecture that combines gesture and speech recognition can reduce recognition errors. This was shown for non-native speakers, in loud environments (Oviatt 2000a, b, c) and for exerted conditions (Kumar et al. 2004).

Known Uses This pattern is manifested in very different application areas including data input (audio-visual speech recognition), scene analysis (Wachsmuth 2001), person identification (Yang et al. 1998, 1999, Hazen et al. 2003, Jain 2003, Snelick et al. 2003), emotion recognition (Nasoz et al. 2002, Lisetti & Nasoz 2002, Busso et al. 2004, Gunes et al. 2004, Zeng et al. 2004) and the like. Following modality combinations are used:

- virtual reality and speech (Kaiser et al. 2003),
- gaze direction and speech (Zhang et al. 2004, Tan et al. 2003, Tanaka 1999, Campana et al. 2001),
- lip-reading in loud environments (Saenko et al. 2004), e.g. to filter out simultaneous speakers (Patterson & Gowdy 2003),
- speech and gesture (Holzapfel et al. 2004, Chai & Qu 2005),
- voice, ink and touchtone (Trabelsi et al. 2002),
- biometrics, voice and face to identify persons (Yang et al. 1999, Hazen et al. 2003).

Related Patterns *Spelling-based Hypothesis Reduction* as well as *Multimodal N-best Selection* are refinements of this pattern.

Multimodal N-best Selection

Context This pattern can be used in multimodal systems that offer speech input of unconstrained text or speech-based selection of items from very large sets such as timetable or navigation systems.

Problem Speech recognition is a statistical process. The recognition of input phrases results in a set of several recognition hypotheses. It is frequently the case that the original input phrase is included in the n-best set but does not coincide with the system's best estimate.

Forces

- When a speech input attempt fails the user can be prompted to repeat or to switch to another interaction modality. But it is inefficient to throw away input data which has failed the goal just by a hair.
- Playing back the *n* best recognition hypotheses, prompting for (spoken) selection and reducing the speech recognition vocabulary to this reduced list of *n* items can correct the error in just one further interaction step. However, items contained in the n-best list are likely to have some acoustic similarity so that they might be mixed up repeatedly by the recogniser.
- Playing back just the item on top of the *n* best recognition hypotheses and prompting for accepting or rejecting may resolve this problem in a few steps, but if the desired item is only the fifth (sixth, seventh ...) best recognition hypothesis five (six, seven ...) error corroboration steps are needed.

Solution Provide the user a means of selecting the correct result from a set of recognition hypotheses via pointing or key presses.

In order to satisfy cases where the desired item cannot be found in the n-best list there has to be a way of explicitly leaving the list selection dialog and to start over the input attempt.

Consequences

- Imperfect recognition results are not thrown away but reused in subsequent interaction steps.
- Recurring recognition problems due to confusability are avoided through alternative selection techniques. Instead of re-speaking the misrecognised item, the user can point to the item displayed in the list.
- Frequently, instead of endless error correction loops, this pattern helps to correct recognition errors in just one additional interaction step.

Rationale Suhm et al. (2001) point out that re-speaking the same word or phrase after recognition failure, although natural, is not the most promising form of error recovery in interactive systems. Changing the input modality to list selection seems to be more promising and has been suggested by Ainsworth & Pratt (1992) and Murray et al. (1993).

Known Uses Directory assistance, timetable information systems, speech-based driver assistance systems support n-best selection.

Related Patterns This pattern can be used in conjunction with *Speech-enabled Form*, *Drop-down Chooser* (Tidwell 2005) and *Autocompletion* (Tidwell 2005) to alleviate error handling in speech-enhanced input forms.

This pattern and *Spelling-based Hypothesis Reduction* are refinements of *Redundant Input*. *Multimodal N-best Selection* makes use of *Drop-down Chooser* (Tidwell 2005) to provide the user a non-speech alternative for selecting from a list of the *n* most likely recognition hypotheses and to avoid repeated errors.

Variant The solution described in this pattern is not restricted to strictly multimodal interaction. Speech-only systems provide similar speech-based approaches of selecting the desired item from a list of hypotheses.

If n-best selection via speech is supported, it is important to offer input modifications to avoid repeated errors. The user should be given the possibility to select the desired option via speaking the line number or re-speaking the item in combination with some distinctive attribute such as the first letter(s).

In these cases, the user has to be prompted in a way that reveals alternative selection strategies apart from simple re-speaking, e.g.: “Did you mean *one* Jonathan Smith, *two* John Griffith, *three* Joseph Reddish or *new input*”.

While the list is being read out the user should have the possibility to interrupt the playback. In full-duplex systems with cleanly separated channels for audio input and output barge-in can be used. That means that the system playback is stopped when the user starts speaking. In half-duplex systems the user should be able to interrupt system playback using a push-to-talk button. In the cases of both barge-in and push-to-talk the interruption moment can be used as a further information source to estimate the selected item (Balentine 1999, Balentine & Morgan 2001).

Spelling-based Hypothesis Reduction

Context Examples where this pattern can be used are systems that offer speech input of unconstrained text or speech-based selection of items from very large sets such as timetable or navigation systems. Errors are particularly likely in cases, when the user has to select from lists with similar sounding words or words with inconsistent pronunciation such as foreign names.

Problem Large recognition vocabularies entail error-prone recognition, especially when many similar sounding words have to be distinguished.

Forces

- Speech input can be used for selecting items from a list that cannot be displayed completely on a small screen, but if the list is too large for speech recognition or includes several similar sounding or problematic words, speech recognition is likely to fail. Problem areas are names in directory assistance systems as well as album and song titles in entertainment systems.
- In the case of recognition failure, the user can switch to text input (typing), but in some applications such as driver assistance systems only unhandy (if any) string input facilities are available.
- Typing the first letters can reduce the size of the selection list so that pointing is possible again, but in some applications such as navigation systems there might be a lot of characters to be input (using an impractical input device) before the list is reduced to a displayable size.
- Some speech recognisers, especially those for small devices, have only restricted resources such that only small vocabularies e.g. for number recognition or for recognising letters of the alphabet are supported. But operating applications this way only provides little added value in contrast to using a small keypad, especially since letters some of which sound similar are likely to be confused by the recogniser.

Solution Offer the user to input the first letters of the input tokens via typing. Use this substring to reduce the size of selectable items (i.e. of the speech recognition vocabulary) to head-matching strings. Using this vocabulary forthcoming speech inputs can be recognised more robustly.

Record speech input attempts and keep this recording available for some forthcoming interaction steps. Failed speech input can thus be reinterpreted afterwards successfully with reduced vocabulary.

Consequences

- By inputting quite few letters, the set of alternatives can be reduced to a size that – although still unsuited for pointing-based list selection – allows robust speech recognition.
- The user needs only to input few letters. This is important in cases where text input is inconvenient.

- There is no need to navigate and scroll through lists.
- Recognition of names, song titles and other problematic words becomes more robust.
- This technique of reducing speech recognition vocabularies simplifies speech recognition on platforms with limited resources.

Rationale The reduction of speech recognition vocabulary can improve recognition performance significantly.

Known Uses Marx & Schmandt (1994) describe the messaging system *Chatter* which allows the user to input contact names via voice spelling, touch-tone spelling and speech-based naming in a combined fashion.

The prototype of a multimodal driver assistance system by Neuss (2001) allows the user to input the first letters using a rotary knob mounted on the centre console, so that the speech recognition vocabulary can be reduced.

Other examples that go in the same direction can be found in Suhm et al. (2001) and Tan et al. (2003).

Related Patterns This pattern can be used in conjunction with *Speech-enabled Form*, *Drop-down Chooser* (Tidwell 2005) and *Autocompletion* (Tidwell 2005) to alleviate error handling in speech-enabled input forms.

This pattern and *Multimodal N-best Selection* are refinements of *Redundant Input*. *Spelling-based Hypothesis Reduction* uses (or is) some kind of variant of *Continuous Filter* (van Welie & Trætteberg 2000). Instead of filtering the items of a selection list, the recognition vocabulary is reduced according to the letters input by the user.

Variant Some systems allow the user to dictate characters (voice spelling) in a purely speech-based way to reduce the recognition vocabulary.

In this case, phonetic alphabets can be used to reduce recognition failures. But this is only feasible if the target user group is expected to be proficient in that phonetic alphabet.

When a phonetic alphabet is supported the user is not proficient in, this might result in spontaneous wordings such as *Motel* instead of *Mike* or *October* instead of *Oscar*. This way, recognition errors might even increase.

5 Related Patterns from other Collections

5.1 Multimodal User Interface Patterns

This is only a fraction of the patterns identified during this work. Other patterns focus on fast / efficient interaction. The following table outlines only the pattern *Speech-enabled Form* (cf. Ratzka 2008), since it is referenced several times in this paper:

Name	**Problem**	**Solution**
Speech-enabled Form	How to simplify string input in form filling applications?	Let the user select the desired form field via pointing and input the desired value via speech.

5.2 Related Patterns from other Authors

These patterns for multi-modal interaction are closely related to traditional user interface patterns described by other authors:

Name Reference	**Problem**	**Solution**
Continuous Filter (van Welie & Trætteberg 2000)	“The user needs to find an item in an ordered set”.	“Provide a filter component with which the user can in real time filter only the items in the data that are of interest.”
Important Message (Tidwell 1999)	“How should the artefacts convey this information to the user?”.	“Interrupt whatever the user is doing with the message, using both sight and sound if possible.”
Form (Tidwell 1999, Sinnig et al. 2004), (www.welie.com)	“The user must provide structural textual information to the application. The data to be provided is logically related”	“Provide users with a form containing the necessary elements. Forms contain basically a set of input interaction elements and are a means of collecting information [...]”.
Drop-down Chooser (Tidwell 2005)	“The user needs to supply input that is a choice from a set [...], a date or time, a number, or anything other than free text typed at a keyboard. [...]”.	“For the Drop-down Chooser control’s ‘closed’ state, show the current value of the control in either a button or a text field. To its right, put a down arrow. [...] A click on the arrow (or the whole control) brings up the chooser panel, and a second click closes it again [...]”.
Autocompletion (Tidwell 2005)	“The user types something predictable, such as a URL, the user’s own name or address, today’s date, or a filename [...]”.	“With each additional character that the user types, the software quietly forms a list of the possible completions to that partially entered string [...]”.

6 Conclusion

This paper outlines an emerging pattern language for multimodal interaction which is far from being complete. Despite the research history of over twenty years, multimodality is still a research-centric field. It begins to reach some dissemination in the fields of automotive, industrial and mobile applications. That is why interaction design support is needed. Interaction design patterns constitute a challenging and exciting approach to this domain.

7 Acknowledgements

I would like to thank my shepherd Michael vanHilst and the members of the workshop group at EuroPLoP 2008 for the numerous suggestions and valuable feedback.

References

Ainsworth & Pratt 1992
AINSWORTH, W.A.; PRATT, S.R.: Feedback strategies for error correction in speech recognition systems. In: *Int J. Man-Mach. Stud.* 36 (1992), Jun., Nr. 6, S. 833–842

André et al. 1996
ANDRÉ, E.; MULLER, A. J.; RIST, T.: The PPP Persona: A Multipurpose Animated Presentation Agent. In: AL., Catarci et (ed.): *Advanced Visual Interfaces*, ACM Press, 1996, S. 245–247

Balentine 1999
BALENTINE, B.: Re-Engineering The Speech menu. In: GARDNER-BONNEAU, D. (ed.): *Human Factors and Voice Interacive Systems*. Norwell, Massachussetts: Kluwer Akademic Publishers, 1999, S. 205–235

Balentine & Morgan 2001
BALENTINE, Bruce; MORGAN, D. P.: *How to Build a Speech Recognition Application. A Style Guide for Telephony Dialogues*. EIG Press, 2001

Benoît et al. 1998
BENOÎT, C.; MARTIN, J.-C.; PELACHAUD, C.; SCHOMAKER, L.; SUHM, B.: Audio-Visual and Multimodal Speech Systems. Version: 1998. http://www.limsi.fr/Individu/martin/publications/download/chmultimodal.ps. In: GIBBON, D. (ed.): *Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume*. 1998

Bernsen 1999
BERNSEN, Niels O.: *Multimodality in Language and Speech Systems - from theory to design support tool.* Lectures at the 7th European Summer School on Language and Speech Communication (ESSLSC). http://www.nis.sdu.dk/ nob/stockholm.zip. Version: Juli 1999, retrieved on: 20.06.2008

Binnie et al. 1974
BINNIE, C. A.; MONTGOMERY, A. A.; JACKSON, P. L.: Auditory and visual contributions to the perception of consonants. In: *Journal of Speech & Hearing Research* 17 (1974), S. 619–630

Bürgy 2002
BÜRGY, Christian: *An Interaction Constraints Model for Mobile and Wearable Computer-Aided Engineering Systems in Industrial Applications*, Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, Diss., 2002

Busso et al. 2004
BUSSO, C.; DENG, Z.; YILDIRIM, S.; BULUT, M.; LEE, C.-M.; KAZEMZADEH, A.; LEE, S.; NEUMANN, U.; NARAYANAN, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2004, S. 205–211

Campana et al. 2001
CAMPANA, E.; BALDRIDGE, J.; DOWDING, J.; HOCKEY, B.-A.; REMINGTON, R.-W.; STONE, L.-S.: Using eye movements to determine referents in a spoken dialogue system. In: *PUI '01: Proceedings of the 2001 workshop on Perceptive user interfaces*. New York, NY, USA: ACM Press, 2001, S. 1–5

Chai & Qu 2005
CHAI, Joyce Y.; QU, Shaolin: A salience driven approach to robust input interpretation in multimodal conversational systems. In: *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2005, S. 217–224

Comerford et al. 2001
COMERFORD, L.; FRANK, D.; GOPALAKRISHNAN, P.; GOPINATH, R.; SEDIVY, J.: The IBM Personal Speech Assistant. In: *Proc. of IEEE ICASSP'01* DARPA, 2001, S. 319–321

Dragičević 2004
DRAGIČEVIĆ, Pierre: *Un modèle d'interaction en entrée pour des systèmes interactifs multi-dispositifs hautement configurables*, Université de Nantes, école doctorale sciences et technologies de l'information et des matérieaux, Diss., Mars 2004. http://www.dgp.toronto.edu/˜dragice/these/html/memoire_dragicevic.html, retrieved on: 20.06.2008

Duarte & Carriço 2006
DUARTE, C.; CARRIÇO, L.: A conceptual framework for developing adaptive multimodal applications. In: *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2006, S. 132–139

Edwards et al. 2004
Edwards, P. J.; Barnard, L.; Emery, V. K.; Yi, J. S.; Moloney, K. P.; Kongnakorn, T.; Jacko, J. A.; Sainfort, F.; Oliver, P. R.; Pizzimenti, J.; Bade, A.; Fecho, G.; Shallo-Hoffmann, J.: Strategic design for users with diabetic retinopathy: factors influencing performance in a menu-selection task. In: *Assets '04: Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM Press, 2004, S. 118–125

Emery et al. 2003
Emery, V. K.; Edwards, P. J.; Jacko, J. A.; Moloney, K. P.; Barnard, L.; Kongnakorn, T.; Sainfort, F.; Scott, I. U.: Toward achieving universal usability for older adults through multimodal feedback. In: *CUU '03: Proceedings of the 2003 conference on Universal usability*. New York, NY, USA: ACM Press, 2003, S. 46–53

Erber 1969
Erber, N. P.: Interaction of audition and vision in the recognition of oral speech stimuli. In: *Journal of Speech & Hearing Research* 12 (1969), S. 423–425

Erber 1975
Erber, N. P.: Auditory-visual perception of speech. In: *Journal of Speech & Hearing Disorders* 40 (1975), S. 481–492

Foster et al. 2005
Foster, M. E.; White, M.; Setzer, A.; Catizone, R.: Multimodal generation in the COMIC dialogue system. In: *ACL '05: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*. Morristown, NJ, USA: Association for Computational Linguistics, 2005, S. 45–48

Gunes et al. 2004
Gunes, Hatice; Piccardi, Massimo; Jan, Tony: Face and body gesture recognition for a vision-based multimodal analyzer. In: *VIP '05: Proceedings of the Pan-Sydney area workshop on Visual information processing*. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2004, 19–28

Hazen et al. 2003
Hazen, Timothy J.; Weinstein, Eugene; Park, Alex: Towards robust person recognition on handheld devices using face and speaker identification technologies. In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 289–292

Holzapfel et al. 2004
Holzapfel, Hartwig; Nickel, Kai; Stiefelhagen, Rainer: Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2004, S. 175–182

Ibrahim & Johansson 2002
Ibrahim, Aseel; Johansson, Pontus: Multimodal dialogue systems: A case study for interactive tv. In: *Proceedings of 7th ERCIM Workshop on User Interfaces for All*. Chantilly, France, 2002, 209–218

Jacko et al. 2004
Jacko, J. A.; Barnard, L.; Kongnakorn, T.; Moloney, K. P.; Edwards, P. J.; Emery, V. K.; Sainfort, F.: Isolating the effects of visual impairment: exploring the effect of AMD on the utility of multimodal feedback. In: *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM Press, 2004, S. 311–318

Jacko et al. 2003
Jacko, J. A.; Scott, I. U.; Sainfort, F.; Barnard, L.; Edwards, P. J.; Emery, V. K.; Kongnakorn, T.; Moloney, K. P.; Zorich, B. S.: Older adults and visual impairment: what do exposure times and accuracy tell us about performance gains associated with multimodal feedback? In: *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM Press, 2003, S. 33–40

Jain 2003
Jain, Anil K.: Multimodal user interfaces: who's the user? In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 1–1

Kaiser et al. 2003
Kaiser, E.; Olwal, A.; McGee, D.; Benko, H.; Corradini, A.; Li, X.; Cohen, P.; Feiner, S.: Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 12–19

Kopp et al. 2004
Kopp, Stefan; Tepper, Paul; Cassell, Justine: Towards integrated microplanning of language and iconic gesture for multimodal output. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2004, S. 97–104

Kumar et al. 2004
Kumar, Sanjeev; Cohen, Philip R.; Coulston, Rachel: Multimodal interaction under exerted conditions in a natural field setting. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2004, S. 227–234

Lisetti & Nasoz 2002
LISETTI, Christine L.; NASOZ, Fatma: MAUI: a multimodal affective user interface. In: *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 2002, S. 161–170

Malaka et al. 2004
MALAKA, Rainer; HÄUSSLER, Jochen; ARAS, Hidir: SmartKom mobile: intelligent ubiquitous user interaction. In: *IUI '04: Proceedings of the 9th international conference on Intelligent user interface*. New York, NY, USA: ACM Press, 2004, S. 310–312

Marx & Schmandt 1994
MARX, Matt; SCHMANDT, Chris: Putting people first: specifying proper names in speech interfaces. In: *UIST '94: Proceedings of the 7th annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM Press, 1994, S. 29–37

Mayhew 1999
MAYHEW, D. J.: *The Usability Engineering Lifecycle*. San Francisco: Morgan Kaufmann, 1999

McCaffery et al. 1998
MCCAFFERY, Fergal; MCTEAR, Michael F.; MURPHY, Maureen: A Multimedia Interface for Circuit Board Assembly. In: *Multimodal Human-Computer Communication, Systems, Techniques, and Experiments*. London, UK: Springer-Verlag, 1998, 213–230

Microsoft
Microsoft: *MiPad: Speech Powered Prototype to Simplify Communication Between Users and Handheld Devices*. `http://www.microsoft.com/presspass/features/2000/05-22mipad.asp`, retrieved on: 20.06.2008

Miyazaki 2002
MIYAZAKI, J.: *Discussion Board System with modality variation: From Multimodality to User Freedom*, Tampere University, Masterarbeit, 2002

Murray et al. 1993
MURRAY, A. C.; FRANKISH, C. R.; JONES, D. M.: Data-entry by voice: Facilitating correction of misrecognitions. In: BABER, C. (ed.); NOYES, J.M. (ed.): *Interactive Speech Technology: Human Factors issues in the Application of Speech Input/Output to Computers*. Bristol, PA: Taylor and Francis, 1993, S. 137–144

Nasoz et al. 2002
NASOZ, Fatma; OZYER, Onur; LISETTI, Christine L.; FINKELSTEIN, Neal: Multimodal affective driver interfaces for future cars. In: *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 2002, S. 319–322

Neely 1956
NEELY, K. K.: Effect of visual factors on the intelligibility of speech. In: *Journal of the Acoustical Society of America* 28 (1956), S. 1275–1277

Neuss 2001
NEUSS, Robert: *Usability Engineering als Ansatz zum Multimodalen Mensch-Maschine-Dialog*, Fakultät für Elektrotechnik und Informationstechnik, Technische Universität München, Diss., 2001

Niedermaier 2003
NIEDERMAIER, Franz B.: *Entwicklung und Bewertung eines Rapid-Prototyping Ansatzes zur multimodalen Mensch-Maschine-Interaktion im Kraftfahrzeug*, Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München, Diss., 2003

Obrenović et al. 2007
OBRENOVIĆ, Z.; ABASCAL, J.; STARČEVIĆ, D.: Universal accessibility as a multimodal design issue. In: *Commun. ACM* 50 (2007), Nr. 5, S. 83–88

Oviatt et al. 1999
OVIATT, S.; BERNARD, J.; LEVOW, G.: Linguistic adaptation during error resolution with spoken and multimodal systems. In: *Language and Speech (special issue on Prosody and Conversation)* 41 (1999), Nr. 3–4, S. 415–438

Oviatt et al. 2000
OVIATT, S.; COHEN, P.; WU, L.; VERGO, J.; DUNCAN, L.; SUHM, B.; BERS, J.; HOLZMAN, T.; WINOGRAD, T.; LANDAY, J.; LARSON, J.; FERRO, D.: Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions. In: *Human Computer Interaction* 15 (2000), Nr. 4, S. 263–322

Oviatt & vanGent 1996
OVIATT, S.; VANGENT, R.: Error reslution during multimodal human-computer interaction. In: *Proc. of the International Conference on Spoken Language Processing* Bd. 2, 1996, S. 204–207

Oviatt 1999
OVIATT, Sharon L.: Mutual disambiguation of recognition errors in a multimodal architecture. In: *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 1999, S. 576–583

Oviatt 2000a
OVIATT, Sharon L.: Multimodal Signal Processing in Naturalistic Noisy Environments. In: YUAN, B. (ed.); HUANG, T. (ed.); TANG, X. (ed.): *Proceedings of the 6th International Conference on SPoken Language Processing (ICSLP)* Bd. 2. Peking: Chinese Friendship Publishers, 2000, 696–699

Oviatt 2000b
OVIATT, Sharon L.: Multimodal system processing in mobile environments. In: *UIST '00: Proceedings of the 13th annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM Press, 2000, S. 21–30

Oviatt 2000c
OVIATT, Sharon L.: Taming recognition errors with a multimodal interface. In: *Commun. ACM* 43 (2000), Nr. 9, S. 45–51

Oviatt & Kuhn 1998
OVIATT, Sharon L.; KUHN, Karen: Referential features and linguistic indirection in multimodal language. In: *Proceedings of the International Conference on Spoken Language Processing* Bd. 6, ASSTA, 1998, 2339–2342

Patterson & Gowdy 2003
PATTERSON, E.K.; GOWDY, J.N.: An audio-visual approach to simultaneous-speaker speech recognition. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* Bd. 5, 2003, 780–783

Pieraccini et al. 2004
PIERACCINI, R.; DAYANIDHI, K.; BLOOM, J.; DAHAN, J.-G.; PHILLIPS, M.; GOODMAN, B. R.; PRASAD, K. V.: Multimodal conversational systems for automobiles. In: *Commun. ACM* 47 (2004), Nr. 1, S. 47–49

Ratzka 2008
RATZKA, Andreas: Design Patterns in the Context of Multi-modal Interaction. In: *To appear in: Proceedings of the 6th Nordic Conference on Pattern Languages of Programs 2007 VikingPLoP 2007*, 2008

Ratzka & Wolff 2006
RATZKA, Andreas; WOLFF, Christian: A Pattern-based Methodology for Multimodal Interaction Design. In: SOJKA, P. (ed.); KOPEČEK, I. (ed.); PALA, K. (ed.): *Proc. of Text, Speech, and Dialogue, TSD'06*. Berlin, Heidelberg: Springer, 2006 (LNAI 4188), S. 677–686

Reithinger et al. 2003
REITHINGER, N.; ALEXANDERSSON, J.; BECKER, T.; BLOCHER, A.; ENGEL, R.; LÖCKELT, M.; MÜLLER, J.; PFLEGER, N.; POLLER, P.; STREIT, M.; TSCHERNOMAS, V.: SmartKom: adaptive and flexible multimodal access to multiple applications. In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 101–108

Saenko et al. 2004
SAENKO, Kate; DARRELL, Trevor; GLASS, James R.: Articulatory features for robust visual speech recognition. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2004, S. 152–158

Schomaker et al. 1995
SCHOMAKER, L.; NIJTMANS, J.; CAMURRI, A.; LAVAGETTO, F.; MORASSO, P.; BENOÎT, C.; GUIARD-MARIGNY, T.; GOFF, B. L.; ROBERT-RIBES, J.; ADJOUDANI, A.; DEFÉE, I.; MÜNCH, S.; HARTUNG, K.; BLAUERT, J.: A Taxonomy of Multimodal Interaction in the Human Information Processing System. Version: Februar 1995. `http://vonkje.cogsci.kun.nl/ miami/reports/reports.html`. 1995. (A Report of the Esprit Basic Research Action 8579 MIAMI). – Forschungsbericht

Sinnig et al. 2004
SINNIG, Daniel; GAFFAR, Ashraf; REICHART, Daniel; SEFFAH, Ahmed; FORBRIG, Peter: Patterns in Model-Based Engineering. In: *CADUI*, 2004, S. 195–208

Snelick et al. 2003
SNELICK, Robert; INDOVINA, Mike; YEN, James; MINK, Alan: Multimodal biometrics: issues in design and testing. In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 68–72

Suhm et al. 2001
SUHM, B.; MYERS, B.; WAIBEL, A.: Multimodal error correction for speech user interfaces. In: *ACM Trans. Comput.-Hum. Interact.* 8 (2001), Nr. 1, S. 60–98

Sumby & Pollack 1954
SUMBY, W. H.; POLLACK, I.: Visual contribution to speech intelligibility in noise. In: *Journal of the Acoustical Society of America* 26 (1954), S. 212–215

Summerfield 1979
SUMMERFIELD, A. Q.: Use of visual information for phonetic perception. In: *Phonetica* 36 (1979), S. 314–331

Tan et al. 2003
TAN, Yeow K.; SHERKAT, Nasser; ALLEN, Tony: Error recovery in a blended style eye gaze and speech interface. In: *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, S. 196–202

Tanaka 1999
TANAKA, Katsumi: A robust selection system using real-time multi-modal user-agent interactions. In: *IUI '99: Proceedings of the 4th international conference on Intelligent user interfaces.* New York, NY, USA: ACM Press, 1999, S. 105–108

Tidwell 1999
TIDWELL, J.: *COMMON GROUND: A Pattern Language for Human-Computer Interface Design.* `http://www.mit.edu/~jtidwell/common_ground.html`. Version: 1999, retrieved on: 20.06.2008

Tidwell 2005
TIDWELL, Jenifer: *Designing Interfaces: Patterns for Effective Interaction Design.* O'Reilly, 2005

Trabelsi et al. 2002
TRABELSI, Z.; CHA, S.-H.; DESAI, D.; TAPPERT, C.: A voice and ink XML multimodal architecture for mobile e-commerce systems. In: *WMC '02: Proceedings of the 2nd international workshop on Mobile commerce.* New York, NY, USA: ACM Press, 2002, S. 100–104

Vitense et al. 2002
VITENSE, H. S.; JACKO, J. A.; EMERY, V. K.: Multimodal feedback: establishing a performance baseline for improved access by individuals with visual impairments. In: *Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies.* New York, NY, USA: ACM Press, 2002, S. 49–56

Wachsmuth 2001
WACHSMUTH, Sven: *Multi-modal Scene Understanding Using Probabilistic Models*, Technischen Fakultät, Universität Bielefeld, Diss., 2001

Wahlster et al. 2001
WAHLSTER, W.; REITHINGER, N.; BLOCHER, A.: SmartKom: Towards Multimodal Dialogues with Anthropomorphic Interface Agents. In: WOLF, G. (ed.); KLEIN, G. (ed.); Projektträger des BMBF für Informationstechnik: Deutsches Zentrum für Luft- und Raumfahrttechnik (DLR) e.V. (Veranst.): *Proceedings of International Status Conference: Lead Projects Human-Computer-Interaction.* Saarbrücken, 2001, 23–32

Wasinger 2006
WASINGER, Rainer: *Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations.* Saarbrücken, Naturwissenschaftlich-Technische Fakultät I der Universität des Saarlandes, Diss., 2006

van Welie & Trætteberg 2000
WELIE, M. van; TRÆTTEBERG, H.: Interaction Patterns in User Interfaces. In: *Proceedings of the Seventh Pattern Languages of Programs Conference.* Monticello, Illinois, USA, 2000

Yang et al. 1998
YANG, J.; STIEFELHAGEN, R.; MEIER, U.; WAIBEL, A.: Visual tracking for multimodal human computer interaction. In: *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems.* New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1998, S. 140–147

Yang et al. 1999
YANG, J.; ZHU, X.; GROSS, R.; KOMINEK, J.; PAN, Y.; WAIBEL, A.: Multimodal people ID for a multimedia meeting browser. In: *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1).* New York, NY, USA: ACM Press, 1999, S. 159–168

Zeng et al. 2004
ZENG, Z.; TU, J.; LIU, M.; ZHANG, T.; RIZZOLO, N.; ZHANG, Z.; HUANG, T. S.; ROTH, D.; LEVINSON, S.: Bimodal HCI-related affect recognition. In: *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces.* New York, NY, USA: ACM Press, 2004, S. 137–143

Zhang et al. 2004
ZHANG, Q.; IMAMIYA, A.; GO, K.; MAO, X.: Overriding errors in a speech and gaze multimodal architecture. In: *IUI '04: Proceedings of the 9th international conference on Intelligent user interface.* New York, NY, USA: ACM Press, 2004, S. 346–348