Multi-Modal User Interaction

Lecture 5: Multimodal Fusion and Design

Zheng-Hua Tan

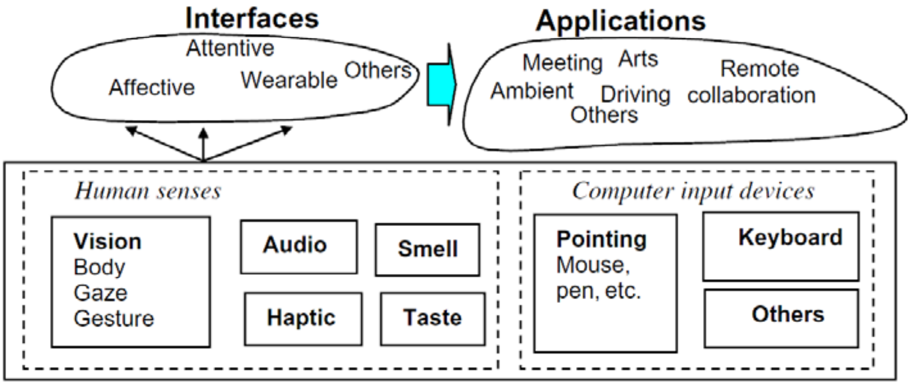
Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk

Part I: Multimodal interaction design

- Multimodal interaction design
- Multimodal fusion
- Decision-level fusion and combining classifiers
- Design guidelines

Multimodal interaction

A human-centered approach



(Jaimes & Sebe, 2007)

Multimodal interaction design

- More than 2 modes –e.g. spoken, gestural, facial expression, gaze; various sensors
- Inputs are uncertain –vs. Keyboard/mouse
  - Corrupted by noise
  - Multiple people
- Recognition is probabilistic
- Meaning is ambiguous

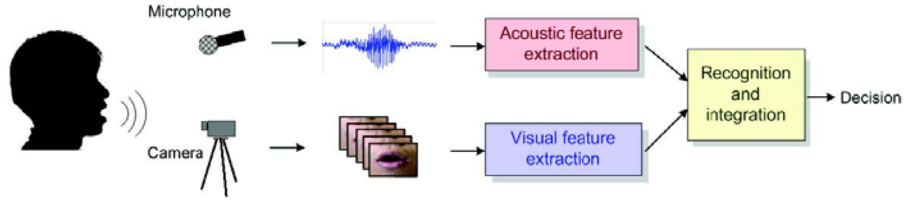
- Design for uncertainty

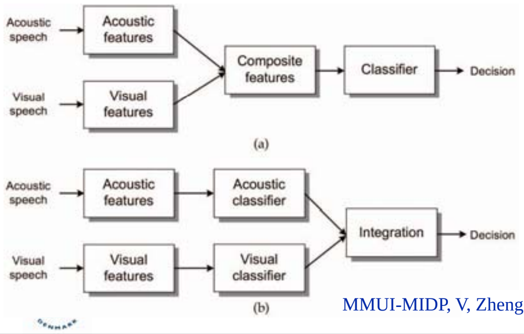
(Cohen, 2005)

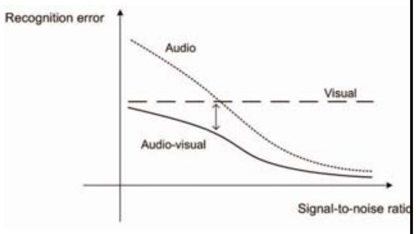
Approach to gaining robustness

- Fusion of inputs from multiple modalities
- Using strengths of one mode to compensate for weaknesses of others—design time and run time
- Avoiding/correcting errors
- Statistical architecture
- Confirmation
- Dialogue context

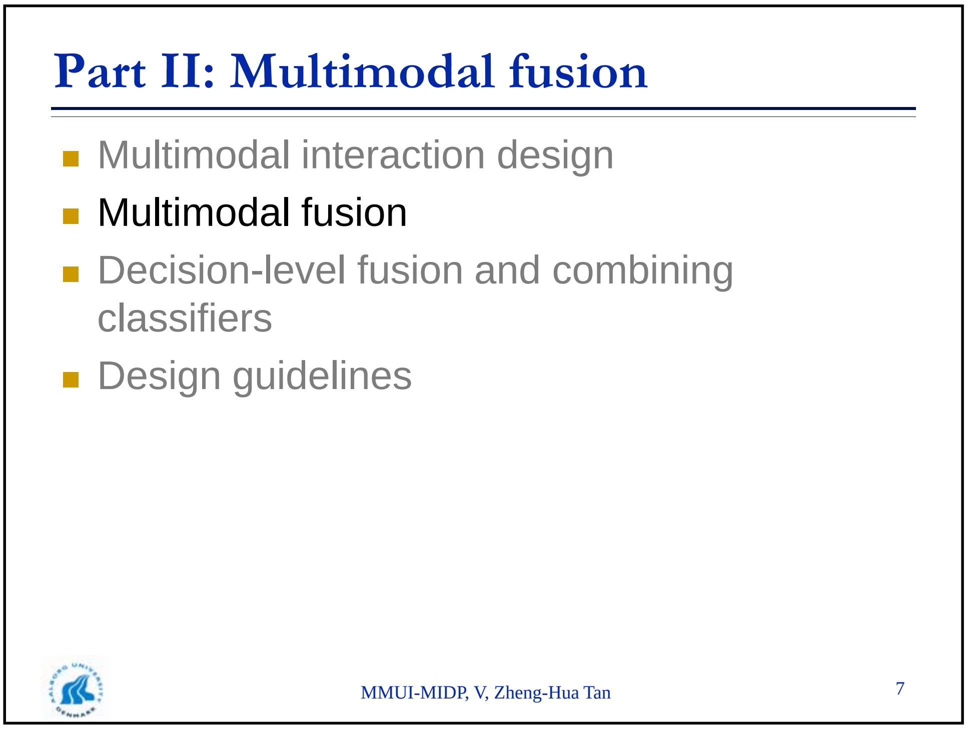
Challenges and fusion approaches


Microphone
Camera
Acoustic feature extraction
Visual feature extraction
Recognition and integration
Decision



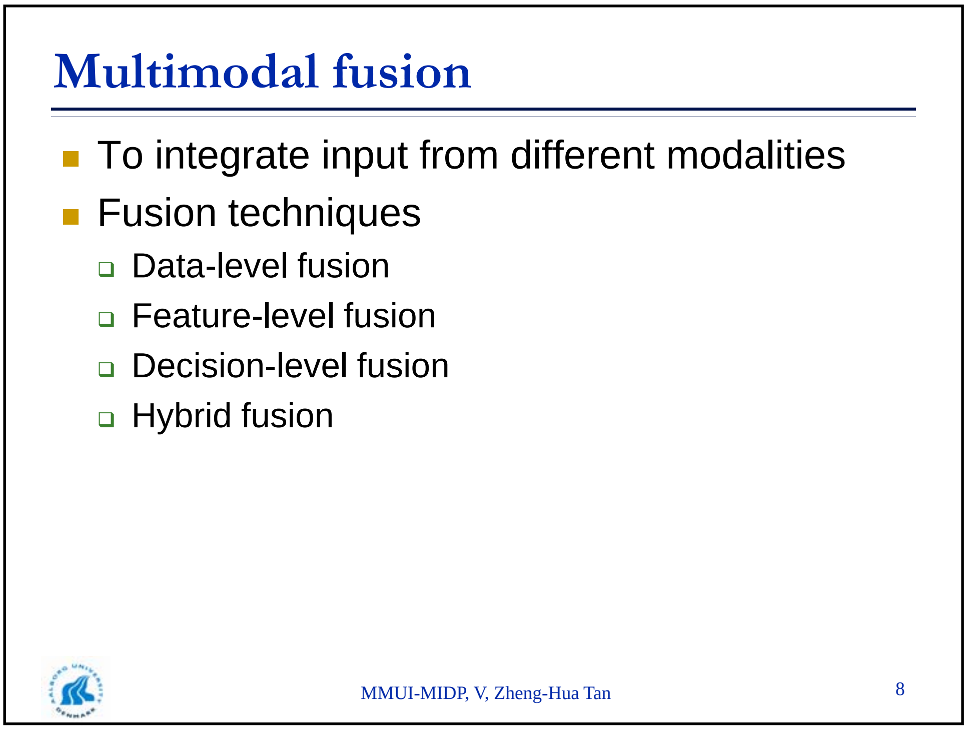


(Lee & Park 2008)



Part II: Multimodal fusion

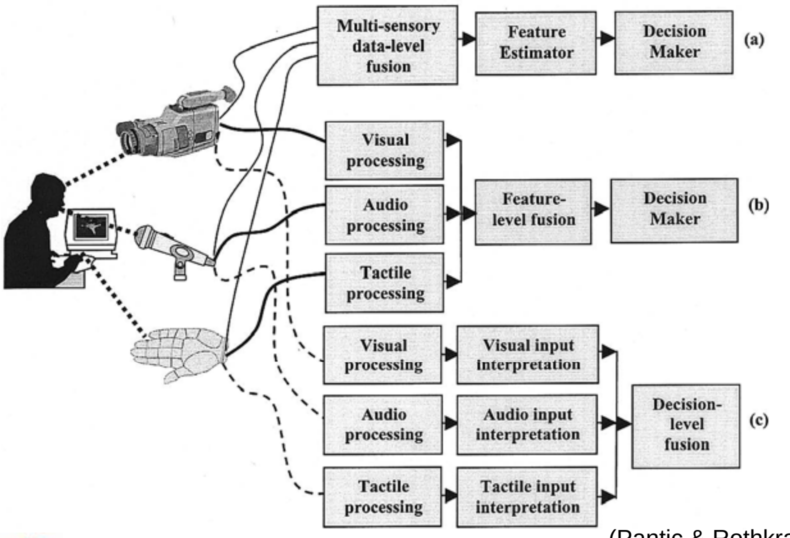
- Multimodal interaction design
- Multimodal fusion
- Decision-level fusion and combining classifiers
- Design guidelines



Multimodal fusion

- To integrate input from different modalities
- Fusion techniques
  - Data-level fusion
  - Feature-level fusion
  - Decision-level fusion
  - Hybrid fusion

Fusion of multiple sensing modalities



(Pantic & Rothkrantz, 2003)

Data-level fusion

- Data-level fusion involves integration of raw sensory observations and can be accomplished only when the observations are of the same type.
- Since the monitored human interactive signals are of different nature and are sensed using different types of sensors, data-level fusion is, in principle, not applicable to multimodal HCI.

Feature-level fusion

- Each stream of sensory information is first analyzed for features and then the detected features are fused, e.g. through *feature concatenation*.

$$\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}\ ,\ \mathbf{o}_{v,t}] \in \mathrm{R}^{l_{av}},\ \text{where } l_{av} = l_a + l_v$$

- Dimensionality reduction techniques such as LDA are usually applied, so called *discriminant feature fusion*.

$$\mathbf{o}_{d,t} = \mathbf{o}_{av,t}\, \mathbf{L}_{av} \qquad \in \mathrm{R}^{l_d}$$

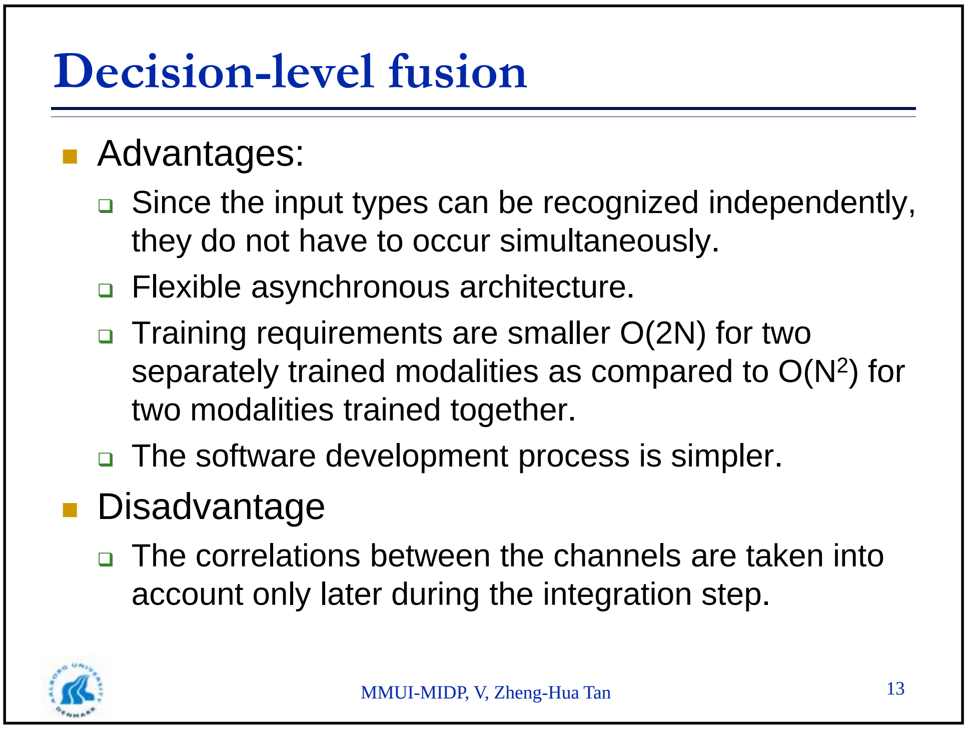
- Using a single classifier avoiding the explicit modeling of different modalities
- The signal-level recognition process in one mode influences the course of recognition in the other.
- More appropriate for closely temporally synchronized input modalities, such as speech and lip movements.

Decision-level fusion

- Aka, later fusion, late intergration
- Integrates common meaning representations derived from different modalities into a combined final interpretation.
- Utilizes independent classifiers, one for each stream, which can be trained independently. The final decision is reached by combining the partial outputs of the unimodal classifiers.

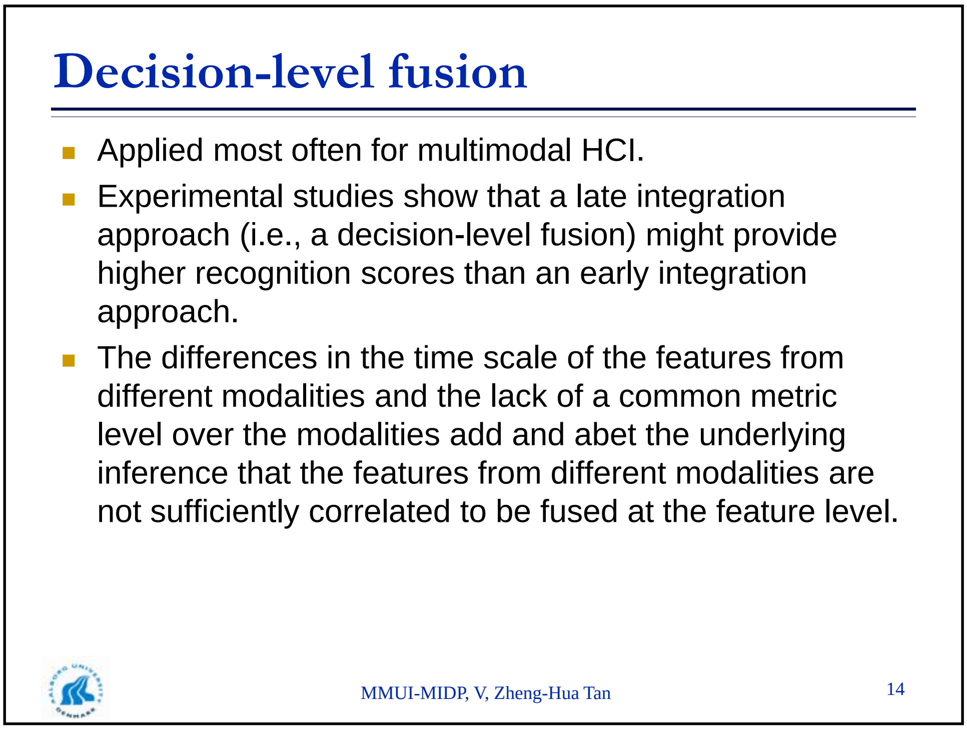
$$C^* = \arg\max_i \left\{ \gamma \log P(O_A \mid \lambda_A^i) + (1-\gamma) \log P(O_V \mid \lambda_V^i) \right\}$$

- Requires a common meaning representation framework for all modalities used and a well-defined operation for integrating partial meanings.



Decision-level fusion

- Advantages:
  - Since the input types can be recognized independently, they do not have to occur simultaneously.
  - Flexible asynchronous architecture.
  - Training requirements are smaller O(2N) for two separately trained modalities as compared to $O(N^2)$ for two modalities trained together.
  - The software development process is simpler.
- Disadvantage
  - The correlations between the channels are taken into account only later during the integration step.



Decision-level fusion

- Applied most often for multimodal HCI.
- Experimental studies show that a late integration approach (i.e., a decision-level fusion) might provide higher recognition scores than an early integration approach.
- The differences in the time scale of the features from different modalities and the lack of a common metric level over the modalities add and abet the underlying inference that the features from different modalities are not sufficiently correlated to be fused at the feature level.

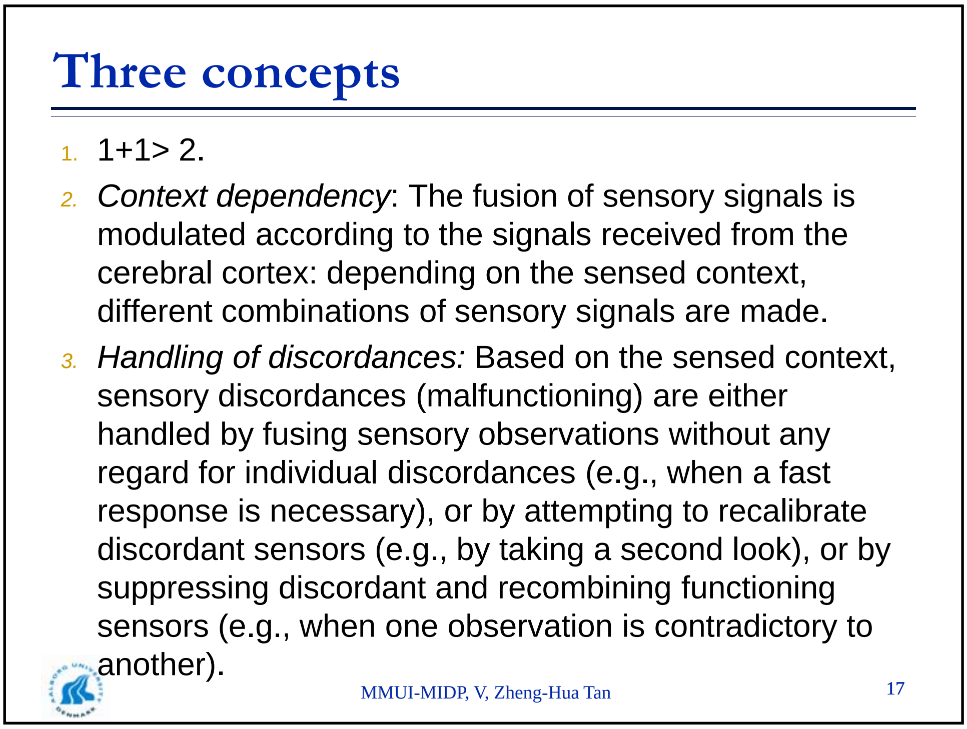
Hybrid fusion

- Problems with early and late fusion techniques:
  - Early fusion techniques fail to model both the fluctuations in the relative reliability and the asynchrony problems between the distinct streams.
  - Human display audio, visual and tactile interactive signals in a complementary and redundatn manner. Input isgnals cannot be considered mutually independent and cannot be combined only at the end of the intended analysis.
- Hybrid fusion
  - combining feature and decision fusion within the framework of the latter

Three concepts from neurological studies

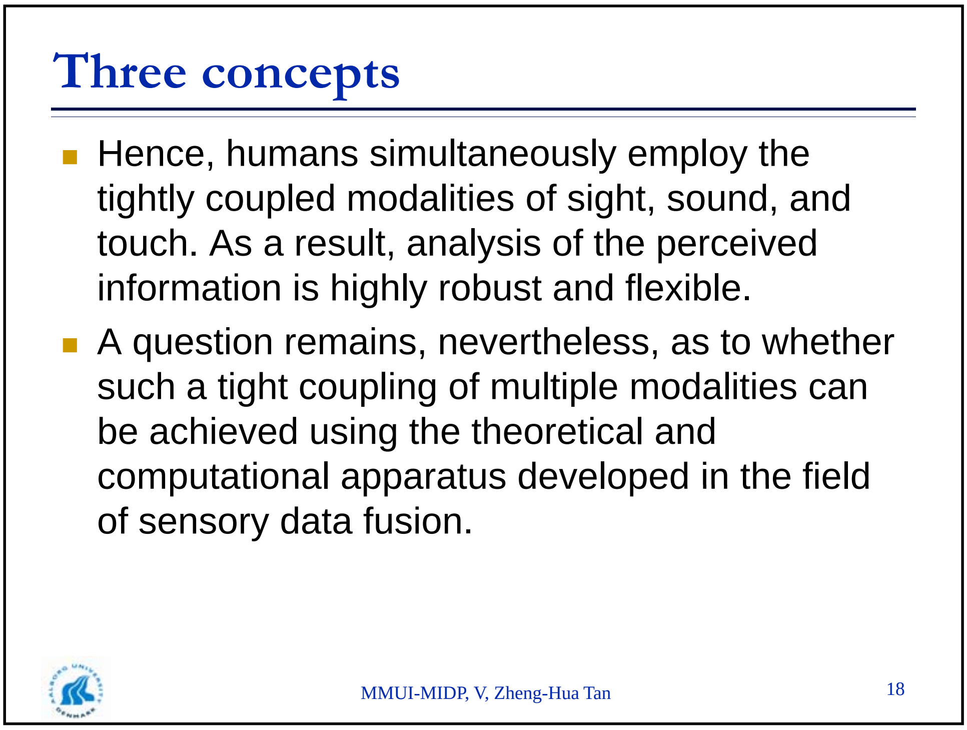
Insight into how the modalities of sight, sound, and touch are combined in human–human interaction can be gained from neurological studies on fusion of sensory neurons:

1. 1+1> 2: The response of multisensory neurons can be stronger for multiple weak input sensory signals than for a single strong signal.
2. *Context dependency*.
3. *Handling of discordances*.



Three concepts

1. 1+1> 2.
2. *Context dependency*: The fusion of sensory signals is modulated according to the signals received from the cerebral cortex: depending on the sensed context, different combinations of sensory signals are made.
3. *Handling of discordances:* Based on the sensed context, sensory discordances (malfunctioning) are either handled by fusing sensory observations without any regard for individual discordances (e.g., when a fast response is necessary), or by attempting to recalibrate discordant sensors (e.g., by taking a second look), or by suppressing discordant and recombining functioning sensors (e.g., when one observation is contradictory to another).

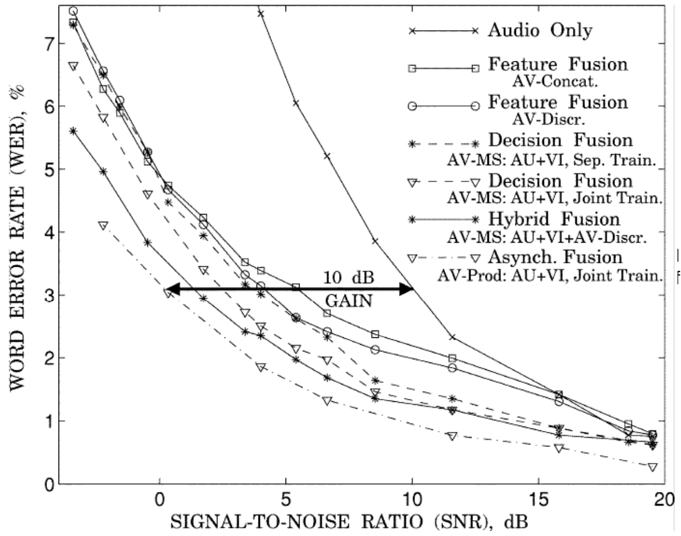


Three concepts

- Hence, humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible.
- A question remains, nevertheless, as to whether such a tight coupling of multiple modalities can be achieved using the theoretical and computational apparatus developed in the field of sensory data fusion.

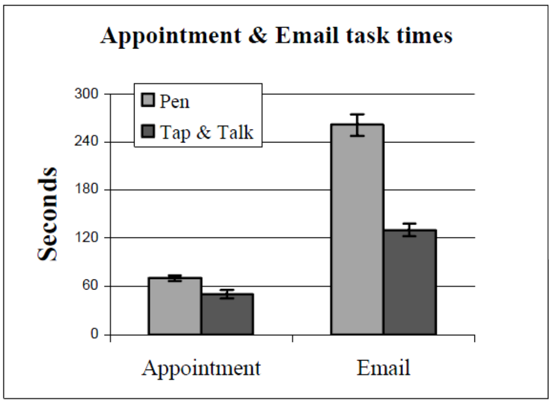
Benefits

Audio-only and audiovisual WER % (Pomianos et al., 2003)



Intermediate fusion

Benefits



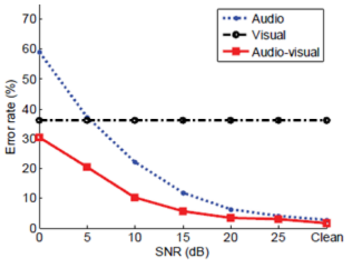
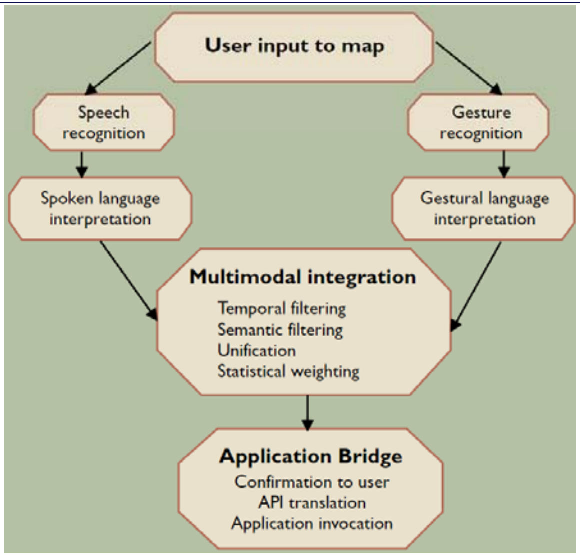


Figure 5 Task completion time of email transcription between the pen-only interface and *Tap and Talk* interface. The standard deviation is also shown above the bar of each performed task.

Part III: DL fusion and combining classifiers

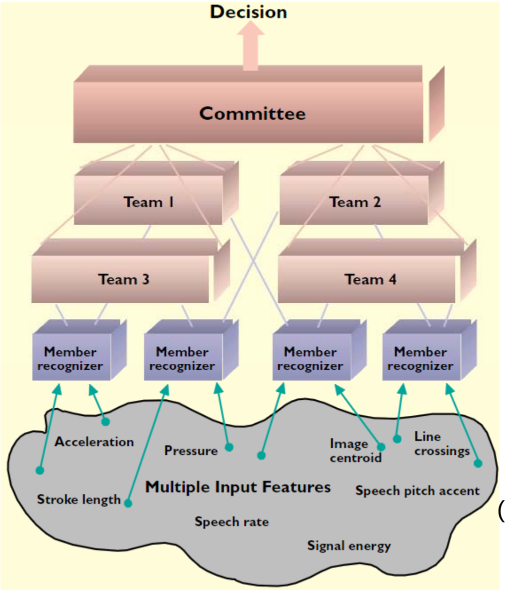
- Multimodal interaction design
- Multimodal fusion
- Decision-level fusion and combining classifiers
- Design guidelines

Info flow with QuickSet's MM architecture



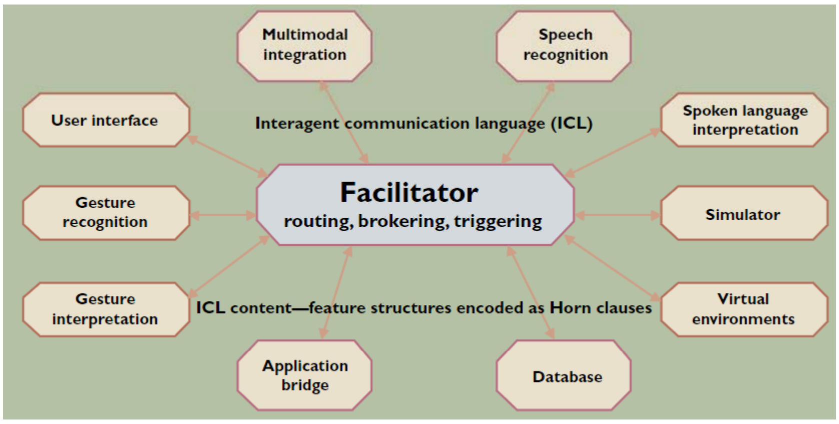
(Oviatt et al. 2000)

Members-Team-Committee recognition approach



(Oviatt et al. 2000)

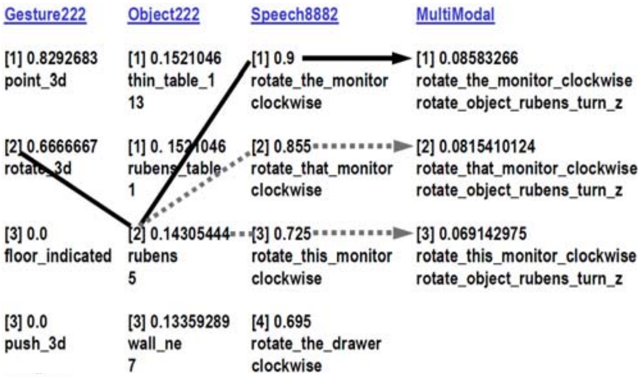
Facilitated multiagent architecture



(Oviatt et al. 2000)

Mutual Disambiguation

- Parallel coordinate plot showing multiple hypotheses for each modality
- MD derives the best joint interpretation by unification of meaning representation fragments
- MD stabilizes performance in challenging environments



Gesture222	Object222	Speech8882	MultiModal
[1] 0.8292683 point_3d	[1] 0.1521046 thin_table_1 13	[1] 0.9 rotate_the_monitor clockwise	[1] 0.08583266 rotate_the_monitor_clockwise rotate_object_rubens_turn_z
[2] 0.6666667 rotate_3d	[1] 0. 1521046 rubens_table 1	[2] 0.855 rotate_that_monitor clockwise	[2] 0.0815410124 rotate_that_monitor_clockwise rotate_object_rubens_turn_z
[3] 0.0 floor_indicated	[2] 0.14305444 rubens 5	[3] 0.725 rotate_this_monitor clockwise	[3] 0.069142975 rotate_this_monitor_clockwise rotate_object_rubens_turn_z
[3] 0.0 push_3d	[3] 0.13359289 wall_ne 7	[4] 0.695 rotate_the_drawer clockwise	

(Kaiser et al., 2003)

Combining classifiers

Now, according to the Bayesian theory, given measurements x_i, $i = 1, \ldots, R$, the pattern, Z, should be assigned to class ω_j provided the a posteriori probability of that interpretation is maximum, i.e.

$$\textit{assign} \quad Z \to \omega_j \quad \textit{if}$$

$$P(\omega_j | \mathbf{x}_1, \ldots, \mathbf{x}_R) = \max_k P(\omega_k | \mathbf{x}_1, \ldots, \mathbf{x}_R) \qquad (1)$$

The Bayesian decision rule states that to use all the avail. information correctly to reach a decision, it is essential to compute the probabilities of the various hypotheses by considering all the measurements simultaneously. This is, of course, a correct statement of the classification problem but it may not be a practicable proposition.

Combining classifiers

- We shall therefore attempt to simplify the above rule and express it in terms of decision support computations performed by the individual classifiers, each exploiting only the information conveyed by vector x_i.
- We shall see that this will not only make rule (1) computationally manageable, but also it will lead to combination rules which are commonly used in practice.

Combining classifiers

- Product rule

$$assign \qquad Z \to \omega_j \qquad if$$

$$P(\omega_j)\prod_{i=1}^{R} p(\mathbf{x}_i|\omega_j) = \max_{k=1}^{m} P(\omega_k)\prod_{i=1}^{R} p(\mathbf{x}_i|\omega_k)$$

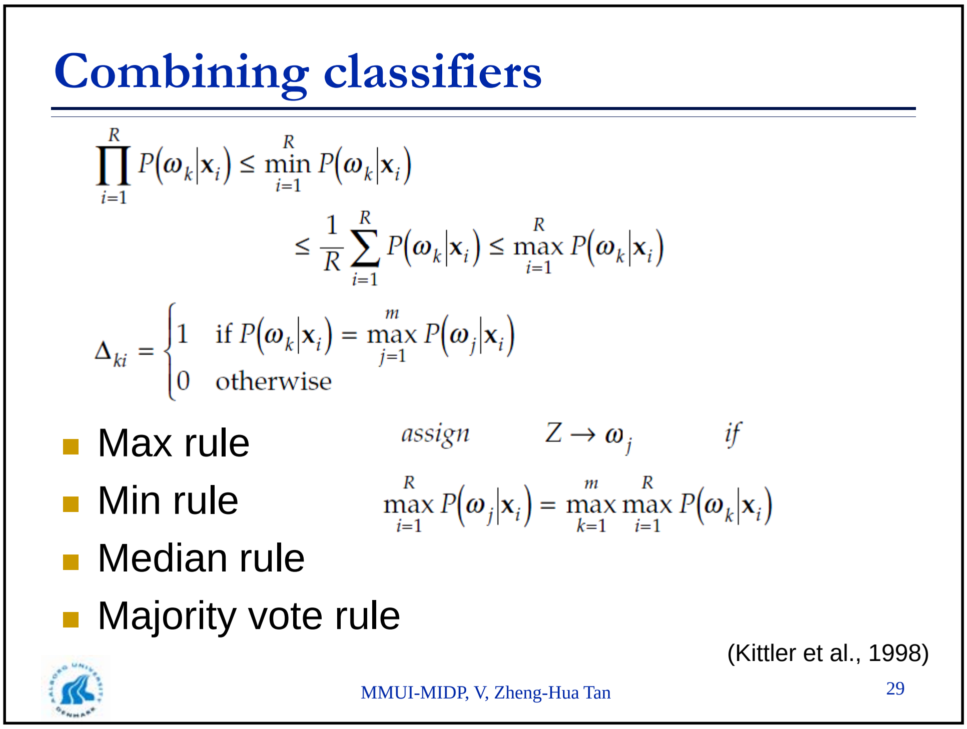
- Sum rule

$$assign \qquad Z \to \omega_j \qquad if$$

$$(1-R)P(\omega_j) + \sum_{i=1}^{R} P(\omega_j|\mathbf{x}_i) =$$

$$\max_{k=1}^{m}\left[(1-R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_k|\mathbf{x}_i)\right]$$

(Kittler et al., 1998)



Combining classifiers

$$\prod_{i=1}^{R} P(\omega_k|\mathbf{x}_i) \le \min_{i=1}^{R} P(\omega_k|\mathbf{x}_i)$$

$$\le \frac{1}{R}\sum_{i=1}^{R} P(\omega_k|\mathbf{x}_i) \le \max_{i=1}^{R} P(\omega_k|\mathbf{x}_i)$$

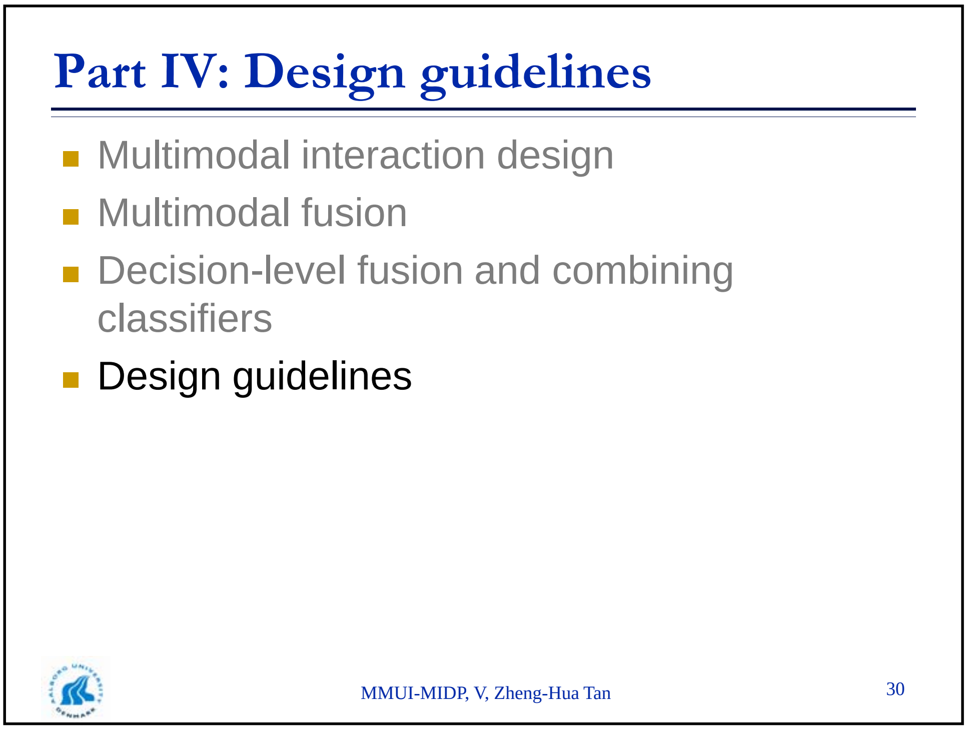
$$\Delta_{ki} = \begin{cases} 1 & \text{if } P(\omega_k|\mathbf{x}_i) = \max_{j=1}^{m} P(\omega_j|\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

- Max rule
- Min rule
- Median rule
- Majority vote rule

$$\textit{assign} \quad Z \to \omega_j \quad \textit{if}$$

$$\max_{i=1}^{R} P(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m}\max_{i=1}^{R} P(\omega_k|\mathbf{x}_i)$$

(Kittler et al., 1998)



Part IV: Design guidelines

- Multimodal interaction design
- Multimodal fusion
- Decision-level fusion and combining classifiers
- Design guidelines

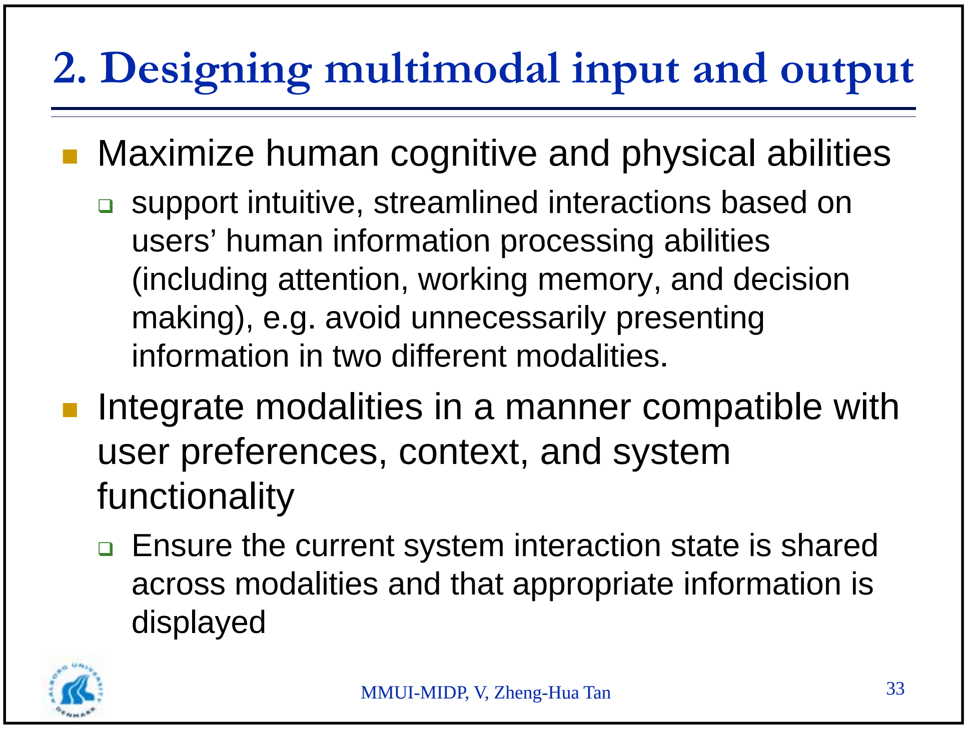
Guidelines for MMUI design

- The goals for MMUI are twofold:
  - to achieve an interaction closer to natural human-human communication, and
  - to increase the robustness of the interaction by using redundant or complementary information.
- Six main categories of guidelines

(Reeves, et al. 2004)

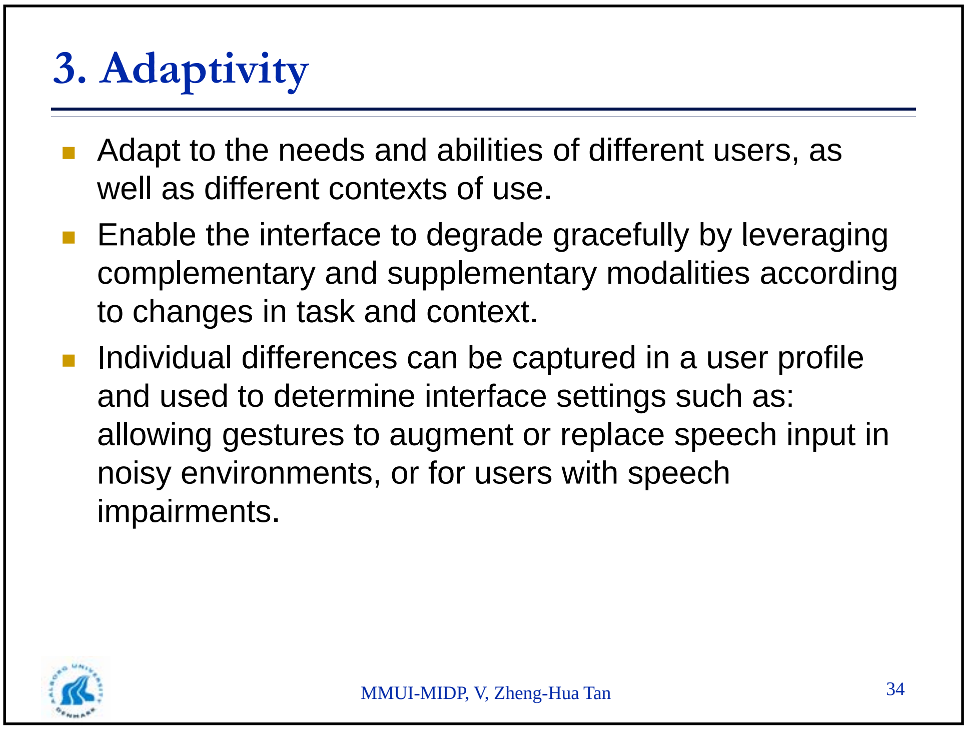
1. Requirements specification

- Design for broadest range of users and contexts of use
  - be familiar with users' psychological characteristics, level of experience, domain and task characteristics, cultural background, as well as their physical attributes (e.g., age, vision).
  - extend the range of potential users and uses, e.g. being applicable in dark and/or noisy environments.
  - support the best modality or combination of modalities in changing environments (e.g. in a car).
- Address privacy and security issues



2. Designing multimodal input and output

- Maximize human cognitive and physical abilities
  - support intuitive, streamlined interactions based on users’ human information processing abilities (including attention, working memory, and decision making), e.g. avoid unnecessarily presenting information in two different modalities.
- Integrate modalities in a manner compatible with user preferences, context, and system functionality
  - Ensure the current system interaction state is shared across modalities and that appropriate information is displayed



3. Adaptivity

- Adapt to the needs and abilities of different users, as well as different contexts of use.
- Enable the interface to degrade gracefully by leveraging complementary and supplementary modalities according to changes in task and context.
- Individual differences can be captured in a user profile and used to determine interface settings such as: allowing gestures to augment or replace speech input in noisy environments, or for users with speech impairments.

4. Consistency

- Presentation and prompts should share common features as much as possible and should refer to a common task including using the same terminology across modalities.

5. Feedback

- Users should
  - be aware of their current connectivity and know which modalities are available to them.
  - be made aware of alternative interaction options without being overloaded by lengthy instructions that distract from the task. Specific examples include using descriptive icons.

6. Error Prevention/Handling

- Provide clearly marked exits from a task, modality, or the entire system, and allow users to undo a previous action or command.
- Provide concise and effective help in the form of task-relevant and easily accessible assistance.
  - Integrate complementary modalities to improve overall robustness during multimodal fusion, thus enabling the strengths of each to overcome weaknesses in others;
  - Give users control over modality selection, so they can use a less error-prone modality for given lexical content;
  - If an error occurs, permit users to switch to a different modality.

Summary

- Multimodal interaction design
- Multimodal fusion
- Decision-level fusion and combining classifiers
- Design guidelines