Multimodal Interaction

Lesson 10
Multimodal Fusion

Maria De Marsico
demarsico@di.uniroma1.it

Credits

Derived from:

- Patrizia Grifoni. **Multimodal Human Computer Interaction and Pervasive Services**. Information Science Reference. 2009
- Niels Ole Bernsen,_Laila Dybkjr. **Multimodal Usability**. Springer 2009

Martin's cooperation and fusion

- In multimodal systems, fusion techniques are mostly applied to **complementary** and **redundant** modalities in order to integrate the information provided by them.
  - Complementary modalities provide the system with non-redundant information that have to be merged in order to get a complete and meaningful message.
  - Redundant modalities require a fusion process that avoids non-meaningful information, increasing, at the same time, the accuracy of the fused message by using one modality to disambiguate information in the other ones.

Maria De Marsico - demarsico@di.uniroma1.it

Fusion approaches

- Current fusion approaches can be considered through two main classifications:
  - according to the **data fusion level** (e.g. the fusion process takes places in the dialogue management system, as well as at grammar level)
  - according to the mathematical method (e.g. based on statistical or artificial intelligence techniques).

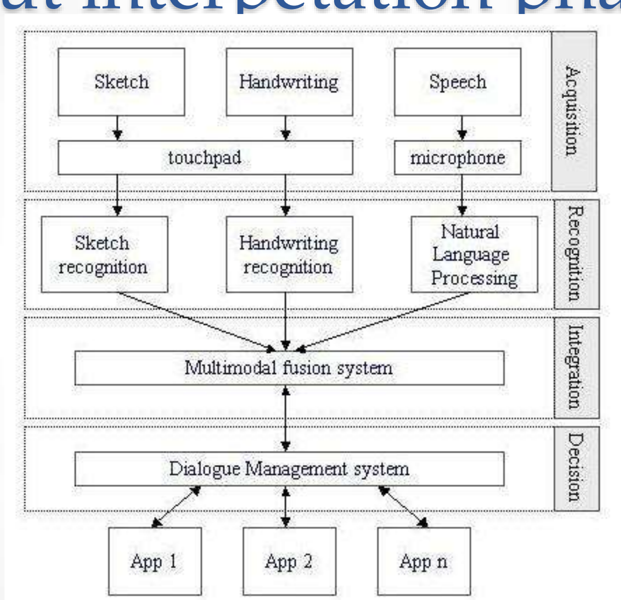
Maria De Marsico - demarsico@di.uniroma1.it

Input interpetation phases

- The mapping between the input message expressed by the user and the corresponding output returned by the system is defined ***input interpretation***.
- The interpretation process involves, generally, four phases, corresponding to the main architectural levels of a multimodal system: the **acquisition**, **recognition**, **integration** and **decision** phases (levels).
- Although the acquisition, recognition and decision are consecutive phases, the same doesn't occur for the **integration** phase (where the **fusion** process takes place), because in some systems the integration phase is prior to the recognition or decision phases, whereas in other systems it's just the opposite.

Maria De Marsico - demarsico@di.uniroma1.it

Input interpetation phases



From: Arianna D'Ulizia. Exploring Multimodal Input Fusion Strategies. In Patrizia Grifoni. **Multimodal Human Computer Interaction and Pervasive Services**. Information Science Reference. 2009

Fusion: when?

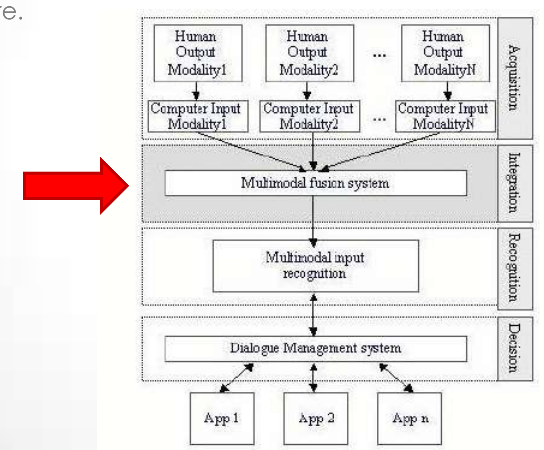
- The following material mostly come from:From: Arianna D'Ulizia. Exploring Multimodal Input Fusion Strategies. In Patrizia Grifoni. **Multimodal Human Computer Interaction and Pervasive Services**. Information Science Reference. 2009

- The integration level, in which the fusion of the input signals is performed, may be placed:
  - immediately after the acquisition level and we refer to the ***fusion at the acquisition***, or **signal**, *level*;
  - Immediately after the recognition level and in this case we refer to the ***fusion at the recognition***, or **feature**, *level*;
  - during the decision level and we refer to the ***fusion at the decision***, or **conceptual**, *level*.

Maria De Marsico - demarsico@di.uniroma1.it

Fusion at acquisition level

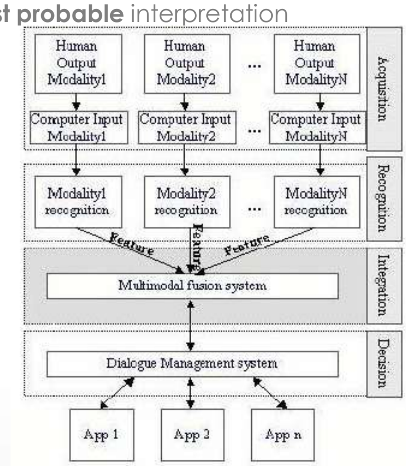
- The ***fusion at the acquisition level*** consists in mixing two or more (generally electrical) signals.
- This kind of fusion may be performed if the signals are **synchronized** and of the **same nature** (two speech inputs, two sketch inputs, etc.)
- It **cannot** be applied to **multimodal** inputs, which may be of different nature.



a1.it

Fusion at recognition level

- The *fusion at* ***the recognition level*** (also ***early fusion*** or *recognition/feature-based fusion*) consists in merging the outcomes of each recognizer by using integration mechanisms (e.g., statistical integration techniques, agent theory, hidden Markov models, artificial neural networks, etc.
- Afterwards, the **integrated** sentence is processed by the decision manager that provides its **most probable** interpretation

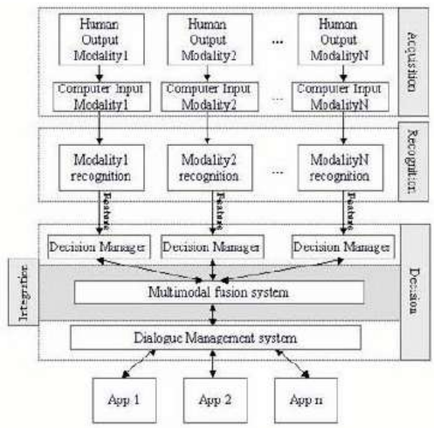


Fusion at recognition level

- A **unimodal recognition** stage and an **integrated decision** stage characterize the interpretation process of the early fusion.
- This strategy is generally preferred for closely and synchronized inputs that convey the same information (**redundant modalities**), as for example speech and lip movements for speech recognition or voice and video features for emotion recognition.
- The main **drawbacks** of the early fusion are the necessity of a large amount of data for the training, and the high computational costs.

Fusion at decision level

- The *fusion at* ***the decision level*** (named also ***late*** *fusion* or *decision/conceptual-based fusion*) means merging directly the semantic information that are extracted from the specific decision managers.
- The outcomes of each recognizer are **separately** interpreted by the decision managers and the extracted semantic meanings are integrated by using specific dialogue-driven fusion procedures to yield the complete interpretation.



Fusion at decision level

- Late fusion is mostly suitable for modalities that **differ** both in their **nature** and in the **time** scale.
- A **tight synchrony** among the various communicative modalities is essential to deliver the correct information at the right time.
- **Reminder: syncronous does not necessarily mean "in the same time"**
- Each input modality is separately recognized and interpreted → the this kind of fusion can rely on the use of standard and well-tested recognizers and interpreters for each modality, as well as on much simpler fusion algorithms.

Hybrid multi-level fusion

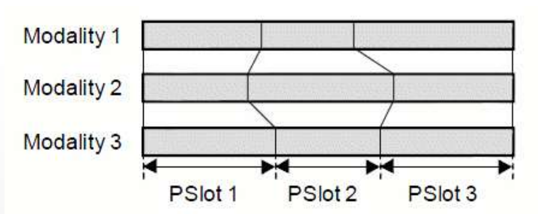
- A fourth level, named ***hybrid multi-level fusion***, can be identified.
- In this kind of fusion the integration of input signals is distributed among the acquisition, the recognition and decision levels.
- The **interdependence** among modalities allows predicting subsequent symbols knowing previous symbols in the input data flow
- **Interdependence** is exploited to improve accuracy of the interpretation process.
- The basis of the hybrid multilevel fusion strategy is a joint multimodal language model, which relies on the symbols acquired during the acquisition phase and is governed by their semantic meanings extracted during the decision phase.

Recognition-Based Fusion strategies

- Integration of input signals at recognition level, requires appropriate structures to represent these signals.
- Three main kinds of representations. Examples:
  - action frame (Vo, 1998)
  - input vectors (Pavlovic et al., 1997)
  - slots (Andre et al., 1998).

Action Frame

- A multimodal input event = a set of **parallel streams** that can be **aligned** and **jointly segmented** such that each part of the segmented input influences part of the interpretation
- **Each stream** represents one **unimodal** input coming from a computer input modality and consists of **elements** associated to a **set of parameters**.
- The **integration** of unimodal inputs consists in producing a **sequence** of **input segments**, named **parameter slots**



Maria De Marsico - demarsico@di.uniroma1.it

Action Frame (cont.)

- A parameter slot **separately** contributes to the multimodal input interpretation, that is called **action frame**.
- An action frame **specifies the action** that has to be performed **in response** to the multimodal input.
- Each **parameter slot** specifies one **action parameter.** The input segments in each parameter slot should contain enough information to determine the value of the corresponding parameter.

Maria De Marsico - demarsico@di.uniroma1.it

Action Frame – Example

- Suppose we have a map navigation system that allows the user to ask for information by speaking and drawing on the screen → The user might say "How far is it from here to there?" while drawing an arrow between two points on the displayed map.
- The speech input stream consists of the words in the utterance whereas the pen input stream contains a pair of *arrow_start* and *arrow_end* tokens.
- The interpretation of this input combination is a *QueryDistance* action frame containing a *QueryDistanceSource* parameter slot followed by a *QueryDistanceDestination* parameter slot.

Maria De Marsico - demarsico@di.uniroma1.it

Action Frame – Example (cont)

- The input streams are segmented and aligned as follows:

Speech:	how far is it from here	to there
Pen:	arrow_start	arrow_end
	QueryDistanceSource	*QueryDistanceDestination*

- If the destination point is somewhere outside the displayed area, the user might say: "How far is it from here to Philadelphia?" and circle the starting point instead.

Speech:	how far is it from here	to philadelphia
Pen:	circle	
	QueryDistanceSource	*QueryDistanceDestination*

- For the utterance "How far is it from Pittsburgh to Philadelphia?" the parameter slots would consist of speech segments only.

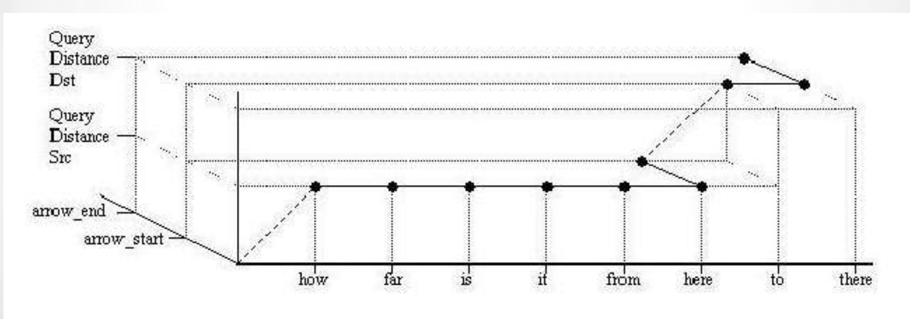
Maria De Marsico - demarsico@di.uniroma1.it

How to integrate input streams

- The integration of the information streams is carried out through the **training** of a **Multi-State Mutual Information Network (MS-MIN)**
- The **MS-MIN** network allows to find **an input segmentation** and a corresponding **parameter slot assignment** in order to extract the actual **action parameters** from the multimodal input.
- A **posteriori** probability of the parameter slot assignment conditional on the input segmentation is introduced.
- This probability is estimated by **output activations** in the MS-MIN network and can be interpreted as the **score of a path** that goes through the segmented parameter slots.

Action Frame – Example (cont)

- An example of path



How to find the path

- A **path score maximization algorithm** is applied to find the input segmentation and the corresponding parameter slot assignment.
- The algorithm creates an extra layer on top of the network.
- Each output unit of the MS-MIN is an output **state** and the top layer of the network produces the best sequence of states that fits the input, according to the path score maximization algorithm.
- Starting point: **Maximum A Posteriori probability** (MAP)

Maria De Marsico - demarsico@di.uniroma1.it

How to find the path

- Suppose we have a sequence of input tokens t_m , *m*= 1 … *M*, that is to be associated with one of several output classes c_n , *n*= *1…N*. It is reasonable to select the *maximum a posteriori* (MAP) hypothesis, or the output class having the greatest *a posteriori* probability given the input:

$$c_{MAP} = \operatorname*{argmax}_{c_n} P(c_n \mid t_1 t_2 \dots t_M)$$

$$= \operatorname*{argmax}_{c_n} \frac{P(t_1 t_2 \dots t_M \mid c_n) P(c_n)}{P(t_1 t_2 \dots t_M)}$$

From Bayes' theorem

Maria De Marsico - demarsico@di.uniroma1.it

How to find the path (cont)

- If we make the simplifying assumption that the input tokens are independent as well as conditionally independent given the target output, i.e.

$$P(t_1 t_2 \dots t_M) = \prod_{m=1}^{M} P(t_m)$$

$$P(t_1 t_2 \dots t_M \mid c_n) = \prod_{m=1}^{M} P(t_m \mid c_n)$$

then it follows that

$$c_{MAP} = \underset{c_n}{\operatorname{argmax}} P(c_n) \prod_{m=1}^{M} \frac{P(t_m \mid c_n)}{P(t_m)}$$

Bayesian classifier applied to a "bag of words" model = the input is considered an unordered collection of independent words.

Maria De Marsico - demarsico@di.uniroma1.it

How to find the path (cont)

- Logarithm function is monotonically increasing → *f* (x) and log $_2$ *f(x)* reach their respective maximum values at the same x value for all *f* (x):

$$c_{MAP} = \underset{c_n}{\operatorname{argmax}} \left(\log_2 P(c_n) + \sum_{m=1}^{M} \log_2 \frac{P(t_m \mid c_n)}{P(t_m)} \right)$$

$$= \underset{c_n}{\operatorname{argmax}} \left(\log_2 P(c_n) + \sum_{m=1}^{M} I(t_m, c_n) \right)$$

mutual information of input token t_m and output class c_n

The right hand side of Equation can be implemented by a connectionist network.

Maria De Marsico - demarsico@di.uniroma1.it

Input Frames

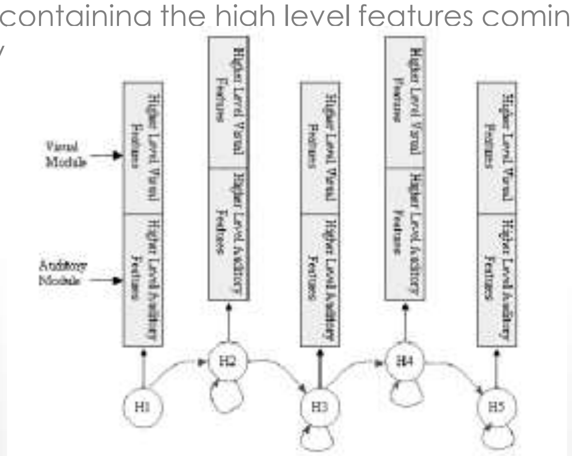
- The input vectors proposed by Pavlovic are used to store the outputs of the visual and auditory interpretation modules.

- The **visual** module firstly tracks the features of the video data by using **skin color region segmentation** and **motion-based region tracking** algorithms and the time series of the tracked features is stored into an input vector.
- Secondly, these features are dynamically classified by using **Probabilistic Independence Networks** (PINs) and **Hidden Markov Models** (HMMs).
- The output of this module consists in a set of higher level features ranged from gestural movement elements, called **visemes** (e.g. "left movement"), to **full gestural words** (e.g. symbol for "rotate about x-axis).

- The **auditory** module has the same architecture and functioning of the visual module applied to audio data.
- A HMM PIN allows to classify the auditory features into auditory elements, called **phonemes**, and **full spoken words**.

Input Frames (cont.)

- The integration of the two interaction modalities is carried out through a **set of HMM PIN structures**, each corresponding to a **predefined audio/visual command**.
- The **state** of each HMM is defined according to the input vectors containing the high level features coming from the auditory

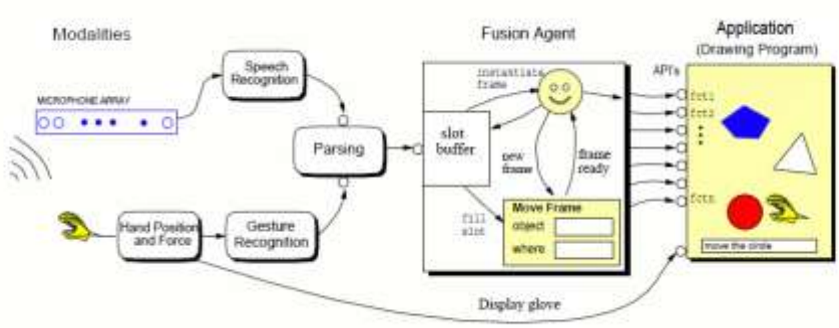


Slots

- In the strategy based on ***slots***, the information inputted by the user is stored into a **slot buffer**, which allows **back referencing** of past lexical units (e.g.: “it” can to reference the previously selected object).
- The command language of the application is encoded in semantic units called **frames**.
- The command frames are composed of **slots**, i.e. lexical units provided by the multimodal input.
- Example: considering the “move frame” two slots can be identified: “object” (to specify the object) and “where” (to specify the final position).
- The frames are predefined (computed off line) and are application-dependent.

Slots

- The parser extracts the lexical units from different input modalities and fills the appropriate slots in the slot buffer.
- The slot buffer is continuously monitored checking for filled frames. Once a frame is filled (enough information to generate a command), the fusion agent sends it to be executed in the current application.



Maria De Marsico - demarsico@di.uniroma1.it

Conclusions about recognition-based strategies

- Main advantages:
  - great coherence with the human-human communication paradigm in which the dialogue is considered as a unique and multimodal communication act; analogously, the recognition-based fusion strategies merge the recognized inputs into a unique multimodal sentence that has to be opportunely interpreted;
  - they allow an easier inter-modality disambiguation.
- Main drawbacks:
  - significant computational load
  - high dependency on time measures; this dependency implies as well a large amount of real data to train the network (both the MS-MIN and the PIN HMM).

Decision-Based Fusion Strategies

- In the decision-based approach, the outcomes of **each** recognizer are **separately** interpreted by specific **decision managers** and then sent to the **dialogue management** system that performs their integration by using specific **dialogue-driven fusion procedures** to yield the complete interpretation.
- To represent the **partial** interpretations coming from the decision managers and achieve the integration of input signals at decision level, several kinds of structures might be employed. Examples:

- *typed feature structures* (Cohen et al., 1997; Johnston, 1998),
- *melting pots* (Nigay and Coutaz, 1995),
- *semantic frames* (Vo and Wood,1996; Russ et al., 2005)

Feature structures

- A feature structure consists of a **collection** of **feature-value pairs**.
- The value of a feature may be an **atom**, a **variable**, or **another feature structure**.
- When two features structures are **unified**, a **composite** structure containing **all of the feature** specifications from each component structure is formed.
- Any feature common to both feature structures **must not clash** in its value.
  - If the values of a common feature are **atoms** they must be **identical**.
  - If **one** is a variable, it becomes **bound** to the **value** of the corresponding feature in the **other** feature structure.
  - If both are **variables**, they become **bound** together, constraining them to **always** receive the **same** value (if unified with another appropriate feature structure).
  - If the values are themselves **feature** structures, the unification operation is applied **recursively**.

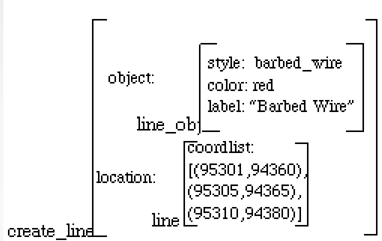
Feature structures

- Importantly, feature structure **unification** can result in a **directed acyclic graph** structure when more than one value in the collection of feature/values pairs makes use of the **same variable**. Whatever value is ultimately unified with that variable thus will fill the value slot of **all** the corresponding features, resulting in a DAG.

Typed feature structures

- ***Typed*** feature structures are an extension of the representation whereby **feature structures** and **atoms** are assigned to **hierarchically ordered types**
- Hierarchy represents domain-specific as well as domain-independent knowledge **using IS-A** and **IS-PART-OF** relations.
- Typed feature structure **unification** requires pairs of feature structures or pairs of atoms which are being unified **to be compatible in type**.
- To be compatible in type, one must be in the **transitive closure of the subtype relation** with respect to the other.
- The result of a **typed unification** is the **more specific feature** structure or atom in the type hierarchy.
- Typed feature structure **unification** is ideally suited to the task of multimodal integration → we want to determine whether a given piece of, say, gestural input is **compatible** with, say, a given piece of spoken input, and if they are compatible, to combine the two inputs into a single result that can be interpreted by the system.
- **Unification** is appropriate for multimodal integration because it can combine **complementary** or **redundant** input from both modes but **rules out contradictory** inputs.

Example



feature structure assigned to the command 'create barbed wire'

Type: **create_line**
Feature: object
Value: feature straucture of type **line_ob**

Feature: location
Value: feature structure of type **line**

Type: **line_ob**
Feature: style
Value: barbed_wire

Feature: color
Value: red

Feature: label
Value: "Barbed Wire"

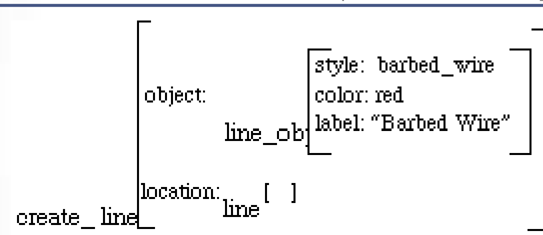
Type: **line**
Feature: coordlist
Value: [(95301, 94360),
(95305, 94365),
(95310, 94380)]

Representation of partial meaning

- The use of feature structures as a **semantic** representation framework facilitates the specification of **partial** meanings.
- Spoken or gestural input which **partially specifies** a command can be represented as an **underspecified** feature structure in which **certain** features **are not instantiated**, **but are given a certain type** based on the semantics of the input

Representation of partial meaning

- For example, if a given speech input can be integrated with a line gesture, it can be assigned a feature structure with an underspecified location feature whose value is required to be of **type *line***

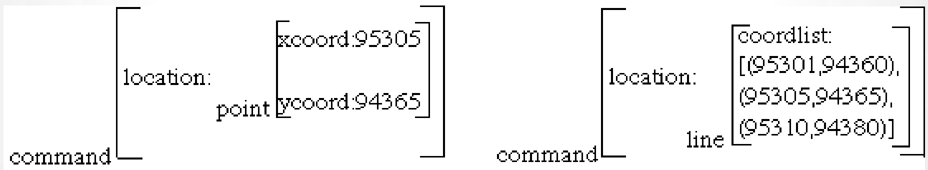


feature structure assigned to the spoken phrase 'barbed wire'

- This phrase is interpreted as a **partially specified** creation command.
- **Before** it can be executed, **it needs a location feature** indicating where to create the line, **which is provided by the user's drawing on the screen.**

Examples of interpretation

- The user's gestures can be assigned a number of interpretations, for example, both a point interpretation and a line interpretation
- Interpretations are represented as typed feature structures.

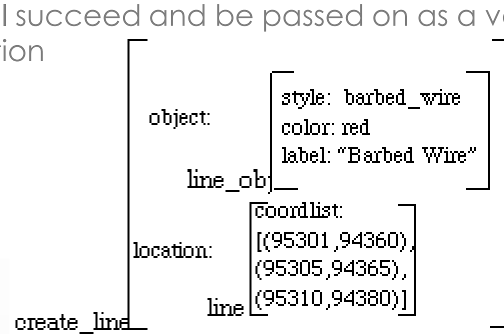


- Continuing our example, **interpretations** of gestures as **location** features are assigned the more general ***command*** type which unifies with all of the commands supported by the system, one of which is *create_line*

Maria De Marsico - demarsico@di.uniroma1.it

Multimodal Compensation.

- In our example, both speech and gesture have only **partial** interpretations, one for speech, and two for gesture.
- The speech interpretation requires its location feature to be of **type *line*** → only unification with the line interpretation of the gesture will succeed and be passed on as a valid multimodal interpretation



- To select the best unified interpretation among the alternative solutions probabilities are associated with each unimodal input.

Maria De Marsico - demarsico@di.uniroma1.it

Grammars with typed feature structures

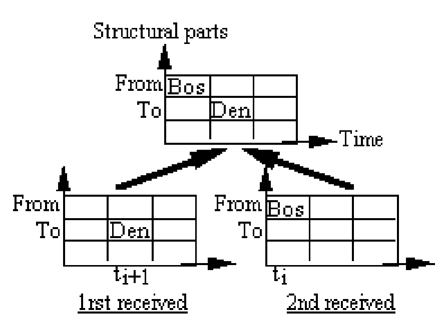
- Going further, Johnston (1998) introduces a grammar representation in which spoken phrases and pen gestures are the **terminal** elements of the grammar, referred to as **lexical edges**.

- Each lexical edge is assigned **grammatical** representations in the form of **typed feature structures**

Maria De Marsico - demarsico@di.uniroma1.it

Melting pots

- A **melting pot** is a 2-D structure, in which the vertical axis contains the "**structural parts**", i.e. the **task objects** generated by the **input actions** of the user, and the horizontal axis is the **time**.
- The fusion is performed within the dialogue manager by using a technique based on agents.

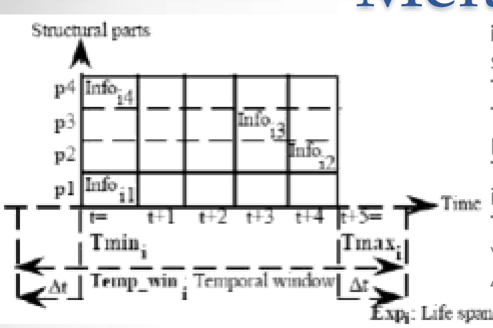


Maria De Marsico - demarsico@di.uniroma1.it

Melting strategies

- Three criteria are used to trigger the fusion of melting pots.
- **Microtemporal** fusion is used to combine information that is produced either in parallel or over overlapping time intervals.
- **Macrotemporal** fusion takes care of either sequential inputs or time intervals that do not overlap but belong to the same temporal window.
- **Contextual** fusion, serves to combine input according to contextual constraints without attention to temporal constraints.

Melting



infoij: piece of information stored in the structural part pj of mi.
Tinfoij: time-stamp of infoij.
Tmaxi: time-stamp of the most recent piece of information stored in mi.
Tmini: time-stamp of the oldest piece of information stored in mi.
Temp_wini: duration of the temporal window for mi.
Δt: Remaining life span for mi.

melting pot mi: mi=(p1, p2,... , pj,..., pn): mi is comprised of n structures p1, p2, ...pn.

The **temporal window** of a melting pot defines the temporal **proximity** (+/- Δt) of two adjacent melting pots: for mi=(p1, p2,...pn), Temp_wini=[Tmini-Δt, Tmaxi+Δt]. Temporal windows are used to trigger **macrotemporal** fusion. The last metrics used to manage a melting pot is the notion of **life span** Expi: Expi=Tmaxi+Δt=Max(Tinfoij)+Δt.
This notion is useful for removing a melting pot from the set of candidates for fusion.

Semantic frames

- Input from each modality is parsed and transformed into a **semantic frame** containing slots that specify command **parameters,** such as the action to carry out or the object to act on.
- The information in these partial frames may be incomplete or ambiguous.
- A domain **independent** frame merging algorithm combines the partial frames into a complete frame by selecting slot values from the partial frames to maximize a **combined score**.

Maria De Marsico - demarsico@di.uniroma1.it

Conclusions about decision-based strategies

- Main advantages:
  - multi-tasking, as different multimodal channels, recognizers and interpreters are arranged for carrying out independent unimodal input processing at the same time
  - the possibility to use standard and well-tested recognizers and interpreters for each modality.
- Main drawbacks:
  - high complexity of the inter-modality disambiguation, particularly when dealing with more complex modalities that need not only pairs item-time but full lattices from each channel to disambiguate the multimodal input.

Maria De Marsico - demarsico@di.uniroma1.it

Hybrid Multi-Level Fusion Strategies

- In the hybrid multi-level approach, the integration of input signals is distributed among the acquisition, the recognition and decision levels.
- Examples of methodologies that have been applied in literature:
  - o *finite-state transducers* (Johnston and Bangalore, 2000)
  - o *multimodal grammars* (Sun et al., 2006; D'Ulizia et al., 2007)

Maria De Marsico - demarsico@di.uniroma1.it

Finite State Transducers

- **Finite-state transducers** (FST) are **finite-state automata** (FSA) where each transition consists of an **input** and an **output** symbol.
- A transition is traversed if its input symbol matches the current symbol in the input and generates the output symbol associated with the transition.
- An FST can be regarded as a **2-tape FSA** with an input tape from which the input symbols are read and an output tape where the output symbols are written.
- A finitestate device **parses** multiple input streams and
- **combines** their content into a single semantic representation.
- For an interface with *n* modes, a finite state device operating over *n+1* tapes is needed (*n* input streams + 1 interpretation output)

Maria De Marsico - demarsico@di.uniroma1.it

Finite State Transducers

- The structure and interpretation of multimodal commands of can be captured declaratively in a **multimodal context-free** grammar.
- In general a context-free grammar can be approximated by an FSA
- The transition symbols of the approximated FSA are the terminals of the context-free grammar and in the case of multimodal CFG these terminals contain *n+1* components (*n* modes + interpretation)
- This approach does not support mutual disambiguation, i.e., using information from a recognized input to enable the processing of any other modality.

Multimodal grammars

- The outcomes of each recognizer are considered as **terminal** symbols of a formal grammar and consequently they are recognized by the parser as a **unique multimodal sentence**.
- In the interpretation phase the parser uses the grammar specification (production rules) to interpret the sentence.
- The unique multimodal input can be represented by using the TFS (Typed Feature Structures)

Conclusions about Hybrid Multi-Level Fusion Strategies

- Main advantages:
  - similarity with the paradigm used in the human-human communication, in which the dialogue is considered as a unique linguistic phenomenon.
- Main drawbacks:
  - high complexity of the inter-modality disambiguation.

Maria De Marsico - demarsico@di.uniroma1.it

Fusion: how?

- **Statistical** methodologies are often applied to decide on the interpretation of a multimodal sentence according to the knowledge of the acquired input signals.
- Classical statistical models applied in the literature are **bayesian networks**, **hidden markov models**, and **fuzzy logic**.
- **Artificial intelligence**-based techniques, such as **neural networks** and **agent theory** are also well-suited for classification and recognition tasks in the multimodal fusion domain.

Maria De Marsico - demarsico@di.uniroma1.it

Statistical methods

- Input signals can be characterized by a certain **degree of uncertainty** associated with the imperfection of data, frequently hard to recognize.
- To deal with this uncertainty statistical models consider **previously observed data** with respect to current data to derive the probability of an input
- Many multimodal systems, especially those that perform the fusion at **recognition** level, rely on statistical fusion strategies that use models of probability theory to combine information coming from different unimodal inputs.
- Three main statistical methods can be applied in the fusion process:
  - bayesian network
  - hidden markov models
  - fuzzy logic

Artificial Intelligence methods

- Often used to perform the fusion of input signals at recognition and decision levels.
- Examples:
- *agent- based techniques* (Nigay and Coutaz, 1995)
- *neural networks* (Meier et al., 2000; Lewis and Powers, 2002).

Readings

- Vo, M.T. (1998). *A framework and toolkit for the construction of multimodal learning interfaces*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA. Available at http://reports-archive.adm.cs.cmu.edu/anon/1998/CMU-CS-98-129.pdf
- Pavlovic, V.I., Berry, G.A., & Huang, T.S. (1997). Integration of audio/visual information for use in human-computer intelligent interaction. In *Proceedings of the 1997 International Conference on Image Processing (ICIP '97)*, (Vol. 1, pp. 121-124). Available at www.cs.rutgers.edu/~vladimir/pub/icip97.ps.gz
- Andre, M., Popescu, V.G., Shaikh, A., Medl, A., Marsic, I., Kulikowski, C., & Flanagan J.L. (1998, January). Integration of speech and gesture for multimodal human-computer interaction. In *Second International Conference on Cooperative Multimodal Communication*, Tilburg, The Netherlands (pp. 28-30). Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.7321

Readings

- Cohen, P.R., Johnston, M., McGee, D., Oviatt, S.L., Pittman, J., Smith, I.A., Chen, L., & Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. ACM Multimedia, 31-40. Available at http://www.uni-mannheim.de/acm97/papers/johnston/ACM.htm
- Johnston, M. (1998, August 10-14). Unification based multimodal parsing. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLINGACL '98), Montreal, Canada (pp. 624-630). Available at http://acl.ldc.upenn.edu/C/C98/C98-1099.pdf
- Nigay, L., & Coutaz, J. (1995). A generic platform for addressing the multimodal challenge. In Proceedings of the Conference on Human Factors in Computing Systems. ACM Press. Available at http://www.sigchi.org/chi95/proceedings/papers/lmn_bdy.htm
- Bouchet, J., Nigay, L., & Ganille, T. (2004). Icare software components for rapidly developing multimodal interfaces. In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04),New York, NY (pp. 251- 258). ACM. Available at http://iihm.imag.fr/publs/2004/ICMI04-bouchet.pdf

Maria De Marsico - demarsico@di.uniroma1.it

Readings

- Vo, M.T., & Wood, C. (1996, May 7-10). Building an application framework for speech and pen input integration in multimodal learning interfaces. In Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP'96), IEEE Computer Society (Vol. 6, pp. 3545-3548). Available at http://www.cs.cmu.edu/afs/cs/user/tue/WWW/ps/icassp96-paper.ps.gz
- Russ, G., Sallans, B., & Hareter, H. (2005, June 20-23). Semantic based information fusion in a multimodal interface. International Conference on Human-Computer Interaction (HCI'05), Las Vegas, NV (pp. 94-100). Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.4023&rep=rep1&type=pdf
- Corradini, A., Mehta, M., Bernsen, N.O., & Martin, J.-C. (2003). Multimodal input fusion in human-computer interaction on the example of the ongoing NICE project. In Proceedings of the NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert, and Response Management, Yerevan, Armenia. Available at http://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CF0QFjAA&url=http%3A%2F%2Fwww.limsi.fr%2FIndividu%2Fmartin%2Fresearch%2Farticles%2FNATO-ASI_1.doc&ei=Tf6nT5DVNcnsOZex6KQD&usg=AFQjCNH-1TcMT08KFDgseWJj7p1xYJHpxQ

Readings

- Johnston, M., & Bangalore, S. (2000). **Finite state multimodal parsing and understanding**. In Proceedings of the International Conference on Computational Linguistics, Saarbruecken, Germany. Available at http://aclweb.org/anthology-new/C/C00/C00-1054.pdf
- Sun, Y., Chen, F., Shi, Y.D., & Chung, V. (2006). **A novel method for multisensory data fusion in multimodal human computer interaction**. In Proceedings of the 20th Conference of the Computer- Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction: Design, Activities, Artefacts, and Environments, Sydney, Australia (pp. 401-404). Available at http://www.ozchi.org/proceedings/2006/sessions/short-papers/modality/sun-p401.pdf
- D'Ulizia, A., Ferri, F., & Grifoni, P. (2007, November 25-30). **A hybrid grammar-based approach to multimodal languages specification**. In OTM 2007 Workshop Proceedings, Vilamoura, Portugal (LNCS 4805, pp. 367-376). Springer-Verlag. Available at http://www-vs.informatik.uni-ulm.de/DE/intra/bib/2007/OTM/papers/4805/48050367.pdf

Readings

- Pérez, G., Amores, G., & Manchón, P. (2005). **Two strategies for multimodal fusion. In Proceedings of Multimodal Interaction for the Visualization and Exploration of Scientific Data**, Trento, Italy (pp. 26-32). Available at http://grupo.us.es/julietta/publications/2005/pdf/Two_Strategies.pdf

- Meier, U., Stiefelhagen, R., Yang, J., & Waibel, A. (2000). **Towards unrestricted lip reading**. International Journal of Pattern Recognition and Artificial Intelligence, 14(5), 571-585. Available at http://swing.adm.ri.cmu.edu/pub_files/pub1/meier_uwe_1999_1/meier_uwe_1999_1.pdf

- Mick Cody, Fred Cummins, Eva Maguire, Erin Panttaja, David Reitter. Research Report on **Adaptive Multimodal Fission and Fusion**. http://web.mit.edu/~erinp/mosaic/MLE/Web/erin/ff-report.pdf

- Jean-Claude MARTIN. **Towards "intelligent" cooperation between modalities. The example of a system enabling multimodal interaction with a map.** http://perso.limsi.fr/Individu/martin/ijcai/article.html