# Virtual Modality: a Framework for Testing and Building Multimodal Applications

**Péter Pál Boda**[1]
Audio-Visual Systems Laboratory
Nokia Research Center
Helsinki, Finland
boda@mit.edu

**Edward Filisko**
Spoken Language Systems Group
CSAIL, MIT
Cambridge, Massachusetts, USA
filisko@csail.mit.edu

## Abstract

This paper introduces a method that generates simulated multimodal input to be used in testing multimodal system implementations, as well as to build statistically motivated multimodal integration modules. The generation of such data is inspired by the fact that true multimodal data, recorded from real usage scenarios, is difficult and costly to obtain in large amounts. On the other hand, thanks to operational speech-only dialogue system applications, a wide selection of speech/text data (in the form of transcriptions, recognizer outputs, parse results, etc.) is available. Taking the textual transcriptions and converting them into multimodal inputs in order to assist multimodal system development is the underlying idea of the paper. A conceptual framework is established which utilizes two input channels: the original speech channel and an additional channel called Virtual Modality. This additional channel provides a certain level of abstraction to represent non-speech user inputs (e.g., gestures or sketches). From the transcriptions of the speech modality, pre-defined semantic items (e.g., nominal location references) are identified, removed, and replaced with deictic references (e.g., here, there). The deleted semantic items are then placed into the Virtual Modality channel and, according to external parameters (such as a pre-defined user population with various deviations), temporal shifts relative to the instant of each corresponding deictic reference are issued. The paper explains the procedure followed to create Virtual Modality data, the details of the speech-only database, and results based on a multimodal city information and navigation application.

## 1 Introduction

Multimodal systems have recently drawn significant attention from researchers, and the reasons for such an interest are many. First, speech recognition based applications and systems have become mature enough for larger-scale deployment. The underlying technologies are gradually exhibiting increased robustness and performance, and from the usability point of view, users can see some clear benefits from speech-driven applications. The next evolutionary step is the extension of the "one dimensional" (i.e., speech-only) interface capabilities to include other modalities, such as gesture, sketch, gaze, and text. This will lead to a better and more comprehensive user experience.

A second reason is the widely accepted, and expected, mobility and pervasiveness of computers. Devices are getting more and more powerful and versatile; they can be connected anywhere and anytime to networks, as well as to each other. This poses new demands for the user interface. It is no longer sufficient to support only a single input modality. Depending on the specific application, the given usage scenario, and the context, for example, users should be offered a variety of options by which to interact with the system in an appropriate and efficient way.

Third, as the output capabilities of devices provide ever-increasing multimedia experiences, it is natural that the input mechanism must also deal with various

[1] Currently a Visiting Scientist with the Spoken Language Systems Group, CSAIL, MIT.

modalities in an intuitive and comprehensive manner. If a map is displayed to the user, it is natural to expect that the user may want to relate to this physical entity, for instance, via gestures, pointing, gazing or by other, not necessarily speech-based, communicative means.

Multimodal interfaces give the user alternatives and flexibility in terms of the interaction; they are enabling rather than restricting. The primary goal is to fully understand the user's intention, and this can only be realized if all intentional user inputs, as well as any available contextual information (e.g., location, pragmatics, sensory data, user preferences, current and previous interaction histories) are taken into account.

This paper is organized as follows. Section 2 introduces the concept of Virtual Modality and how the multimodal data are generated. Section 3 explains the underlying Galaxy environment and briefly summarizes the operation of the Context Resolution module responsible for, among other tasks, resolving deictic references. The data generation as well as statistics is covered in Section 4. The experimental methodology is described in Section 5. Finally, the results are summarized and directions for future work are outlined.

# 2 Virtual Modality

This section explains the underlying concept of Virtual Modality, as well as the motivation for the work presented here.

## 2.1 Motivation

Multiple modalities and multimedia are an essential part of our daily lives. Human-human communication relies on a full range of input and output modalities, for instance, speech, gesture, vision, gaze, and paralinguistic, emotional and sensory information. In order to conduct seamless communication between humans and machines, as many such modalities as possible need to be considered.

Intelligent devices, wearable terminals, and mobile handsets will accept multiple inputs from different modalities (e.g., voice, text, handwriting), they will render various media from various sources, and in an intelligent manner they will also be capable of providing additional, contextual information about the environment, the interaction history, or even the actual state of the user. Information such as the emotional, affective state of the user, the proximity of physical entities, the dialogue history, and biometric data from the user could be used to facilitate a more accommodating, and concise interaction with a system. Once contextual information is fully utilized and multimodal input is supported, then the load on the user can be considerably reduced. This is especially important for users with severe disabilities.

Implementing complex multimodal applications represents the chicken-and-egg problem. A significant amount of data is required in order to build and tune a system; on the other hand, without an operational system, no real data can be collected. Incremental and rule-based implementations, as well as quick mock-ups and Wizard-of-Oz setups (Lemmelä and Boda, 2002), aim to address the application development process from both ends; we follow an intermediate approach.

The work presented here is performed under the assumption that testing and building multimodal systems can benefit from a vast amount of multimodal data, even if the data is only a result of simulation. Furthermore, generating simulated multimodal data from textual data is justified by the fact that a multimodal system should also operate in speech-only mode.

## 2.2 The concept

Most current multimodal systems are developed with particular input modalities in mind. In the majority of cases, the primary modality is speech and the additional modality is typically gesture, gaze, sketch, or any combination thereof. Once the actual usage scenario is fixed in terms of the available input modalities, subsequent work focuses only on these input channels. This is advantageous on the one hand; however, on the other hand, there is a good chance that system development will focus on tiny details related to the modality-dependent nature of the recognizers and their particular interaction in the given domain and application scenario.

Virtual Modality represents an abstraction in this sense. The focus is on what semantic units (i.e., meaningful information from the application point of view) are delivered in this channel and how this channel aligns with the speech channel. Note that the speech channel has no exceptional role; it is equal in every sense with the Virtual Modality. There is only one specific consideration regarding the speech channel, namely, it conveys deictic references that establish connections with the semantic units delivered by the Virtual Modality channel.

The abstraction provided by the Virtual Modality enables the developer to focus on the interrelation of the speech and the additional modalities, in terms of their temporal correlation, in order to study and experiment with various usage scenarios and usability issues. It also means that we do not care how the information delivered by the Virtual Modality arose, what (actual/physical) recognition process produced them, nor how the recognition processes can influence each other’s performance via cross-interaction using early evidence available in one channel or in the other - although we acknowledge that this aspect is important and desired, as pointed out by Coen (2001) and Haikonen (2003), this has not yet been addressed in the first implementation of the model.

The term "virtual modality" is not used in the multimodal research community, as far as we know. The only occurrence we found is by Marsic and Dorohonceanu (2003), however, with "virtual modality system" they refer to a multimodal management module that manages and controls various applications sharing common modalities in the context of telecollaboration user interfaces.

### 2.3 Operation

The idea behind Virtual Modality is explained with the help of Figure 1. The upper portion describes how the output of a speech recognizer (or direct natural language input from keyboard) and a sequence of words, $\{w_1 \ldots w_N\}$, is transformed into a corresponding sequence of concepts, $\{C_1 \ldots C_M\}$. The module responsible for this operation, for the sake of simplicity and generality, is called a classifier (CL). In real-life implementations, this module can be a sophisticated natural language understanding (NLU) unit, a simple semantic grammar, or a hybrid of several approaches.

The middle part of Figure 1 exhibits how the Virtual Modality is plugged into the classifier. The Virtual Modality channel (VM) is parallel to the speech channel (Sp) and it delivers certain semantic units to the classifier. These semantic units correspond to a portion of the word sequence in the speech channel. For instance, the original incoming sentence might be, "*From Harvard University to MIT*." In the case of a multimodal setup the semantic unit, originally represented by a set of words in the speech channel (e.g., "*to MIT*"), will be delivered by the Virtual Modality as $m_i$.

The tier between the two modality channels is a deictic reference in the speech channel ($d_i$ in the bottom of Figure 1). There are various realizations of a deictic reference, in this example it can be, for example, "here", "over there", or "this university". Nevertheless, in all cases for these input combinations (i.e., speech only, speech and some other modality) the requirement is that the very same sequence of semantic concepts is to be produced by the classifier.

There is one more criterion: there must be a temporal correspondence between the input channels. The deictic reference can only be resolved if the input delivered by the other modality channel is within certain time frames. This is indicated tentatively in the figure as $m_i$ occurring either in synchrony with, prior to, or following the deictic reference in time (see Section 4.3).

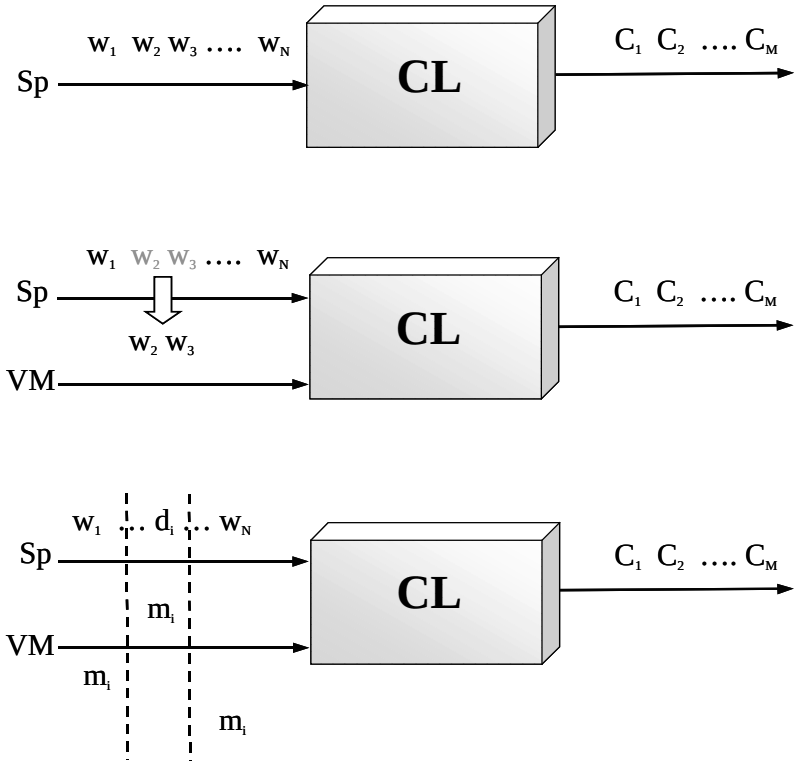


**Figure 1.** The concept of Virtual Modality (Sp and VM stand for the Speech and Virtual Modality channels, respectively. CL is a classifier and integrator that transforms and fuses a sequence of words, $w_i$, and Virtual Modality inputs, $m_i$, into a corresponding sequence of concepts $C_k$.).

In the above described model the speech channel will always have a deictic replacement when a semantic unit is moved to the Virtual Modality channel, although Oviatt, DeAngeli and Kuhn (1997) reported their findings that in a given application domain users are not even using spoken deictic references in more than half of the multimodal input cases. Therefore, to conform to this, we keep in mind that $d_i$ can have a void value, as well.

### 2.4 The use of Virtual Modality

The framework described above enables two steps in the development of multimodal systems. First, with the introduction of the Virtual Modality, modules designed to resolve inputs from multimodal scenarios can be tested. Quite often, these inputs alone represent ambiguity and the combination of two or more input channels are needed to resolve them.

On the other hand, with the removal of pre-defined semantic units to the Virtual Modality channel, a multimodal database can be generated from the speech-only data. For instance, in a given application domain, all location references can be moved to the Virtual Modality channel and replaced by randomly chosen deictic references. Furthermore, the temporal relation between the deictic reference and the corresponding semantic unit in the Virtual Modality can be governed by external parameters. This method facilitates the generation of a large amount of "multimodal" data from only a limited amount of textual data. This new database can then be used for the first task, as described above, and equally importantly, it can be used to train statistically motivated multimodal integrator/fusion modules.

As it was pointed out by Oviatt *et al.* (2003), predictive and adaptive integration of multimodal input is necessary in order to provide robust performance for multimodal systems. Availability of data, even if it is generated artificially, can and will help in the development process.

### 2.5 Further considerations

The primary goal of an interactive system is the full understanding of the user's intention in the given context of an application. Processing all active inputs from the user can only attain this task: recognizing and interpreting them accurately. Additionally, by considering all passively and implicitly available information (e.g., location, sensory data, dialogue history, user preferences, pragmatics), the system can achieve an even fuller understanding of the user's intention.

The Virtual Modality can be used to simulate the delivery of all the previously described information. From a semantic interpretation point of view, an implicitly available piece of information, i.e., the physical location of the user (detectable by a mobile device, for instance), is equal to an active user input generated in a given modality channel. The only difference might be the temporal availability of the data: a location information derived from a mobile device is continuously available over a longer period of time, while a user gesture over a map specifying, for example the value for a "from here" deictic reference, is present only for a relatively short time.

## 3 System Architecture

Researchers in the Spoken Language Systems group at MIT have been developing human-computer dialogue systems for nearly two decades. These systems are implemented within the Galaxy Communicator architecture, which is a multimodal conversational system framework (Seneff *et al.*, 1998). As shown in Figure 2, a Galaxy system is configured around a central programmable hub, which handles the communications among various human language technology servers, including those that handle speech recognition and synthesis, language understanding and generation, context resolution, and dialogue management.

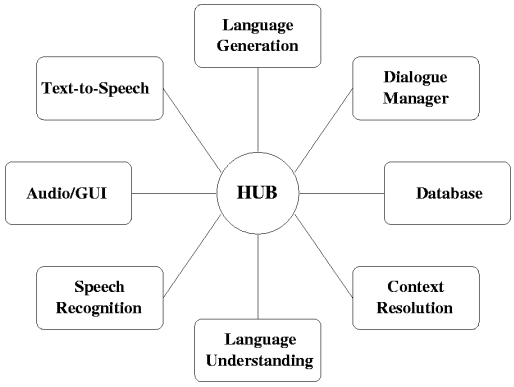


**Figure 2.** The Galaxy Communicator architecture.

Several Galaxy domains are currently under development at MIT (Zue *et al.*, 1994; Seneff *et al.*, 2000; Zue *et al.*, 2000; Seneff, 2002), but the research effort presented here concerns only Voyager, the traffic and city guide domain (Glass *et al.*, 1995; Wang, 2003) - although the Virtual Modality concept is applicable for other domains as well. Voyager's map-based interface provides opportune conditions for the use of multimodal input and deictic references. For example, a typical user input may be, "*How do I get from here to there?*" which is spoken while the user clicks on two different locations on a graphical map.

After the utterance has been recognized and parsed, the semantic frame representation of the utterance is sent to the Context Resolution (CR) server (Filisko and Seneff, 2003). It is the CR server's duty to interpret the

user's utterance in the context of the dialogue history, the user's physical environment, and limited world knowledge, via a resolution algorithm. This protocol includes a step to resolve any deictic references the user has made.

In addition to the user's utterance and dialogue history, all gestures for the current turn are sent to the CR server. All of this contextual information can then be utilized to make the most appropriate resolutions of all the deictic references. The context-resolved semantic frame is finally sent to the dialogue manager, where an appropriate reply to the user is formulated.

The simulation of such an interaction cycle has been facilitated by the use of a Batchmode server, developed by Polifroni and Seneff (2000). The server receives an input (e.g. the text representation of the spoken utterance) from a file of logged or pre-formatted data. After the input has been processed, the next input is obtained from the input file, and the cycle continues (more details in Section 5).

# 4 Data Generation

## 4.1 Application domain

The original data used for generating multimodal simulated inputs are taken from the log files of the Voyager application. The Voyager application provides information about city landmarks (e.g. universities, museums, sport arenas, subway stops), gives navigation guidance and up-to-date traffic information over the phone and via a graphical interface. Geographically it covers the area of Boston and Cambridge in Massachusetts. Users can use natural language in the queries and dialogue management takes care of user-friendly disambiguation, error recovery and history handling. A typical dialogue between Voyager (V) and a user (U) is given below:

U: Can you show me the universities in Boston?
V: *Here is a map and list of universities in Boston…*

U: What about Cambridge?
V: *Here is a map and list of universities in Cambridge…*

U: How do I get there <click Harvard> from here <click MIT>?
V: *Here are directions to Harvard University from MIT…*

## 4.2 Defining a user population

As mentioned earlier, the data to be generated can be used both for testing and for system development. In both scenarios, real dialogues should be simulated as closely as possible. Therefore a virtual user population was defined for each experiment.

First, the distribution of various user types was defined. A user type is specified in terms of the delay a user exhibits with the Virtual Modality data delivery, relative to the speech channel. The following six user types were defined: outspoken, precise, too-fast, quickie, slowhand and everlate. *Outspoken* is an imaginary user who never uses the Virtual Modality, and communicates with the system using only the speech modality. *Precise* always issues the Virtual Modality input in synchrony with the spoken deictic reference. *Too-fast* always issues the Virtual Modality input significantly earlier than the corresponding deictic reference in the speech channel, while *Quickie* issues the Virtual Modality input only slightly earlier than the deictic reference. Similar rules apply for *Slowhand* and *Everlate*, except that they issue the Virtual Modality input slightly later or much later, respectively, than the deictic reference.

Once the composition of the user population has been determined, the corresponding temporal deviations must be specified. In a real system the exact instances are typically given as elapsed time from a reference point specified by a universal time value (with different devices synchronized using the Network Time Protocol). However, such accuracy is not necessary for the experiments. Rather, a simplified measurement is introduced in order to describe intuitively how the Virtual Modality input deviates from the instant when the corresponding deictic reference was issued. The unit used here is a word distance, more precisely the average length of a word (how many words are between the deictic reference and the input in the Virtual Modality channel). A 0 means that the deictic reference and the Virtual Modality event are in synchrony, while a –1 (+1) means that the Virtual Modality input was issued one word earlier (later) than the corresponding deictic reference.

Using this formalism, the following deviation pattern for the five user types is defined as a starting point for the experiments:

| | |
|---|---|
| Too-fast | -2 |
| Quickie | -1 |
| Precise | 0 |
| Slowhand | +1 |
| Everlate | +2 |

**Table 1.** Temporal deviation parameters for the user types that use the Virtual Modality.

## 4.3 Generation of the multimodal data

Generating multimodal data is, in a sense, the reverse process of the multimodal integration step. Since it is known how the deictic references are realized in a given domain, generating sentences with deictic references

once the actual definite phrases are found, seems straightforward.

The idea is simple: find all instances of a given type of semantic unit (e.g., location reference) in the input sentences, move them to the Virtual Modality channel with timing information and, as a replacement, put sensible deictic references back into the original sentences.

The implementation, however, reveals several problems. First, identification of the location references is not necessarily an easy task. It may require a complex parsing or keyword-spotting algorithm, depending on the application in question. In our case, the log files include the output of the TINA Natural Language Understanding module, meaning that all semantically relevant units present in an input sentence are marked explicitly in the output parse frame (Seneff, 1992). Figure 3 gives an example of the parse frame.

```
{c directions
  :subject 1
  :domain "Voyager"
  :input_string " give me directions from harvard to mit "
  :utterance_id 6
  :pred {p from
        :topic {q university
              :name "Harvard University" } }
  :pred {p to
       :topic {q university
       :name "Massachusetts Institute of Technology" } } }
```

**Figure 3.** Parse frame for the input sentence "*give me directions from harvard to mit*".

The movement and time marker placement step represents no problem.

The third step, namely the replacement of the removed semantic units with sensible deictic references, requires certain manipulation. Performing the replacement using only deictic references, such as, "here", "over here", "there", and "over there", would result in a rather biased data set. Instead, depending on the topic of the location reference (e.g., city, road, university), definite noun phrases like "this city" and "that university" were also used. Eventually, a look-up table was defined which included the above general expressions, as well as patterns such as "this $" and "that $" in which the variable part (i.e., $) was replaced with the actual topic. The selection for a sentence was randomly chosen, resulting in good coverage of various deictic references for the input sentences. For the example depicted in Figure 3, the following sentence is generated:

"*give me directions from there to this university*"

The following is a summary of the overall multimodal data generation process:

1. Define the distribution of the user population (e.g., outspoken 20%, precise 40%, quickie, 20%, slow-hand 15%, everlate 5%);
2. Define the corresponding deviations (see Table 1);
3. Randomly allocate turns (sentences) to the pre-defined user types (e.g. 40% of all data will go for the precise user type with deviation 0);
4. Identify all location references in the input sentence based on the parse frame;
5. Remove all or a pre-defined quantity of location expressions from the original sentence and replace them with deictic markers;
6. Place the removed location phrases into the Virtual Modality channel;
7. Place time markers to the Virtual Modality channel referring to the original position of the location phrases in the input sentence;
8. Issue the pre-determined time shift, if needed, in the Virtual Modality channel;
9. Randomly select an acceptable deictic reference and insert it into the original sentence in place of the deictic marker;
10. Repeat 4-9 until all data has been processed.

An example of the generated Virtual Modality data and the corresponding sentence is shown in Figure 4.

### 4.4 Statistics

Table 2 below details some statistics on data obtained from Voyager's original log files.

| | |
|---|---|
| Number of sessions | 1099 |
| Number of turns | 6077 |
| Average number of turns per session | 5.53 |
| All location references | 6982 |
| Average number of references per turn | 1.14 |
| Number of different location reference patterns | 203 |
| The five most frequent location reference patterns: | |
| *in* <….> | 9.95% |
| *of* <….> | 8.15% |
| *show_me* <….> | 6.39% |
| *on* <….> | 5.69% |
| *from* <….> *to* <….> | 3.95% |

**Table 2.** Overview of the original Voyager data (turn = sentence).

Although the above table covers the original data, the newly generated Virtual Modality database has the same characteristics since the location references there become deictic references.

## 5 Experiments

The experimental setup is depicted in Figure 4. The core of the system is the Galaxy Communicator architecture extended with the Batchmode server (as explained in Section 3 and shown in more details in Figure 2). It must be noted that although the sentences are taken from dialogues, each sentence is processed independently so the focus of attention is the new aspect introduced by the Virtual Modality.

There are two runs for each experiment. First, the original sentences are input to the Batchmode server and then passed to the Voyager application via the Galaxy architecture. The outcomes are the corresponding frames from the Language Understanding server (the Context Resolution server is not invoked due to the absence of context in this case). The second run takes the Virtual Modality data, namely the new sentences with the deictic references and the accompanying data for the Virtual Modality channel (semantic value, begin-end markers). The output frames are produced by the Language Understanding module and further processed by the Context Resolution server to resolve deictic references.

The last step of the execution is the comparison of the frame pairs: one frame for the original sentence and the other for the Virtual Modality data.

The results presented below are from the very initial tests; clearly more work is needed to justify the concept of Virtual Modality, as well as to fully investigate the utilization of the generated data in testing.

The initial experiments were run on 436 sentences, which represent a small portion of the entire database. The results indicate that if only one deictic reference per sentence is used with zero deviation, the generated output frames are identical to the original sentence output frames in 82.03% of the cases. The erroneous results occurred when the preposition and a chosen deictic form together formed an ungrammatical expression (e.g. "*how about on over there?"*). The data generation process requires further refinements to decide whether a preposition can be used with a randomly selected deictic expression.

In sentences with two deictic references only 78.26% agreement was achieved. The major reason for this is the incorrect replacement of highways and highway numbers with deictic references by the data generation process. Also, awkward combinations of deictic references result in incorrect resolution. All these problems will be addressed in future work.

Additionally, since the current version of the Context Resolution server has no built-in time limits for resolving deictic references, future work will aim to incorporate some kind of temporal considerations and adaptivity. The Virtual Modality data creation process supports the generation of a large amount of time-shifted versions of the original data, which can be used for further testing of the system's temporal robustness.

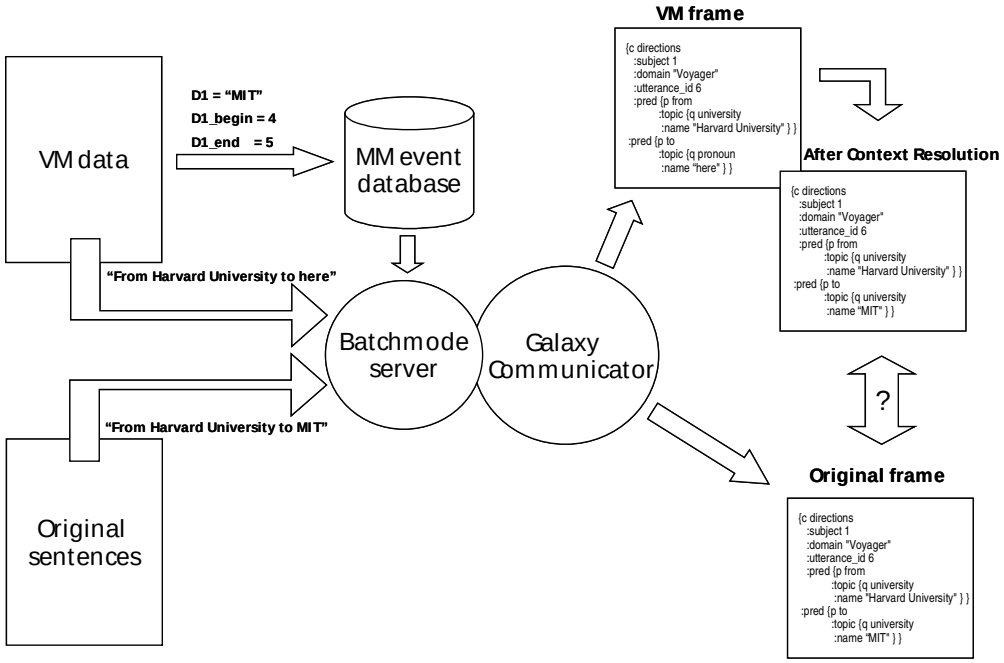


**Figure 4.** The evaluation procedure

## 6 Summary and Future Work

An experimental framework has been introduced, called Virtual Modality, which aims to assist in the development and testing of multimodal systems. The paper explained the motivation behind generating (simulated) multimodal data from available textual representations of logged user inputs. The procedure of replacing location references with deictic expressions, as well as the placement of the referenced semantic value along with temporal information to the Virtual Modality channel were explained. Lastly, the evaluation scenario and preliminary test results were presented.

Future work will investigate the following topics:

- how the generated Virtual Modality data can be utilized to train a statistical (or hybrid) multimodal integration module;
- how adaptivity in the integration module can be achieved using a vast amount of training data;
- how N-best choices in the Virtual Modality input can be utilized;
- whether disambiguation can be achieved with training examples covering all temporal cases;
- how evidence in one modality can help to resolve some ambiguity in the other modality, and ultimately, how to provide an accurate interpretation of the overall user intention.

## 7 Acknowledgements

Thanks to Stephanie Seneff for her comments, to Mitchell Peabody for setting up the database server and to D. Scott Cyphers for his always-available help with the Galaxy system.

The first author acknowledges the financial support from the Nokia Foundation, Suomen Akatemia and the Ella and Georg Ehrnrooth Foundation.

## References

Michael H. Coen. 2001. "Multimodal Integration A Biological View." *17th International Joint Conference on Artificial Intelligence, IJCAI 2001*, Seattle, Washington, USA, August 4–10, pp. 1417–1424.

Edward Filisko and Stephanie Seneff. 2003. "A Context Resolution Server for the Galaxy Conversational Systems." *Eurospeech'2003*. Geneva, Switzerland, September 1–4, pp. 197–200.

J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue. 1995. "Multilingual Spoken-Language Understanding in the MIT Voyager System," *Speech Communication*, 17(1-2):1–18.

Pentti O. Haikonen. 2003. *The Cognitive Approach to Conscious Machines*. Imprint Academic, Exeter, UK.

Saija-Maaria Lemmelä and Péter Pál Boda. 2002. "Efficient Combination of Type-In and Wizard-of-Oz Tests in Speech Interface Development Process." *ICSLP'2002*, Denver, CO, September 16–20, pp. 1477–1480.

Ivan Marsic and Bogdan Dorohonceanu. 2003. "Flexible User Interfaces for Group Collaboration". *International Journal of Human-Computer Interaction*, Vol.15, No.3, pp. 337-360.

Sharon Oviatt, Antonella DeAngeli and Karen Kuhn. 1997. "Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction." *Conference on Human Factors in Computing Systems, CHI '97*. ACM Press, New York.

Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson, and Lesley Carmichael. 2003. "Toward a Theory of Organized Multimodal Integration Patterns During Human-Computer Interaction." *5th International Conference on Multimodal Interfaces, ICMI'2003*, Vancouver, British Columbia, Canada, pp. 44–51.

Joseph Polifroni and Stephanie Seneff. 2000. "Galaxy-II as an Architecture for Spoken Dialogue Evaluation." *2nd International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Stephanie Seneff. 1992. "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, 18(1):61–86.

Stephanie Seneff. 2002. "Response Planning and Generation in the MERCURY Flight Reservation System," *Computer Speech and Language*, 16:283–312.

Stephanie Seneff, Chian Chuu, and D. Scott Cyphers. 2000. "ORION: From On-line Interaction to Off-line Delegation." *ICSLP'2000*, Beijing, China, October, pp. 142–145.

Stephanie Seneff, Ed Hurley, Raymond Lau, Christine Pao, P. Schmid, and Victor Zue. 1998. "Galaxy-II: A Reference Architecture for Conversational System Development." *ICSLP'1998*, pp. 931–934.

Sy Bor Wang. 2003. *A Multimodal Galaxy-based Geographic System*. S.M. thesis, MIT Department of Electrical Engineering and Computer Science.

Victor Zue, Stephanie Seneff, Joseph Polifroni, Michael Phillips, Christine Pao, D. Goodine, David Goddeau, and James Glass. 1994. "PEGASUS: A Spoken Dialogue Interface for On-line Air Travel Planning," *Speech Communication*, 15(3–4):331–340.

Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and I. Lee Hetherington. 2000. "JUPITER: A Telephone-based Conversational Interface for Weather Information," *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.