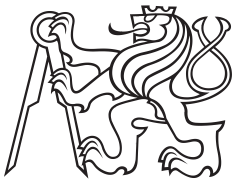
Doctoral Thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Telecommunication Engineering

Response Time Improvement of Multimodal Interactive Systems

Ing. Roman Hák

Supervisor: Ing. Tomáš Zeman, Ph.D.
Ph.D.Programme: Electrical Engineering and Information Technology
Branch of study: Telecommunication Engineering

Prague, February 2017

Acknowledgements

I would like to thank Dr. Tomáš Zeman for his guidance, supervision and helpful comments and advices during the course of my doctoral study.

My thanks also go to all colleagues from the Department of Telecommunication Engineering at the CTU in Prague for creating a pleasant, friendly and inspiring environment. Special thanks belong to Prof. Boris Šimák for his leadership and support in research work.

Finally, I would like to thank my parents, family and all friends that provided me with support, understanding and motivation during my studies and beyond.

Declaration

I hereby submit for the evaluation and defence the disertation thesis elaborated at the CTU in Prague, Faculty of Electrical Engineering.

I declare I have accomplished my final thesis by myself and I have named all the sources used in accordance with the Guideline on ethical preparation of university final theses.

February 20, 2017
Prague, Czech Republic

Roman Hák

Abstract

Multimodal interaction, permitting our highly skilled and coordinated communicative behavior to control computer systems, has been proven as a key to natural and very flexible human-computer interaction. However, multimodal input processing submits great research and development challenges in contrast to the traditional user interfaces. Besides processing of complex input signals from individual modality sensors (e.g. speech recognition, image processing, etc.), it also requires more detailed understanding of human communication paradigms and interaction schemes.

The submitted thesis deals with an analysis of users' integration patterns observed during multimodal interaction and explores possibilities of their utilization to increase accuracy and robustness of algorithms for multimodal input processing.

The work contains three main parts. The first one is dedicated to an analysis of the most fundamental multimodal integration patterns, which is followed by a quantitative research and evaluation of import characteristics of the patterns in the form of own conducted user study. In the context of the new findings, a definition of a classification of one of the most important integration patterns, i.e. synchronization pattern dividing users to simultaneous (SIM) and sequential (SEQ) integrators, is modified and readjusted. The modified classification addresses issues with consistency and accuracy and offers a significantly superior solution to the original definition provided in the related literature.

Based on the evaluations and results obtained in the quantitative empirical research, a method for multimodal integration patterns modeling with utilization of machine learning algorithms, namely Bayesian Networks, is designed and proposed in the following part of the thesis. The constructed probability model is capable of very precise and robust multimodal input prediction with accuracy of 99%.

A procedure for applying the predictive capabilities of the constructed classification model to address the multimodal input segmentation is then introduced. The proposed procedure is subjected to tests and measurements in order to evaluate the segmentation accuracy and impact of the procedure employment on response time of the system. Experiments with a selection of training sets and a comparison of four approaches to encode continuous input variables in the model are conducted as a part of the measurements. The results show that the introduced segmentation method provides a significant improvement in response time (to 0.8 s for SEQ and under 0.5 s for SIM integrators) over the state-of-the-art approaches, while maintaining remarkably high accuracy (98–99%). This significant decrease in response time allows a system to respond more instantly on user's multimodal input with nearly real-time feedback and brings very important improvement in terms of usability, which should positively influence users' experience and satisfaction with the multimodal interaction interface.

Keywords: multimodal interaction, input segmentation, integration patterns, user modeling, response time

Abstrakt

Multimodalní interakce umožňuje plně využít naše velmi zdatné a vysoce koordinované komunikační schopnosti k ovládání počítačových systémů. Představuje tak cestu k přirozené a velmi flexibilní interakci člověka s počítačem. Zpracování multimodálního vstupu však oproti tradičním uživatelským rozhraním představuje mnohé náročné výzkumné i vývojové úkoly. Kromě zpracování složitých signálů od jednotlivých senzorů (např. rozpoznání řeči, obrazu apod.) vyžaduje také mnohem detailnější znalost a porozumění lidským komunikačním paradigmatům a interakčním schématům.

Předložená práce se zaobírá analýzou uživatelských integračních vzorců pozorovaných při multimodální interakci a zkoumá možnosti jejich využití ke zvýšení přesnosti a robustnosti algoritmů pro zpracování multimodálních vstupů.

Práce obsahuje tři stěžejní části. První z nich je věnována analýze nejpodstatnějších multimodálních integračních vzorců, na kterou navazuje kvantitativní výzkum důležitých charakteristik těchto vzorců v podobě vlastní uživatelské studie. V rámci nově získaných poznatků je modifikována definice pro klasifikaci jednoho z nejdůležitějších vzorců, tj. synchronizační vzor dělící uživatele na simultánní (SIM) a sekvenční (SEQ) integrátory. Nová klasifikace řeší zejména problémy v konzistenci a přesnosti, a významně tak kvalitativně přesahuje původní definici uváděnou v související literatuře.

Na základě zjištění a výsledků dosažených v rámci kvantitativního výzkumu je v další části práce navržena metoda pro modelování multimodálních integračních vzorců pomocí algoritmů strojového učení, jmenovitě Bayesovských sítí. Zkonstruovaný pravděpodobnostní model poskytuje velmi přesnou a robustní predikci multimodálního vstupu dosahujícího 99% úspěšnosti.

Následně je popsán postup aplikování predikčních schopností modelu při řešení segmentace spojitého multimodálního vstupu. Představená metoda je podrobena testům a měřením s ohledem na přesnost a dopad jejího použití na zlepšení doby odezvy systému. V rámci měření jsou provedeny experimenty s volbou trénovací množiny a porovnání čtyř přístupů ke kódování spojitých vstupních proměnných v modelu. Výsledky ukazují, že navržená metoda poskytuje významné zlepšení v době odezvy systému (0,8 s pro SEQ a pod 0,5 s pro SIM integrátory) v porovnání s nejmodernějšími publikovanými postupy při zachování pozoruhodně vysoké přesnosti (98–99 %). Toto výrazné snížení umožňuje systému zareagovat na multimodální uživatelský vstup s odevzvou téměř v reálném čase. Přináší tak důležité zlepšení ve smyslu použitelnosti, které by mělo pozitivně ovlivnit celkovou uživatelskou zkušenost a spokojenost s multimodálním interakčním rozhraním.

Klíčová slova: multimodální interakce, segmentace vstupu, integrační vzory, modelování uživatelů, doba odezvy

Překlad názvu: Snížení doby odezvy u multimodálních interaktivních systémů

Contents

Figures

Tables

Abbreviations

3D three-dimensional

ASR automatic speech recognition

BBN Bayesian Belief Network

DARPA Defense Advanced Research Projects Agency

DTM Dynamic Time Windows

FG fine-grained

GUI graphical user interface

HCI human-computer interaction

HMM Hidden Markov Models

NLU natural language understanding

PDA personal digital assistant

SEQ sequential

SIM simultaneous

UI user interface

WIMP windows, icons, menus, pointer

WoZ Wizard of Oz

Chapter 1

Introduction

Natural communication between a human and computer has been always a great challenge for researchers, developers and designers of interactive systems. While significant advances to improve speech and gesture recognition performance, natural language understanding (NLU), and more recently, motion tracking and image processing performance as well have been made in recent years, the systems capable of natural human-computer interaction (HCI) have still not found widespread acceptance in everyday life. An important reason for this is the inflexibility and impropriety of each input mode (or modality) when used alone. The key to permitting our highly skilled and coordinated communicative behavior to control computer systems is utilization of a multiplicity of communication channels and signals working in concert to supply complementary information or increase robustness with redundancy [VW96]. To address these issues and offer more natural, flexible, transparent and efficient interaction between a human and computer, a multimodal interaction has to be employed.

From the research perspective, multimodal interaction represents a new direction in the field of HCI and a paradigm shift away from conventional graphical user interfaces (GUIs), also referred to as WIMP (windows, icons, menus, pointer) systems. Besides processing of complex input signals from individual modality sensors (e.g. speech recognition, image processing, etc.), difficult research and development challenges are also covered behind multimodal fusion (or integration). The fusion is a complex process responsible for integration of multiple related inputs into an integrative interpretation expressing real intents of a user and requires more detailed understanding of human communication paradigms and interaction schemes in contrast to the conventional GUIs.

1.1 Advantages of Multimodal Interaction

As already stated, multimodal interaction offers more natural, flexible, transparent and efficient communication between a human and computer. There are, however, other advantages, and myths as well, associated with the multimodal interaction and multimodal interactive systems.

Many empirical user studies related to multimodal interaction have been conducted in order to guide the design of multimodal interfaces. For example, Johnston et al. in [JCM+97] prepare a user study on a map-based application where users can perform the same tasks using only speech, only pen or a combination of both (i.e. multimodally). Users' multimodal inputs resulted in 10% faster task completion, 23% fewer words, 35% fewer spoken disfluencies, and 36% fewer task errors compared to unimodal spoken input. Another interesting fact was that all, or 100%, of users indicated a preference for multimodal interaction over unimodal (speech-only or pen-only interaction). Based on this study, other related researches (e.g. [ODK97, CJM+97a, Ovi99a]) and extensive experiences in the area, Prof. Oviatt [Ovi99b] exposes common engineering myths regarding how people interact multimodally. The myths are quoted in the following list:

1. If you build a multimodal system, users will interact multimodally.
2. Speech and pointing is the dominant multimodal integration pattern [or combination].
3. Multimodal input involves simultaneous signals.
4. Speech is the primary input mode in any multimodal system that includes it.
5. Multimodal language does not differ linguistically from unimodal language.
6. Multimodal integration involves redundancy of content between modes.
7. Individual error-prone recognition technologies combine multimodally to produce even greater unreliability.
8. All users' multimodal commands are integrated in a uniform way.
9. Different input modes are capable of transmitting comparable content.
10. Enhanced efficiency is the main advantage of multimodal systems.

Considering these myths helps significantly to avoid wrong assumptions and misunderstandings when prototyping, developing or constructing multimodal systems.

More recently, research has also focused on designing and deploying multimodal interfaces. Following this trend, Reeves et al. [RMM+04] proposed and defined the *"guidelines for multimodal user interface design"*. The most important of them are outlined in the list bellow.

- Multimodal systems should be designed for the broadest range of users and contexts of use, since the availability of multiple modalities supports flexibility. For example, the same user may benefit from speech input in a car, but pen input in a noisy environment.
- Designers should take care to address privacy and security issues when creating multimodal systems: speech, for example, should not be used as a modality to convey private or personal information in public contexts.
- Modalities should be integrated in a manner compatible with user preferences and capabilities, for example, combining complementary audio and visual modes that users can co-process more easily.
- Multimodal systems should be designed to adapt easily to different contexts, user profiles and application needs.
- Error prevention and handling is a major advantage of multimodal interface design, for both user- and system-centered reasons. Specific guidelines include integrating complementary modalities to improve system robustness, and giving users better control over modality selection so they can avoid errors.

These guidelines make a primary overview and summarization of the most significant aspects of multimodal interactive systems.

1.2 Multimodal Input Processing

Successful processing of multimodal input signals and their accurate fusion into an integrative interpretation of users' intents is a very complex task, which submits substantial requirements to design and development of the system. Thus, an advanced system architecture is fundamental. We explored

advanced architectures of multimodal interactive systems in our previous research and recently proposed a framework Manitou [HZ13, HDZ12]. It is a feature rich and flexible application framework for rapid prototyping and development of multimodal interfaces.

A basic architecture of a typical multimodal system (built on top of the Manitou framework) is depicted in Figure 1.1. A detailed description of the most important parts is discussed in the following paragraphs (more details can be found in [HZ13]).

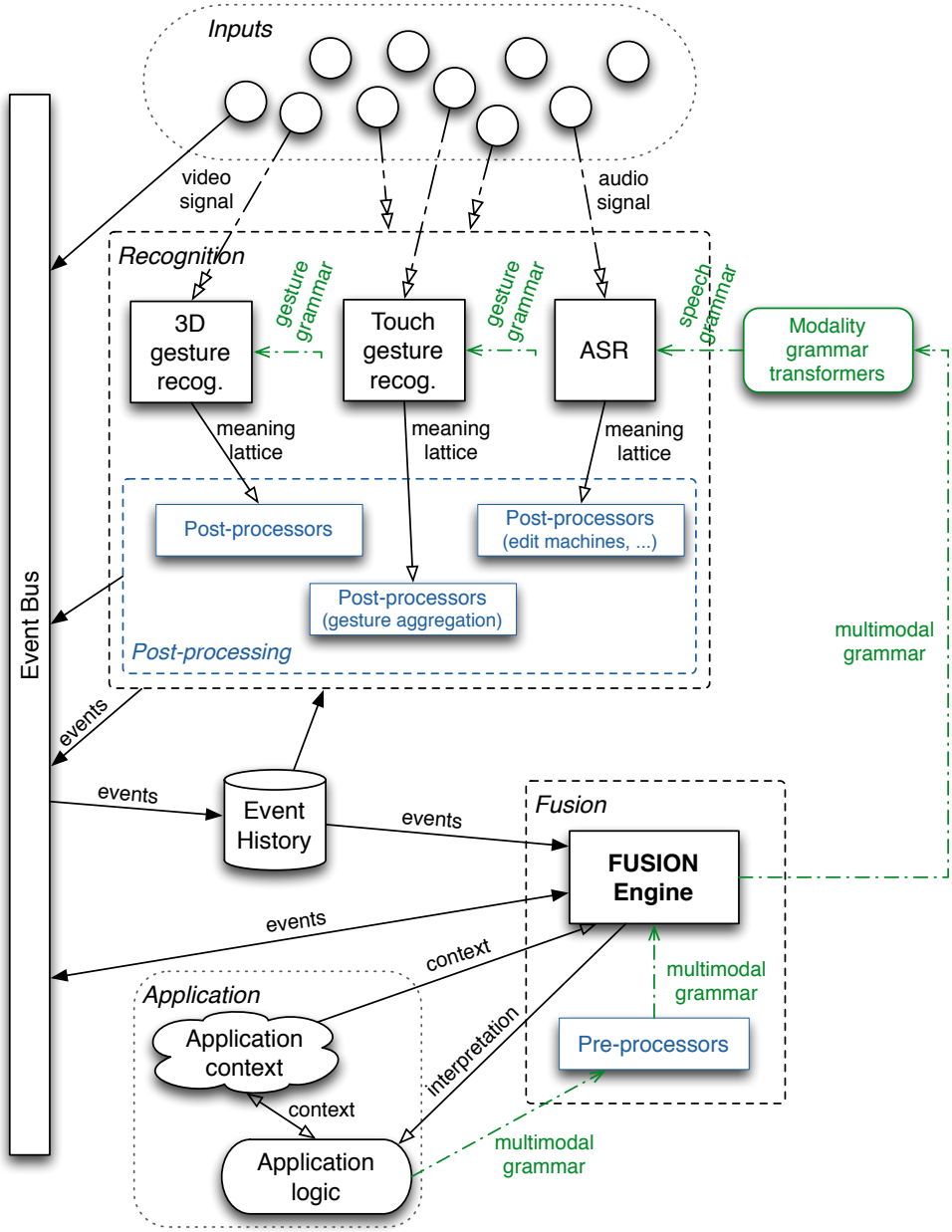


Figure 1.1: A basic architecture of a typical multimodal interactive system.

1.2.1 Input Processing & Recognition

The processing initiates in input devices where diverse signals (acoustic, visual, etc.) and events (e.g. a key press or a pointer movement) are captured. The input data is then transferred to appropriate recognizers for processing (e.g. a automatic speech recognition (ASR), three-dimensional (3D) gesture recognition etc.). It involves extracting of relevant features and results in an assignment of derived interpretations. The results are packed and wrapped into relevant event objects and sent to an *event bus* or passed on for additional processing (*post-processing*).

1.2.2 Communication & Events

A central communication platform across the system is realized through the *event bus*. Every important data flow within the system travels over this bus in a form of events. Individual components can observe, monitor and utilize information of events arising in different levels of processing. At the same time, components are able to create and sent events notifying about new facts obtained within their activity.

1.2.3 Multimodal Fusion

Multimodal fusion (or integration) is the most important stage and one of the core procedures in the multimodal input processing. It is responsible for fusion (integration) of input data received from multiple input sources (e.g. sensors and other input devices). A result of the procedure is an integrative interpretation that is assigned to a combination of inputs belonging to a single interaction. Such an interaction combination is called a *multimodal unit* or *group*. The units typically consist of two or more inputs, but they can also contain data from only a single input source. In that case the interaction is referred to as *unimodal*. A component in the multimodal system responsible for this activities is typically referred to using the term *fusion engine* (other different terms can be found in less related literature — for more details on terminology see [LNP+09]).

Fusion methods can be classified by a level of processing of input data they operates on. According to Sharma et al. [SPH98], there are three distinctive classes of multimodal fusion methods:

- *Data-level* – the fusion is performed on "uprocessed", or "raw", data obtained directly from the input sensors.
- *Feature-level* – input sensory data are analyzed for features before the fusion is performed. The extracted features are used for the fusion instead of the original data.
- *Decision-level* – this type of fusion is based on the integration of individual mode decisions or interpretations. For instance, a hand gesture is interpreted as a deictic gesture with information about a particular object located at the pointed coordinates.

The data-level fusion is rare in multimodal interaction since the data from individual modalities are typically of a different nature (e.g. speech and gestures) originating from different types of sensors (e.g. a camera, microphone, accelerometer, etc.), and therefore it is not possible to fuse them effectively in a raw form. The feature- and decision-level fusions are more common in HCI with the latter being the most frequently used.

In multimodal systems, the fusion engine is directly controlled by the *application logic*, which provides a set of supported inputs and input combinations typically in the form of multimodal grammars (e.g. [Joh98, JB05, DLI10]). Various methods of pre-processing can be performed before the fusion engine parses the grammar. A common task of pre-processing is enrichment of the grammar leading towards greater robustness of the system and reduction of grammar complexity resulting in lower demands on its development. An example of such pre-processor within the framework is a component that searches the lexical database (e.g. WordNet[1]) for synonyms, which are then inserted to the original grammar as alternative tokens.

A number of techniques for the multimodal fusion processing have been proposed and introduced in the related research. Their fusion strategies are based on one of the following approaches:

1. Frame-merging [VW96]
2. Unification of feature structures [JCM+97, Joh98]
3. Finite-state machines processing multimodal context-free grammars [JB05]
4. Machine learning and statistical approaches [DSL12]

[1] http://wordnet.princeton.edu/

1.2.4 Multimodal Input Segmentation

In addition to the fusion itself, there is another essential task that has to be addressed within the procedure of integration. That task is *multimodal input segmentation* and its objective is to determine temporal relations between sequential and/or parallel inputs from multiple modalities, or to *segment* them, into either integrative multimodal or separate unimodal units. The process of input segmentation is illustrated in Figure 1.2.

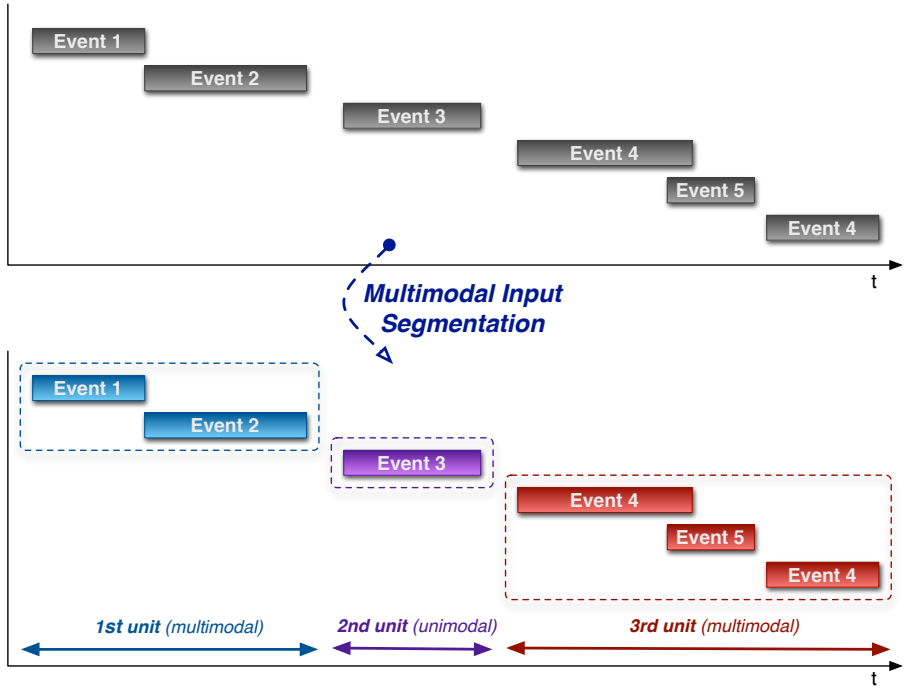


Figure 1.2: Process of multimodal input segmentation.

The input segmentation is crucial for two main reasons. First, accuracy of the segmentation affects an error rate of a consecutive fusion process, and second, decision latency (or a delay) affects response time of the system. Both properties are very important in terms of reliability and usability of a multimodal system, which directly influences experience and satisfaction of users and ultimately their general willingness to use the system.

The *decision delay* (or *wait period*) is a fundamental part of the input segmentation and is deliberately included in the process in order to accurately determine, which inputs belong to the current and which ones to the next unit in the series of multimodal input events (see Figure 1.3). Without their employment the multimodal units would be segmented prematurely (see the problem of *under-segmentation* in the following part).

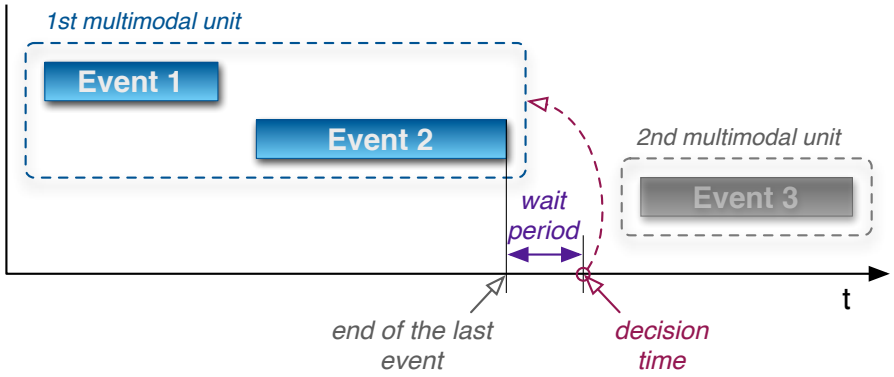


Figure 1.3: Illustration of decision delay (or wait period) in multimodal input segmentation.

Over- and Under-Segmentation

There are two primary problems associated with selection and assessment of a wait period. Both problems are illustrated in Figure 1.4. The first one, *over-segmentation* (1.4b), occurs when the wait period is too short and a segment (multimodal unit) is ended prematurely due to a longer time interval between related events.

The second problem is *under-segmentation* (1.4c) and occurs if the wait period is conversely too long. In this case, a multimodal unit is incorrectly extended with events from the next unit since an interval between unrelated events is shorter than the wait period. Both problems lead to an incorrect segmentation and eventually to an improper interpretation of users' intents, as a consequence.

Under-segmentation is typically a less severe issue, since users, thanks to their natural intelligence [Ovi99b], intuitively start to increase pauses between multimodal units when confronted with this type of error while interacting with the system. This fact could create a wrong assumption that resolving multimodal input segmentation only requires the selection of satisfactory long interval for the wait period to avoid over-segmentation. Unfortunately, the increase of the wait period comes with another negative and undesirable effect — growth of response time. The response time is a very important property of interactive systems in terms of usability and users' experience. Therefore, there should be an effort to minimize the response length — and thus the wait period — in order to lessen its negative impact on usability.

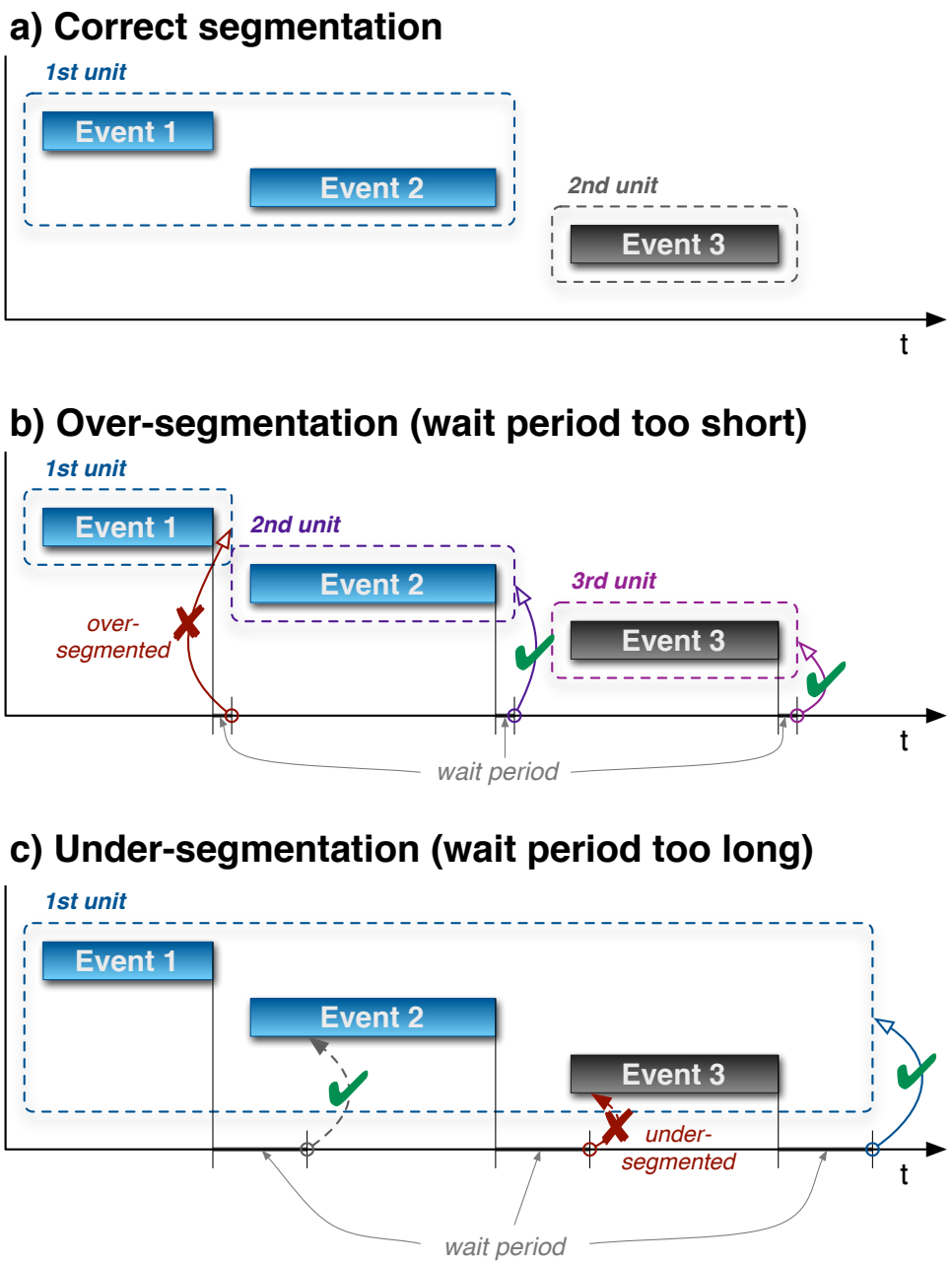


Figure 1.4: Problem of *over-* and *under-segmentation*. The top figure (a), shows a correctly segmented inputs. Figure b) illustrates an example of *over-segmentation* (due to too short wait period) and the bottom figure (c) demonstrates *under-segmentation* (wait period is too long).

1.3 Thesis Organization

The thesis is organized as follows. In the next chapter, related works from the perspective of individual topics covered in this thesis are provided. Chapter 3 describes primary motivations and main objectives behind the research presented in this work. An analysis of the most important multimodal integration patterns and details of our user study on the topic are introduced in Chapter 4. The next chapter (Chapter 5) is dedicated to modeling of the integration patterns using a combination of machine learning and probabilistic graphical models. A description of an application of the developed model to improve response time of multimodal interactive systems follows in Chapter 6. The final part presents general conclusions of the thesis and outlines future directions of the research.

Chapter 2

Related Work

Basic principles and ideas behind multimodal interaction were first broadly demonstrated to the research community in Bolt's seminal work "Put-that-there" [Bol80]. In the following years, rather engineering approaches prevailed in the designing of new methods and algorithms involved in integration, synchronization and interpretation of multiple input modes. In this period, scientific research was primarily driven by experience gained in related domains (e.g. computational linguistics) yielding in the development of a number of false presumptions, misunderstandings, and myths [Ovi99b].

2.1 Empirical Evidence of Individual Differences and Integration Patterns

It took almost two decades since the Bolt's work [Bol80] until the appearance of the first studies investigating individual differences and variations in users when interacting with multimodal systems. A research of pivotal importance in the perspective of integration and synchronization of multimodal input was introduced by Oviatt et al. [ODK97]. Most notably, the authors identified *sequential (SEQ)* and *simultaneous (SIM)* multimodal integration patterns in adult subjects. This started a series of consecutive research, where new and previously unobserved interaction phenomena were discovered. Xiao et al. followed the aforementioned work and analyzed speech and pen-based multimodal integration patterns first in the case of children [XGO02] and later in seniors [XO03].

Another study demonstrated that the dominant integration pattern remains resistant to change (with the consistency of 97%), even when strong reinforcement is delivered to encourage a user to switch the pattern [OCT+03]. Instead, further *entrenching* of the dominant pattern was recognized (i.e. increasing the intermodal *lag* for the SEQ integrators and *overlap* for the SIM integrators). Similar conclusions were confirmed in a more recent study [HPSM11]. In [OCL04], authors shown the ability of flexible multimodal interfaces to support users in managing cognitive load as the frequency of multimodal interactions substantially increased over unimodal with the task difficulty. Evidence of stability and persistence of the integration patterns over time was provided in [OLC05]. Authors collected data from tested subjects for more than six weeks and observed 95–96% consistency of their dominant integration pattern.

In a more recent experimental study, Schüssel et al. [SHS+14] used individual user interaction behavior for error detection and recovery. The researchers reported significant inconsistency of the simple classification scheme (i.e. SIM and SEQ integrations as described by Oviatt et al. [ODK97, OLC05]) in their experimental multimodal application combining speech and touch input. Instead, they introduced different metrics and temporal distributions derived from onsets and offsets between modality combinations derived from the previous interactions (stored in an "Interaction History") to detect errors, resolve conflicts, and recover from them. Applying the proposed approach, they were able to improve the robustness of the system and reduce initial error rate from 4.9% to a minimum of 1.2%. The reported inconsistency surrounds the previous evidence with controversy and suggests that the classification as defined by Oviatt et al. could be oversimplified and its appropriateness may be limited to a specific task or multimodal input combination.

2.2 Prototypes and Demonstrational Systems

Employing multimodal interaction in visual/spatial domains proved to offer numerous performance advantages over unimodal interaction [Ovi99b, Ovi03a]. In addition, the portion of multimodal constructions produced by users is higher in these domains in comparison with others. Therefore, map-based interfaces and systems were commonly used in the past to test and study multimodal interaction and also to demonstrate its principles.

One of the first full-featured multimodal interactive systems, QuickSet, was developed by Cohen et al. [CJM+97a, CJM+97b]. It was a prototype of a distributed interactive application applied to collaborative military training

used to control a simulator and a 3D virtual terrain visualization system using pen/voice input. A derivative of QuickSet's multimodal technology was recently used in Sketch-Thru-Plan [CKB+15], an advanced command and control system enabling rapid creation of operational plans for military ground operations supported by the Defense Advanced Research Projects Agency (DARPA).

MATCH (Multimodal Access To City Help) [JBV+02] represents another interesting testbed multimodal application running on personal digital assistants (PDAs) offering a city guide and navigation system that enables mobile users to access restaurants and subway information for New York City. Interaction is provided through speech and pen input (i.e. drawing on a display with a stylus) or by synchronous multimodal combinations of the two modes.

More recently, Speak4it [EJ12], a cloud-based mobile application, was claimed to be the first commercially deployed multimodal system. It supports true multimodal input through a combination of speech and touch-gestures on a map interface and offers local search capabilities. Only a very low usage rate of multimodal commands (only 3% compared to 19.2% previously reported by Oviatt et al. [Ovi99b]) were observed during the studied period. Authors provided a number of explanations for this phenomenon, including insufficient alerting and education of users about the multimodal functionality, a bad interface design conflicting with existing interaction paradigms used in other map-based mobile applications, or simply that multimodal commands were assigned to rather sporadically used functionalities.

Popularity of multimodal interaction in visual/spatial domains is declared by an increasing number of employments in virtual and augmented reality [BL12, LBB+13].

2.3 Testing methods

Regarding the testbeds and evaluation systems, researchers of multimodal integration patterns — and multimodal interaction in general — have several options to assess and verify their hypotheses. Two main approaches were frequently used for purposes of usability testing [Ovi03b]. The first and most straightforward is to create a fully functional prototype of a tested system. The second approach is to design only elementary parts of the system, which are simple to implement (typically a graphical user interface), and simulate

the complex and sophisticated functionalities (e.g. multimodal fusion engines, speech recognizers etc.). To this end the Wizard of Oz (WoZ) technique has regularly been used and adopted to the specifics of multimodal interaction [CJM+97a, OCL04, OLC05]. An interesting enhancement is the Dual-Wizard technique introduced and described in [OCT+03]. The simulation technique involves two "wizards". The first one (*Input Wizard*) observes, recognizes and records the user's synchronization pattern (i.e. SIM or SEQ). This information is then passed to the second one (*Output Wizard*), who monitores the content of the user's input, identifies the intended action, and responds with appropriate feedback.

In order to facilitate developing and testing of non-fully functional prototypes based on WoZ techniques, Serrano and Nigay offered OpenWizard [SN10], a component-based approach for rapid prototyping and testing built upon the OpenInterface toolkit.

Facilitating of rapid prototyping, which allows the quick assessment of the usability of prototype systems or their complex parts in early stages of a design process, is considered the most pronounced advantage of the WoZ simulation technique. Sometimes these techniques might be the only choice due to unfeasibility of important system components for researchers and practitioners. Nevertheless, the WoZ simulation could also bring some negative aspects to the testing. Most importantly it is the introduction of a human element ("wizard") into the experiment that could influence behavior of the system through unexpected feedback, variable response time, and other unpredictable human errors. All of these uncertainties could have impact on the final results, which could consequently draw misleading conclusions about the subjects of testing. Admittedly, many of these shortcomings can be avoided through good experiment design and by employing experienced researchers.

2.4 Multimodal Input Segmentation

Supplied with evidence from first empirical studies [ODK97], authors of early multimodal interaction prototypes (e.g. QuickSet [CJM+97a]) used fixed thresholds to find temporal compatibilities between corresponding inputs. Gupta and Anastasakos addressed the input segmentation using adaptive wait periods, Dynamic Time Windows (DTM) [GA04], which use a probabilistic method, Bayesian Belief Network (BBN), to predict properties of the next expected input. The wait period time is computed based on the predictions. Using DTM, the delay was reduced to 1.3 s from fixed time delays of 2 and 4

seconds employed in a user study by the authors. Reported accuracy of the predicted variables was around 80% when longer training periods were used ($10k$ epochs). Similarly, Huang et al. combined user modeling and machine learning to predict the multimodal integration patterns using BBN with an intention to develop *user-adaptive temporal thresholds* [HO06, HOL06]. Using these techniques the designed learning models correctly decided between unimodal and multimodal input with 85% accuracy and classified users' multimodal input as either *sequential* or *simultaneous* in 82% of cases. More recently, Miki et al. [MKM+14] used a probability distribution of time intervals between onsets of deictic gestures and the accompanying speech utterances to segment multimodal input. Of all utterances, 93.8% were correctly associated with gestures in their preliminary experiment.

A different solution in comparison to the previous works is presented in [KB06]. Rather than focusing on prediction of durations to wait for the end of user's turn, the proposed approach employs an underlying parsing mechanism to filter out input variants not acceptable by a given multimodal grammar using the *edge-splitting* technique. Likewise, there is a class of multimodal fusion methods based on machine learning techniques and modeling (e.g. Hidden Markov Models (HMM)) [DSL12] that handles the segmentation by design to some extent. The crucial disadvantage of these solutions resides in a limitation they impose on multimodal input grammars. The accepted grammar cannot contain any rule that is same as an initial part of another rule in the grammar. Instances of such rules representing two commands for zooming in in a map-based interface follows:

$$\begin{aligned} C_1 &\rightarrow \text{speech}(\textit{“zoom in”}) \\ C_2 &\rightarrow \text{speech}(\textit{“zoom in”})\ \text{gesture}(p_1) \end{aligned} \tag{2.1}$$

The first command consist of a single speech input whereas the seconds one is accompanied by a deictic gesture (p_1) pointing to a place of interest. The problem arises when a user wants to use the shorter command (C_1). Instead of simply accepting the rule C_1 after the user uttered "zoom in" the system would indefinitely wait for a deictic gesture since it cannot filter out the rule C_2. These situations are commonly resolved using fixed temporal thresholds, which brings us back to the problem of wait periods.

Chapter 3

Motivation and Objectives

3.1 Motivation

While accuracy and robustness of input modality sensors and recognizers as well as multimodal fusion algorithms and methods related to the input integration have reached a satisfactory level in the recent years, there is still one important property of multimodal input processing that have been rather overlooked. That is response time. As already mentioned in the introduction, the response time is a very important property in terms of reliability and usability of a multimodal system, which directly influences experience and satisfaction of users and ultimately their willingness to use the system.

There are two particular causes of the delays that affect response time in multimodal interactive systems:

1. *Sensory delays* – delays interconnected with individual input modality sensors and/or related signal processing (e.g. feature extraction, recognition or interpretation assignment etc.).
2. *Wait periods* – decision delays related to the segmentation of inputs into multimodal units.

Sensory delays are typically only tenths to few hundredths of a second long and could be possibly minimized by utilization of modern input sensors and

recognizers with real-time response time and/or by increasing of computational power of the system (or by optimization of signal processing algorithms).

The wait periods, on the other hand, are not directly dependent on computational efficiency of used processing algorithms or the underlying hardware. They form a fundamental part of the multimodal input segmentation and are deliberately included in the process in order to accurately determine the end of one multimodal unit and the beginning of another. Without their employment the multimodal segmentation would be incorrect due to the problem of over-segmentation (for more details see Section 1.2.4).

The essential objective of this work is to minimize response time of multimodal interactive systems caused by these wait periods. The emphasis also has to be given on robustness, reliability and accuracy of the provided solution. We plan to achieve the objective through the optimization of the multimodal input segmentation procedure.

The empirical evidence from various studies (see Section 2.1) revealed that there are differences between individual users in an approach they integrate inputs during multimodal interaction. Fortunately, a number of distinctive patterns and schemes were observed and identified in users' multimodal integration style. Importantly, the patterns were found very stable over time (with consistency of 97%) and resistant to change even when strong selective reinforcement was intentionally delivered [OCT+03, OLC05].

These facts suggest the particular advancements in the minimization of response time (and enhancements of reliability and robustness of the multimodal systems in general) could be gained through adaptation to the specific integration patterns and synchronization schemes of individual users.

3.2 Objectives

In the previous section, the essential task of this work was defined as the minimization of response time caused by the wait periods (or decision delays) related to the multimodal input segmentation. This task can be divided into a number of particular objectives as follows:

- To analyze the multimodal integration patterns and their characteristics, and, as a result of the process, identify the most promising candidates

that could be effectively employed in terms of the modeling of specific users' interaction behavior.

- To design and develop an accurate, reliable and robust interaction model capable to profit from adaptation to the multimodal integration patterns of diverse users.
- To apply the developed interaction model on optimization of a multimodal input segmentation process with focus to effectively improve response time (i.e. to decrease the wait periods).

This thesis deals with all of these objectives. They will be addressed in the following chapters in the respective order.

Chapter 4

Analysis of Integration Patterns in Multimodal Interaction

In this chapter, the most interesting multimodal integration patterns (see Section 2.1) will be introduced and described. Then, details of measurements and results of our user study of the integration patterns will be provided. The main objective for the study was to analyze and evaluate the integration patterns, their important characteristics and statistics.

4.1 Multimodal Integration Patterns

Although multimodal interaction offers a more natural communication paradigm in opposition to the traditional user interfaces (UIs), it introduces further research and development challenges, since it requires more detailed understanding of human interaction schemes. Empirical studies revealed that there are differences between users in an approach they integrate inputs during multimodal interaction. Fortunately, a number of distinctive patterns were observed and identified in users' integration behaviour allowing to classify users accordingly to their integration patterns. Importantly, the patterns were found very stable over time (97% consistency) and resistant to change even when strong reinforcement was intentionally delivered [OLC05, HPSM11]. These facts offer multimodal systems with possibilities to enhance their robustness and accuracy through adaptation to the specific characteristics of individual users. Two main and the most important identified patterns are temporal modality precedence and a temporal synchronization pattern.

4.1.1 Modality Precedence Pattern

In the perspective of modality precedence, users either use preferably one of the input modes in temporal precedence over others for the great majority of their multimodal interactions or stay neutral and combine the modes in a different order for different commands [OLC05, HZ17]. The former class of users can be denoted as modality dominant (e.g. *speech-dominant*, *gesture-dominant* etc.) and the latter as *neutral*.

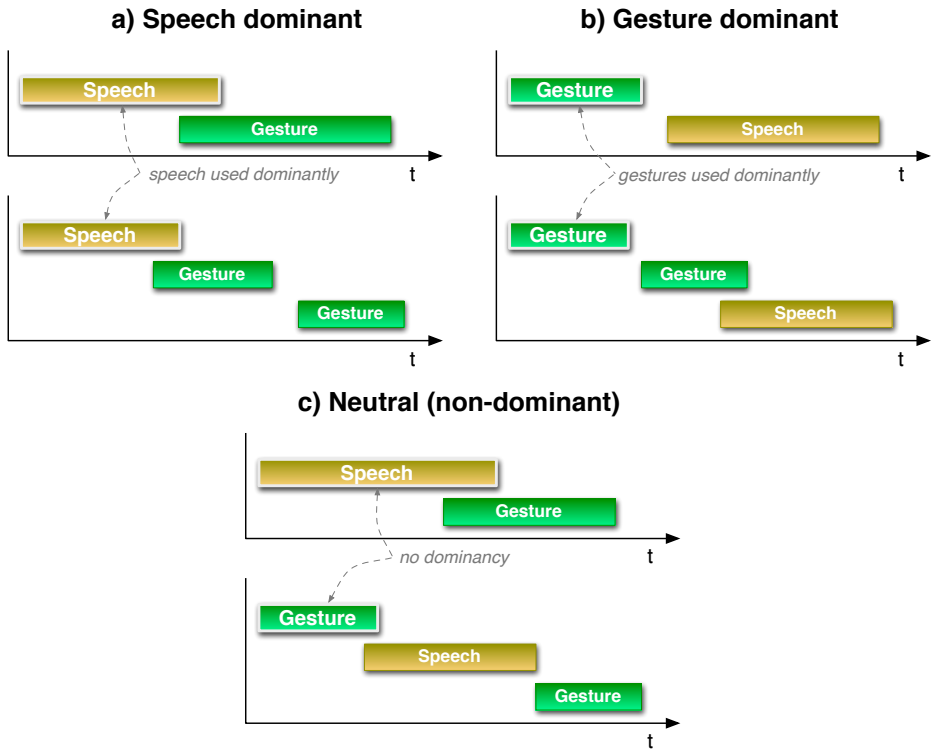


Figure 4.1: Illustration of the modality precedence patterns.

4.1.2 Temporal Synchronization Pattern

In terms of multimodal input synchronization, two primary integration patterns, SIM and SEQ, were discovered by Oviatt et al. [ODK97], and lately confirmed and subjected to further studies in [OCT+03, XO03, OLC05]. According to the research reports, SIM integrators carry multimodal input simultaneously (i.e. there is a temporal overlap between input signals or events), whereas SEQ integrators deliver input signals strictly sequentially (i.e. with a lag between signals). Subjects are considered SIM or SEQ dominant if at least 60% of their integrations follow the same pattern [OCL04]. Figure 4.2 illustrates examples of integrated inputs for each synchronization pattern.

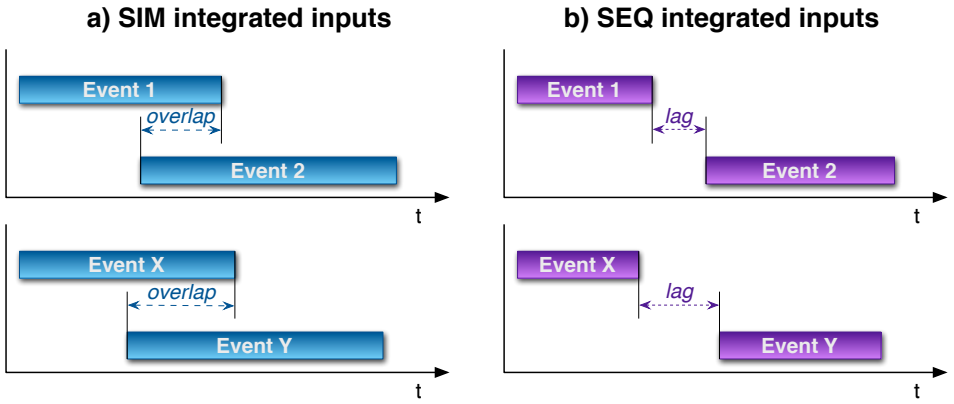


Figure 4.2: Illustration of the temporal synchronization patterns.

Although this classification was successfully utilized in other works, Schüssel et al. [SHS+14] recently reported difficulties regarding accuracy of the SIM/SEQ classification as defined by Oviatt et al. With respect to this report, an investigation of the reported inaccuracy will be included into the main goals of the following study.

4.2 User Study on Integration Patterns

4.2.1 Testbed (Testing System)

In the early stages of this study, we considered utilizing a WoZ simulation in our testing system. However, a functional prototype if feasible, is a better alternative to WoZ techniques as discussed earlier. Therefore, we decided to implement a fully featured multimodal system as a testbed for our testing purposes.

Multimodal interaction is notably popular and proved to offer performance advantages especially in the visual/spatial domains. We decided to follow this trend and developed a multimodal interactive system that combines speech and gesture inputs. The gesture input is provided through a computer mouse (instead of a pen) as we wanted to evaluate a multimodal system that is capable of running on a standard workstation without the need of specific hardware components. The testing system is implemented and built on top of the multimodal framework Manitou [HZ13], which provides a base architecture and other advanced multimodal functionalities. The core of the system consists of two input recognizers (i.e. for gestures and speech) and a

multimodal integration component (*fusion engine*).

$1 Recognizer from Wobbrock et al. [WWL07] offers a powerful yet simple to implement and computationally efficient recognition algorithm for single stroke gestures. Our implementation uses a slightly modified version of the algorithm that additionally allows to distinguish a gesture orientation[1].

Speech recognition capabilities were integrated using PocketSphinx [HDKC+06]. PocketSphinx is an open-source lightweight speech recognition engine with a real-time recognition performance developed as a part of the CMUSphinx project at Carnegie Mellon University. The US English generic acoustic and language models (also provided by CMUSphinx) were used in the running configuration of our speech system.

The fusion engine uses a finite-state multimodal integration method, as described in [JB00, JB05]. Additionally, a 4-edit machine [BJ08, BJ09] built only with in-grammar entries was integrated in order to enhance robustness of speech interpretation. The finite-state method does not address a solution for input segmentation. Hence, a simple threshold strategy was used for segmentation of multimodal constructions with the fixed threshold set to 4 seconds. The strategy was slightly enhanced with an ability to accept input segments eagerly (i.e. without waiting) when there are no other possible hypotheses available and the current input combination cannot be accompanied with any other input to form different or alternative interpretations.

4.2.2 Methods

Subjects

10 adult subjects aged 25 to 49 years ($M = 31.09$, $SD = 7.33$), three female and seven male, participated in the study. All were unpaid volunteers with varying degrees of computer experience, but no-one was a computer scientist or had any previous experience with multimodal interactive systems. Although none of the participants were native speakers of English, their command of the language ranged from very good to near native in both spoken and written forms. One subject was left-handed, 9 were right-handed.

[1] The original algorithm is rotation invariant.

Task

Participants were introduced to a multimodal map-based application (very similar to Speak4it [EJ12] described earlier) offering standard location services as searching for points of interest (e.g. restaurants, hotels, emergencies, etc.), providing a route planner and basic public transport information. A dominant part of application's user interface was a map view spread over the majority of the window (see Figure 4.3). Besides standard GUI elements, the application was controlled using speech commands, mouse gestures and their multimodal combinations.

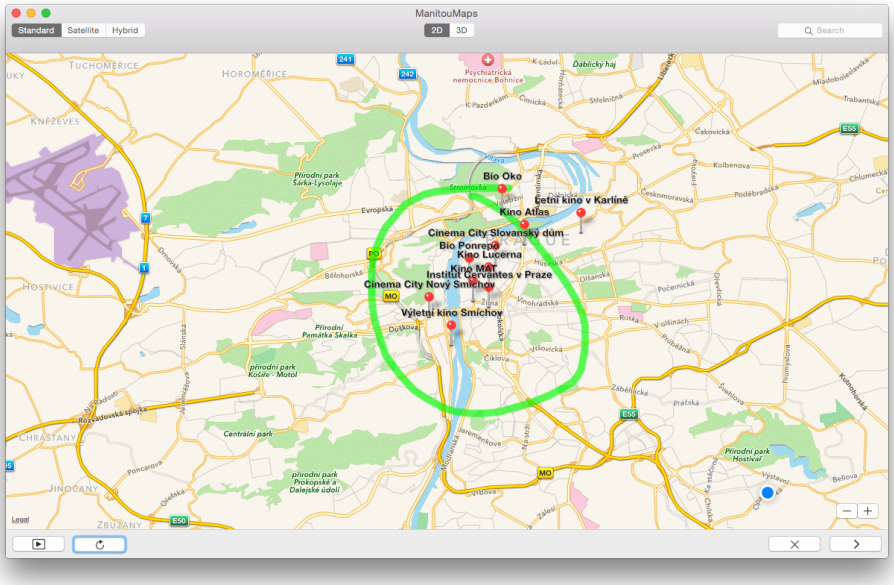


Figure 4.3: Map-based application user interface. The green path indicates a region gesture produced by a user in order to select an area of interest.

The map application supported 3 main types of gestures:

1. *points*,
2. rectangular or rounded *regions*, and
3. *symbols*.

The pointing gestures enabled users to define new or select existing points on the map. Specifying areas or selections of points was possible through the region gestures. The software also supported a set of 12 gesture symbols to execute predefined actions.

Difficulty	Goal	# of steps	Involved modalities
Low	Zoom out the map view	1	S (1x)
Moderate	Get information about a cinema in the view	2	S (2x) + G (2x)
Intermediate	Get directions between Prague and Ostrava	3	S (3x) + G (4x)
High	Find out phone numbers and postal addresses of libraries in the surrounding area of Czech Technical University in Prague	4	S (4x) + G (3x)
Very High	Find estimated travel time and distance between a theatre in the downtown of Prague and the closest airport	5	S (4x) + G (5x)

Table 4.1: Examples of the task difficulty. S and G in the last column denote speech and gesturing, resp., with an indication of a total number of distinct inputs conveyed through the given modality (in the brackets).

A concept of implicit contextual information (e.g. the last selected or found object, the current location, etc.) was intentionally not implemented into the application. This slight inconvenience obliged users to explicitly define all data needed to accomplish a task and resulted in significantly higher rates of multimodal interaction usage, but without forcing them to alter their interaction behavior. For instance, if a goal is to get a phone number of a restaurant in a specific area, a user has to find a restaurant in the area and then, in the second step, explicitly select the found restaurant using a pointing gesture and request its phone number using speech.

Difficulty of the test scenarios was divided into five groups:

- *low*,
- *moderate*,
- *intermediate*,
- *high* and
- *very high*.

Tasks with *low difficulty* required the user to provide only non-spatial information conveyed over a single modality and consisted of one or two steps. *Moderate difficulty* tasks involved two pieces of information (one spatial and one non-spatial) per action transferred over multiple modalities and comprised of no more than 3 steps to complete the goal. Tasks with *intermediate level*

of difficulty required three elements of input data (i.e. two elements with spatial/location information and one non-spatial) conveyed multimodally. *High* and *very high difficulty* scenarios were composed of a combination of low to intermediate tasks involving 3–5 and 4–6 steps, respectively, with the latter including more complex tasks. A list of sample scenario goals from each difficulty level is shown in Table 4.1.

Procedure

Moderated usability testing with quantitative objectives [Lew12] was arranged to assess all important measurements. All tests were done in a smaller multimedia lab specifically reorganized for the purposes of the study. The arrangement of the lab room is depicted in Figure 4.4. A moderator was present in the room throughout the whole session. In the initial phase, he was briefing subjects, introducing them to the system and its controls and providing feedback or help as needed. His role in the main session was changed to observing and taking notes about the test and its progress. In order to minimize impact of the moderator to tested subject's behavior, the communication was limited only to providing the necessary help when explicitly requested by the subject. All subjects were informed about the moderator's role during the main session in the initial phase and reminded once more right before the beginning of the session. No strict timeouts were given for task or session completion, i.e. the participants were given as much time as they needed.

Introduction and training phase. Subjects were first oriented to the lab by a moderator. For the purposes of the short training, five diverse scenarios were carefully chosen to help users become acquainted with the system and its overall control. The participants were instructed to repeat scenarios (or the selected subtasks) until they were fully oriented and ready to proceed to the main session. Critical in this phase was to emphasize that temporal order of modalities was not decisive and any combination should be correctly interpreted by the system. The subjects were asked to try changing the order of modalities and other characteristics of their input in order to identify the interaction style that they felt the most natural and comfortable. Typical training phase took 15–25 minutes.

Main Session. A total number of 16 isolated test scenarios were prepared for the main session. The scenarios were provided to the volunteers in textual form printed on paper. The first four scenarios were simple with

Figure 4.4: Lab room layout during the usability testing.

low to moderate difficulty and comprised of 2–3 steps to accomplish the task. The difficulty of the remaining scenarios was evenly distributed with a level fluctuating between moderate and very high. The participants were instructed to use exclusively the interaction style that they feel is the most natural to them and to progress through the scenarios in a given order and successfully complete every task at least three times before moving to the next. These reiterations were included in order to measure possible variations in the interaction characteristics observed while doing the same task repetitively. This session took the participants 45 minutes on average. During this session, only 3 subjects took the opportunity (in a total of 5 cases) to request feedback from the moderator. In all cases, subjects were verifying if they had already completed a sufficient number of repetitions of a given task.

Final Interview. Upon completion of the main session, volunteers were interviewed about their interactions and asked to subjectively evaluate the system performance. Afterwards, they were debriefed on the purpose of the study.

Durations of individual phases as well as total session duration for each participant are shown in Table 4.2.

Subject	Intro & Training	Main Session	Debrief	Total
1	18′	42′	10′	70′
2	15′	41′	8′	64′
3	20′	42′	7′	69′
4	19′	46′	11′	76′
5	16′	45′	8′	69′
6	17′	50′	9′	76′
7	25′	44′	9′	78′
8	21′	44′	9′	74′
9	19′	49′	7′	75′
10	22′	51′	10′	83′
Average	19′	45′	9′	73′

Table 4.2: Total session durations and durations of particular phases for individual subjects measured in minutes.

Data Recording and Annotation

Two video recordings were captured from all sessions:

1. *screen recording* – capturing an application UI and audio input
2. *camera recording* – documenting participants as they interact with the test application.

Apart from the video capture, all important events, hypotheses provided by underlying recognizers, final interpretations and other data were timestamped and recorded by a data-logger.

At the beginning of the recording a high-pitch sound was played using the application itself in order to provide a sync reference. This reference was then used in the post-processing phase to synchronize both recordings and data from the logger.

Before analysis, the captured data went through specific post-processing procedures. One of the essential objectives was to annotate the data. Since most of the events were already logged and properly timestamped, it was only needed to mark errors and divide them into the following groups:

1. *user errors* – errors caused by a user (e.g. a subject forgets to select an object using a pointing gesture while asking for its address),
2. *recognition errors* – low level errors produced by underlying recognizers (e.g. when a speech recognizer provides an incorrect interpretation of an utterance),
3. *interpretation errors* – high level errors experienced when the system fails to interpret a correct meaning even if all underlying components (e.g. recognizers, etc.) provided appropriate recognition results.

Another task was to perform revision and adjustments of speech utterance timestamps since the temporal information provided by the underlying speech recognizer were inaccurate and limited in precision. The temporal precision of the PocketSphinx engine is limited to .01 second. Beyond that limitation, it showed significant inaccuracies in determination of temporal aspects of speech activity.

According to the required post-processing a proper equipment was necessary. Although there are powerful and feature-rich annotation tools available for researchers (e.g. ANVIL[2]), we decided to implement our own advanced annotation tool due to specific needs and internal data structure of logged data. Its UI consists of several temporally synchronized views (see Figure 4.5) — two views for the recorded video data (the upper part of the window) and a timeline view (in the bottom part) showing all important events and visualisation of an audio stream waveform. The tool offers standard playback controls to review the session or selected parts, allows the annotation of individual events on the timeline and adjustment of their timestamps both directly by dragging the event over the timeline and indirectly by editing its value using standard input fields.

4.3 Results of User Study

A total of 1417 individual constructions were expressed by participants during the study. Of them, 978, or 69%, were correctly interpreted by the system. User errors accounted for 83 (or 6%) of uninterpreted expressions and the remaining 355 (or 25%) were recognition errors. Most of the recognition errors (98%) were caused by the speech recognizer. Total accuracy of the system (after excluding user errors) was 73.3%.

[2] http://www.anvil-software.org

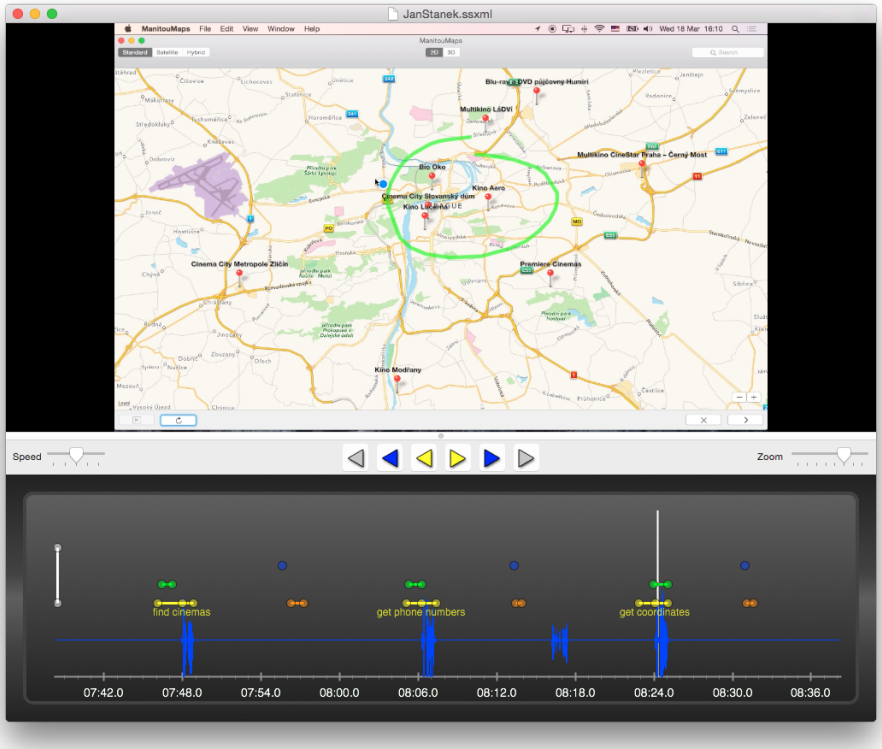


Figure 4.5: Annotation tool user interface. The upper part contains a playback view with a captured screen recording. The bottom part of the window shows a timeline view, which is synchronized with the playback view.

Of 978 correctly interpreted commands, 729 (or 75%) were expressed multimodally and 249 (or 25%) unimodally. The higher frequency of multimodal commands corresponds to the composition of tested scenarios, which were intentionally focused more on multimodal tasks.

4.3.1 Speech

Linguistic analysis of spoken utterances revealed that all participants used syntactically simpler command-style language constructions or switched to this style very early during the training phase and remained consistent throughout the whole session. For example, a user was searching for restaurants in a designated area by selecting the location using a region gesture and speaking: *"Find restaurants in the selected area."* However, when repeating the same task later the user completed the action using the following utterance: *"Find restaurants,"* and kept the same simplified phrase from then on. Similar observations about linguistic differences associated with multimodal interaction

were detected and examined in previous studies [ODK97, Ovi99b].

Total accuracy of the speech subsystem was 73.9%. There were numerous causes of the limited success rate. Probably the major cause was small deviations in pronunciation and intonation introduced by subjects, since all of them were non-native speakers of English. This could be resolved in future by adaptation of the used acoustic model. However, the results of the study were not influenced, as only valid and correct interpretations were taken into account for further analyses.

A paired t-test confirmed (normality verified using the Shapiro-Wilk test) that speech duration did not change significantly between the first and second half of the main session, $t(91) = 1.45$, $p = .075$. At the same time no increase or decrease in duration was observed from the first attempt to the deeper repeats, a paired t-test, $t(87) < 1$, ns.

4.3.2 Gestures

Symbol gestures were used rather rarely by participants, since they were assigned exclusively to auxiliary actions. Thus, without the context of another modality, only regions were interesting to analyze in more depth.

Although the testing application supported two forms of region gestures (*rounded* and *rectangular*), subjects exclusively used the rounded variant. Detailed investigation of gesture shapes showed interesting similarities and paradigms in gesture style of individual participants. All 365 (or 100%) produced region gestures had a counter-clockwise direction. Relative locations of the initial points as well as of the last points of the region gestures displayed only fractional differences in the subjects.

Three main shape feature patterns were observed:

1. *Opened* – an open-loop circle (or ellipse); an angle distance between the beginning and end of a gesture path is >20° in the opposite direction to the path.
2. *Closed* – a closed-loop ellipse without significant overlap; an angle distance is within <-20°, +20°>.
3. *Overlapping* – a closed-loop ellipse with an overlap; an angle distance is >20° in the same direction as a gesture path.

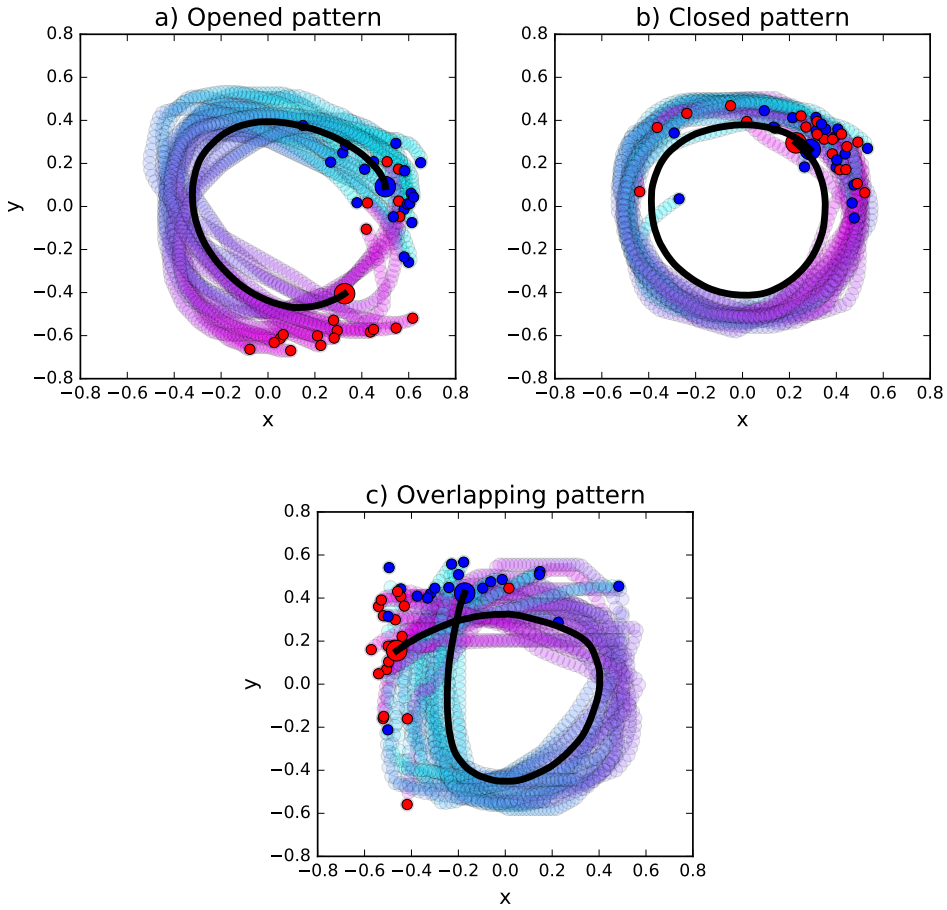


Figure 4.6: Visualization of region gesture shape patterns.

Sample gestures from one subject with opened, closed and overlapping pattern are visualized in Figure 4.6a), 4.6b) and 4.6c), respectively. The gestures were aligned and appropriately scaled for better legibility. The blue dots indicate the heads (initial points) and the tails (last points) are marked with red. An average gesture path is plotted with a black line.

Average angles of heads and tails, sample means and medians of their angle distances, and percentage of closed gestures for each participant is shown in Table 4.3.

More than half of the participants, or 7, had a closed-loop pattern, 2 participants displayed an overlapping pattern and 1 subject used exclusively open-loop regions.

Subject	Average angle		Angle distance		Closed
	Head	Tail	Mean	Median	
1	121.3°	164.1°	**42.8°**	**37.1°**	96.8%
2	59.4°	88.1°	**28.7°**	**29.5°**	83.3%
3	78.3°	89.5°	19.3°	15.4°	65.9%
4	88.2°	104.9°	16.7°	13.4°	88.5%
5	104.7°	115.5°	10.9°	7.1°	88.2%
6	52.5°	46.1°	11.7°	6.6°	77.5%
7	84.2°	95.6°	11.5°	6.3°	72.7%
8	67.8°	80.2°	12.4°	5.6°	61.5%
9	50.9°	60.1°	9.2°	5.6°	90.5%
10	88.7°	270.9°	**-95.3°**	**-104.6°**	**0.0%**

Table 4.3: Analysis of region gestures.

4.3.3 Multimodal Commands

To analyze common properties and identify main patterns and other relationships, multimodal constructions were divided into the following groups by content involved in a gesture modality:

1. *single region*,
2. *single point*, and
3. *two points*.

Figure 4.7 reveals three main groups and the percentage of their appearance in the data set.

Multimodal Command Duration

Multimodal command duration was measured for each command as duration from the start of the first signal to the end of the final signal. Command duration of multimodal constructions involving a region gesture did not change significantly from the first to the second half of the session, a paired t-test, $t(108) < 1$, ns, nor did it change between the original and deeper repeats (i.e. between the first and other successive repetitions of the same task), $t(92) = 1.21$, $p = .11$.

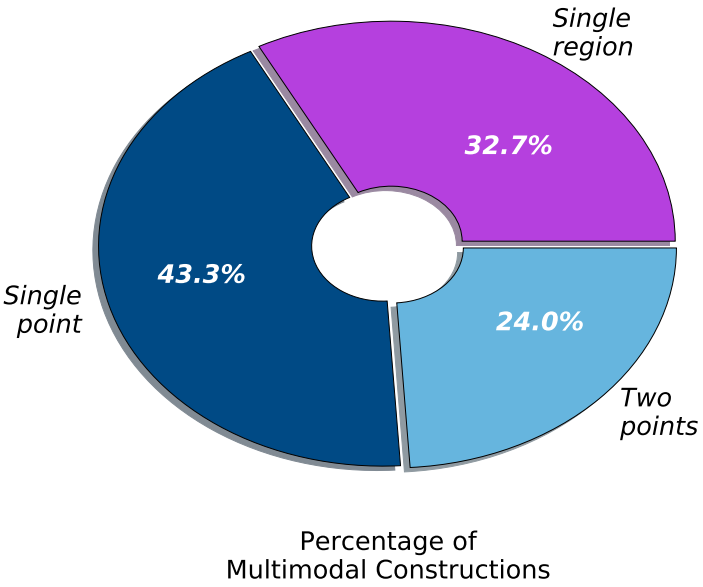


Figure 4.7: Percentage of multimodal constructions represented by different type of gesture-based content.

Command duration of multimodal constructions containing a point specification in a gesture modality likewise did not change significantly by a paired t-test between the first and second half of the session, $t(131) = 1.32$, $p = .09$. During each testing scenario, no significant change was found between the original input and deeper repeats, a paired t-test, $t(107) = 1.08$, $p = .14$.

Similarly, there was no evidence of change in multimodal command duration incorporating speech and two pointing gestures between the first and second half, $t(64) < 1$, ns, as well as from the original to deeper repeats, $t(47) < 1$, ns, both confirmed by a paired t-test. The assumption of normality was confirmed by the Shapiro-Wilk test for all tested duration differences.

Modality Precedence

A temporal analysis of all multimodal expressions across all three aforementioned groups revealed that 6 out of 10 participants used speech as their dominant modality (i.e. 75% or more of their interactions started with speech). Two subjects preferably delivered gestures in temporal precedence over speech, and the remaining 2 subjects showed no significant modality precedence, or were non-dominant.

Subject	1st half	2nd half	Complete session	Precedence
1	96.3%	**100.0%**	98.5%	*Speech*
2	96.3%	**97.5%**	97.0%	
3	88.0%	**100.0%**	95.0%	
4	86.1%	**97.6%**	92.2%	
5	73.3%	**89.8%**	83.5%	
6	79.5%	**81.8%**	80.7%	
7	64.5%	72.5%	71.8%	*Non-dominant*
8	*81.5%*	45.5%	59.2%	
9	*60.5%*	**18.4%**	42.0%	*Gesture*
10	*50.0%*	**10.8%**	29.6%	

Table 4.4: Percentage of speech precedence versus gesture precedence in multimodal constructions.

A percentage of modality delivery in the first and second half of the session, as well as over the complete session for each participant is presented in Table 4.4.

Percentage of dominant modality typically slightly increased between the first and second half underlining and entrenching the participants' dominant scheme. More pronounced pattern changes were observed for 3 subjects (rows 8, 9, 10 in Table 4.4). A detailed examination unveiled the appearance of a pattern switch in the first half of the main session indicating the training phase was not long enough for these individuals to build up and stabilize their dominant delivery patterns.

Temporal Synchronization

Table 4.5 presents an evaluation of the SIM/SEQ pattern, as suggested by Oviatt et al. [OCL04], performed on our collected data set. In accordance with the SIM/SEQ pattern, all 10 participants were classified as SIM integrators and no participant was either a SEQ integrator or non-dominant with consistency of 88.2% when focusing solely on multimodal constructions of speech and a region gesture (see Speech+Region column in Table 4.5). However, compared to other combinations, the obtained results vary substantially leading to the conclusion that the classification is inconsistent across the different multimodal constructions (i.e. *speech+region*, *speech+single point* and *speech+2 points*).

Subject	SIM / SEQ ratio		
	Speech+Region	Speech+Point	Speech+2 Points
1	100.0%	62.5%	71.4%
2	100.0%	6.3%	100.0%
3	100.0%	45.8%	33.3%
4	100.0%	26.9%	100.0%
5	96.9%	37.5%	44.4%
6	90.0%	50.0%	50.0%
7	85.7%	14.8%	33.3%
8	81.0%	53.3%	85.7%
9	66.7%	8.7%	16.7%
10	61.5%	5.3%	33.3%

Table 4.5: Percentage of SIM-integrated versus SEQ-integrated commands represented by a type of multimodal construction.

These findings are contrary to the previous discoveries and research results introduced by Oviatt et al. [OCT+03, OCL04, ODK97] and Xiao et al. [XGO02, XO03]. Schüssel et al. [SHS+14] recently also reported difficulties regarding the SIM/SEQ classification. They concluded the SIM/SEQ pattern is not distinctive enough (at least for data in their experiment) and decided to avoid the classification. Instead, they introduced different metrics as a replacement.

Our detailed analysis revealed numerous cases where subjects tending to the SEQ integration pattern started with their subsequent modality by the end of the previous input signal with short overlap and not necessarily after it with noticeable lag, as suggested in the related literature. Similarly, users tending to the SIM pattern started in some of their interactions with the subsequent modality after the end of the previous signal without any overlap – typically when the duration of the preceding modality signal was very short (e.g. a pointing gesture).

These trends are visible in the temporal histograms below, where data from the same six participants are presented in both diagrams. Two subjects from each group according to their dominant modality were selected (subj. 1–2 — non-dominant, 3–4 — speech-dominant and 5–6 — gesture-dominant). Histograms in Figure 4.8 represent an onset difference between speech and gesture signals. The mean value of the difference is located around 0 seconds for the first four subjects (1–4), whereas there is a considerable disparity for subjects 5 and 6. Figure 4.9 shows histograms with differences of intermodal lag / overlap (i.e. between the end of the preceding signal and the onset of the following one). In contrast with the previous, subjects 1-4 delivered their multimodal inputs with significant signal overlap, while the other two subjects (5–6) conveyed their inputs with a short lag or overlap between the signals with a mean value located around 0 s.

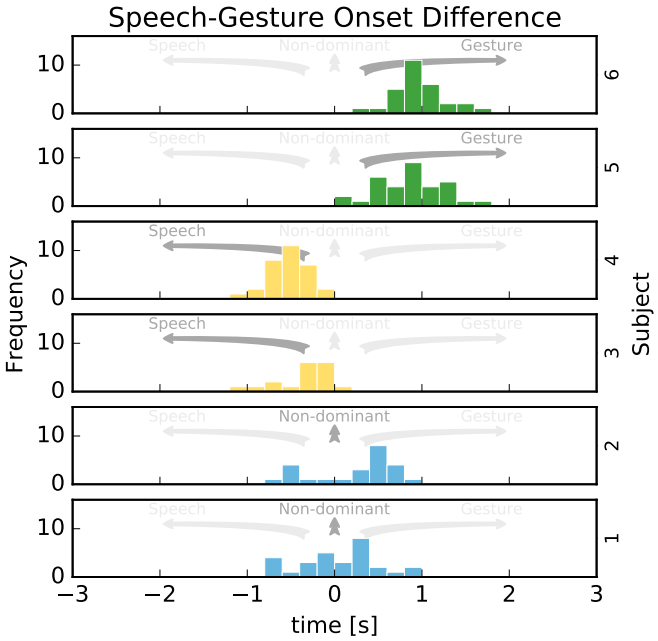


Figure 4.8: Histogram of temporal onset differences between speech and gesture signal for selected subjects. The blue color denotes non-dominant, yellow speech-dominant and green gesture-dominant subjects.

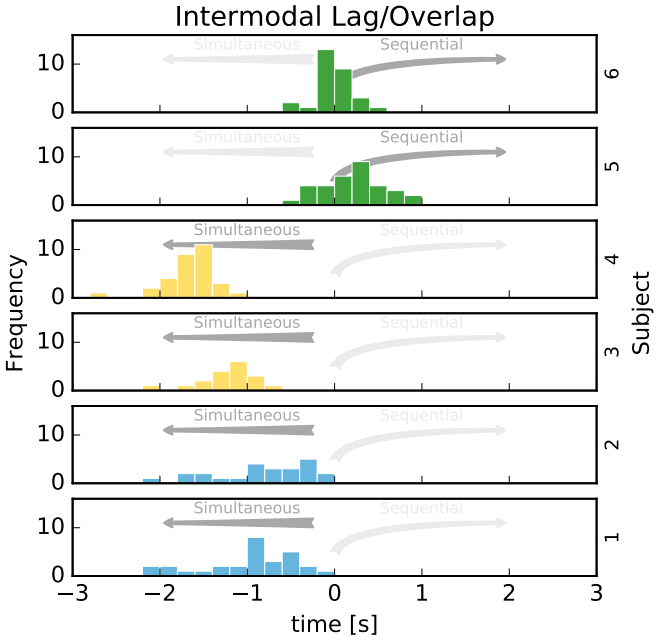


Figure 4.9: Histogram of temporal signal differences between the first modality signal offset and the following signal onset. The blue color denotes non-dominant, yellow speech-dominant and green gesture-dominant subjects.

In order to sufficiently describe subjects' multimodal synchronization pattern that is consistent over the multimodal constructions, on the one hand, and fits to all subjects on the other hand, we redefined conditions of the SIM/SEQ classification. In our new definition of simultaneous and sequential synchronization patterns, a multimodal construction is considered:

1. ***Sequential*** (SEQ_R)[3] – if the onset-distance between two signals is greater than or equal to their overlap and the onset-distance is greater than a defined minimal threshold.
2. ***Simultaneous*** (SIM_R) – if the onset-distance of two signals is less then their overlap or if the onset-distance is lower than or equal to the minimum.

The following equation describes a rule to determine if a multimodal input is classified as SIM_R or SEQ_R:

$$\begin{array}{lll} \Delta t_{on} \geq \Delta t_{ovr} \wedge \Delta t_{on} > \Delta t_{min} & \Longrightarrow & \text{input} \in SEQ_R \\ \Delta t_{on} < \Delta t_{ovr} \vee \Delta t_{on} \leq \Delta t_{min} & \Longrightarrow & \text{input} \in SIM_R \end{array} \tag{4.1}$$

Δt_{on} denotes the onset-distance and Δt_{ovr} the overlap between the signals. They are computed as follows:

$$\begin{aligned} \Delta t_{on} &= t_{s_2} - t_{s_1}, \\ \Delta t_{ovr} &= t_{e_1} - t_{s_2}, \end{aligned} \tag{4.2}$$

where t_{s_1} and t_{e_1} denote time of the beginning and the end of the first modality signal, respectively, and t_{s_2} is time of the beginning of the following modality. Demonstrative illustrations are depicted in Figure 4.10.

Δt_{min} in (4.1) is a minimal defined threshold for the onset-distance. If the onset of two inputs lies inside the threshold they are considered simultaneous even if they do not overlap. This condition addresses those situations where the signal of preceding modality is very short (e.g. pointing gestures) and thus the comparison between onset and overlap is not appropriate. Based on our dataset, we found a value of 750 ms as optimal for a gesture-speech multimodal combination.

[3] In order to distinguish between the two definitions, the one from Oviatt et al. will be denoted as SEQ_O/SIM_O and our redefined as SEQ_R/SIM_R in the rest of the work.

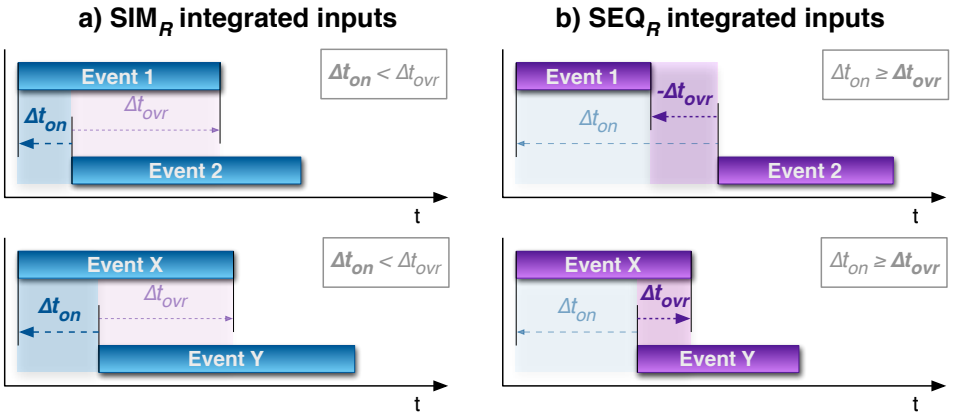


Figure 4.10: Illustration of the redefined temporal synchronization pattern classification.

Table 4.6 provides an evaluation of the SIM/SEQ pattern for individual subjects using the redefined classification. Comparison of consistencies of individual multimodal groups as well as average consistency evaluated for the original SIM_O/SEQ_O and redefined SIM_R/SEQ_R classifications is shown in Figure 4.11. The redefined classification displays strong consistency in all types of multimodal combinations. It offers average consistency of 95.5%, compared to only 79.1% of the original classification, a relative improvement of 20.7%. Advancement in consistency is especially considerable for combinations with a single point and two points in a gesture modality, where the relative improvement is 29.7% and 25.8%, respectively.

Subject	SIM_R / SEQ_R ratio		
	Speech+Region	Speech+Point	Speech+2 Points
1	100.0%	97.0%	90.9%
2	100.0%	100.0%	100.0%
3	*0.0%*	*4.0%*	*11.1%*
4	100.0%	96.3%	92.3%
5	96.9%	90.0%	88.9%
6	94.1%	93.3%	85.7%
7	*0.0%*	*3.6%*	*0.0%*
8	95.2%	100.0%	88.9%
9	*5.0%*	*8.3%*	*0.0%*
10	*0.0%*	*0.0%*	*9.1%*

Table 4.6: Percentage of SIM-integrated versus SEQ-integrated commands represented by a type of multimodal construction using the redefined classification.

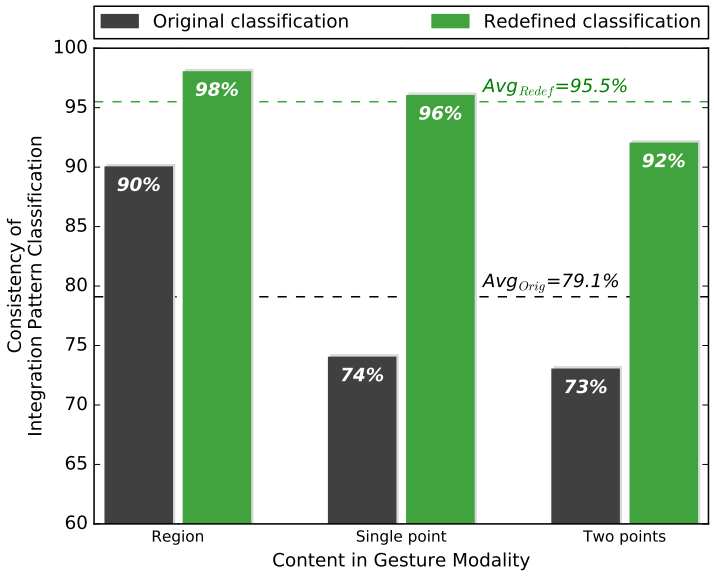


Figure 4.11: Consistency of the original $\text{SIM}_O/\text{SEQ}_O$ and redefined $\text{SIM}_R/\text{SEQ}_R$ multimodal pattern classification.

4.3.4 New Categorization Combining $\text{SIM}_R/\text{SEQ}_R$ Pattern and Dominant Modality

In order to provide a single classification of users' integration patterns, which describes the important individual differences more coherently, a new and more detailed categorization is introduced. To this end, we combined the two most distinctive features in a single categorization, i.e. the redefined $\text{SIM}_R/\text{SEQ}_R$ classification with dominant modality precedence. In our case, a combination of speech and gestures, it forms a total of 4 basic categories:

1. *SEQ/Speech*
2. *SEQ/Gesture*
3. *SIM/Speech*
4. *SIM/Gesture*

Apart from these basic combinations, there are two additional categories:

5. *SIM/Neutral*
6. *SEQ/Neutral*

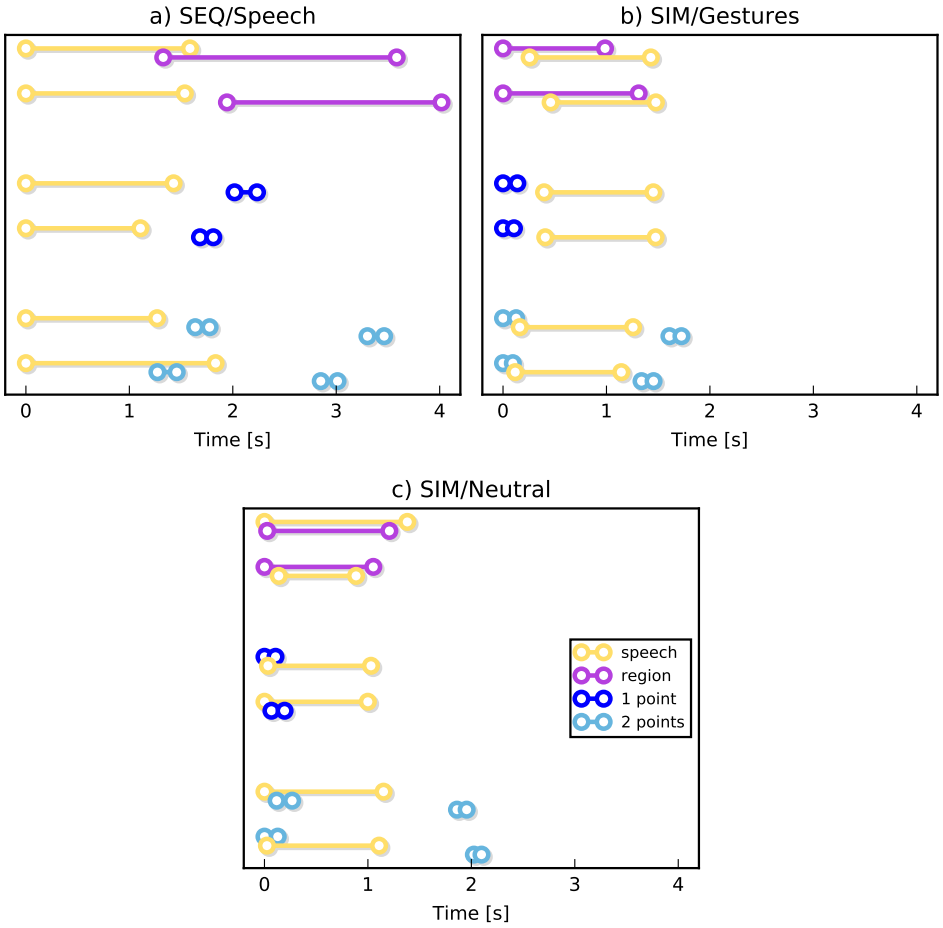


Figure 4.12: Visualization of temporal delivery of input signals for selected multimodal pattern classifications.

These two additional categories occur in subjects with a non-dominant modality (see subjects 1 and 2 in Figure 4.8), although we presume that presence of the SEQ/Neutral category is very uncommon.

Examples of event sequences representing three integration pattern categorizations using real samples selected from our dataset obtained during the study are depicted in Figure 4.12.

Evaluation of multimodal synchronization patterns with respect to the new categorization showed that 4 participants exhibited the SEQ/Speech pattern, 2 were SIM/Gesture integrators and remaining 4 subjects were classified as SIM/Neutral integrators. Table 4.7 provides average differences in onsets (for SIM_R integrators) and lags/overlaps (SEQ_R integrators) between constructions containing a single region and a single pointing gesture.

Pattern	# of subjects	Average difference Region [s]	Point [s]	Relative change [s]	[%]
SIMULTANEOUS		onset			
SIM/Gesture	2	0.51	0.47	*.04*	*-7% (ns)*
SIM/Neutral	4	0.35	0.42	*.07*	*20% (ns)*
SEQUENTIAL		lag(+)/overlap(–)			
SEQ/Speech	4	-0.09	0.35	**.44**	**-506%**

Table 4.7: Average difference in modality signals between commands with a region and a pointing gesture.

In the case of SIM/Gesture integrators, a paired t-test (this and all of the following tested onset and lag/overlap differences are normally distributed according to the Shapiro-Wilk test) confirmed that average signal onset between modalities did not change when comparing multimodal constructions with a region with those containing a single pointing gesture, $t(18) < 1$, ns.

Likewise, there was no significant change in average onset between input signals from both multimodal command groups produced by participants with a SIM/Neutral integration pattern, confirmed by a paired t-test, $t(35) < 1$, ns.

Average signal lag/overlap changed for SEQ/Speech integrators from -0.09 (overlap) to 0.35 seconds (lag), or by 0.44 seconds (-506%), between the constructions with a region and a single pointing gesture, significant by a paired t-test, $t(31) = 5.7$, $p < .0005$, one-tailed.

Further investigation and video analysis provided insights into this result, which is related to some specifics of mouse input. In our data set, a measurement/log of gesture signal was initiated by a mouse press and finished with its release. However, the analysis revealed that input was initiated little sooner with movement of a mouse cursor to an intended position. Since pointing gestures typically require greater precision and accuracy (and thus time) compared to region gestures, there is alleged increase in the signal lag.

To complete a multimodal integration model, distributions of the temporal differences should be derived separately for every user. Since different temporal aspects are relevant for SIM_R and SEQ_R integrators, we propose two independent models. One that captures intermodal onset difference of the signals for SIM_R subjects and another one that captures intermodal lag/overlap for subjects with the SEQ_R pattern. Figure 4.13 provides sample

distributions for 4 subjects from our dataset (subjects 1-2 are SIM/Neutral and 3-4 are SIM/Gesture integrators) and figure 4.14 shows distributions of intermodal lag/overlap for 2 subjects representing the SEQ/Speech category.

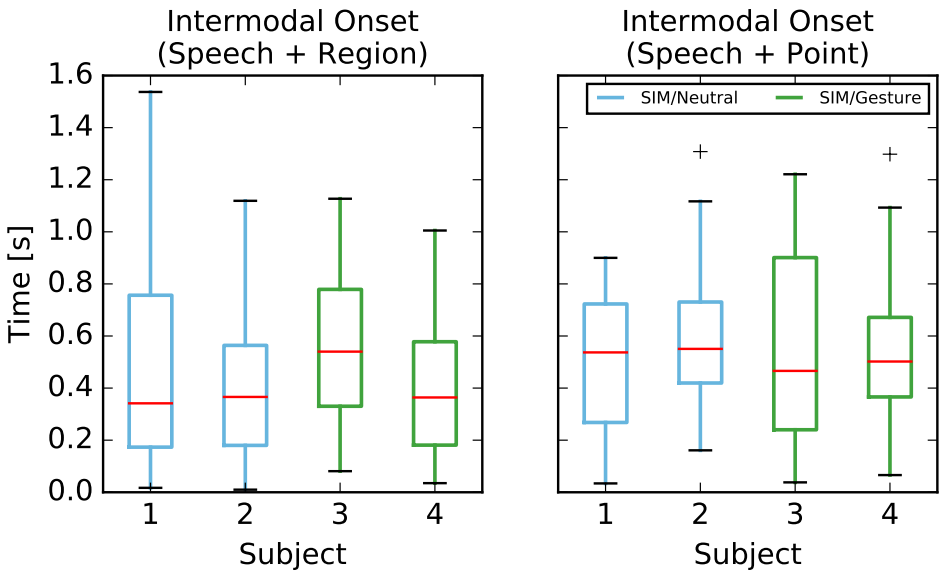


Figure 4.13: Distribution of intermodal onset for subjects with SIM/Neutral and SIM/Gesture integration categorization.

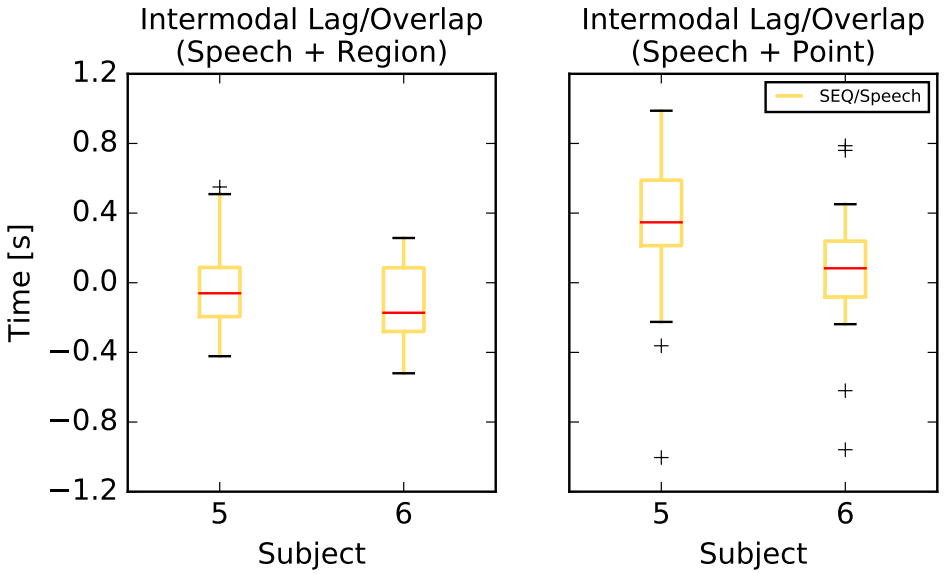


Figure 4.14: Distribution of intermodal lag/overlap for subjects with SEQ/Speech integration categorization.

4.4 Discussion

As expected, duration of a speech signal did not exhibit any significant changes either in the course of a session or during deeper repeats. Total multimodal command duration was consistent and stable throughout the session for each individual subject. Based on previous reports, it is expected to be very stable over longer periods of time and resistant to changes even when strongly enforced [OCT+03].

Three anticipated patterns were confirmed in the perspective of modality precedence. Participants were classified as speech-dominant in 6 cases (60%), neutral (or non-dominant) in 2 cases (20%), and two subjects, or 20%, were gesture-dominant.

The classification of subjects' multimodal integration pattern to *simultaneous* (SIM_O) and *sequential* (SEQ_O) strictly based on the existence of an intermodal lag/overlap, introduced in the previous literature, did not offer satisfying measure, since it was strongly inconsistent over the different multimodal constructions. A detailed analysis revealed that some *sequential* integrators tend to start with a subsequent modality by the end of the previous one with a small overlap between the signals. As a consequence, many of the interactions from those subjects were inaccurately evaluated as *simultaneous*. Based on our findings, we provided redefined conditions and introduced the improved SIM_R/SEQ_R classification leading to a significant improvement of average consistency from 79.1% to 95.5% (a relative improvement of 20.7%).

In order to provide more coherent classification of multimodal integration patterns we combined the redefined SIM_R/SEQ_R classification with dominant modality precedence and introduced a new compound integration pattern categorization.

According to the newly introduced categorization of multimodal integration patterns, 2 participants of our study were classified as SIM/Gesture integrators, 4 exhibited the SIM/Neutral integration pattern and the other 4 were SEQ/Speech integrators. We did not observe any subject with SIM/Speech, SEQ/Gesture or SEQ/Neutral patterns. In our presumption, the last mentioned pattern, SEQ/Neutral, would be very rare as the majority of subjects with the non-dominant modality precedence are expected to be SIM integrators (i.e. belonging to the SIM/Neutral category). However, it would be interesting to evaluate a distribution of the categories in the population.

Based on previous research and empirical studies, we expect that our new integration pattern categorization would be very stable and consistent over time [OLC05] and predictable very shortly from the beginning of interaction. According to [HO06], it is possible to classify users' dominant SIM_O/SEQ_O pattern using the first 15 commands with 100% prediction accuracy. Nevertheless, user's dominant delivery patterns should be already fully developed and stabilized (i.e. the presumption of short predictability should not be applied on users that have no previous experience with similar multimodal systems). Robustness of multimodal interactive systems should increase significantly with employing adaptation techniques to these patterns resulting in a boost in users' experience and overall usability of the system.

Precise distributions and values of onsets (for SIM_R integrators) and lags/overlaps (for SEQ_R integrators) should be derived for every user separately. The values could also vary for different combinations of multimodal inputs even for one user, however, the main SIM_R/SEQ_R scheme and dominant modality should remain the same.

According to our findings, the subjects with the SIM_R integration pattern exhibit different temporal aspects in comparison with SEQ_R integrators, and more interestingly, some of the characteristics are contradicting (e.g. intermodal lag/overlap versus onset) resulting in the possible degradation of the single-model solution. An approach with separate models for both SIM_R and SEQ_R integrators should offer superior results to the single-model concept suggested by Schüssel et al. For instance, we suppose that if the authors would employ two different sets of metrics in their Interaction history, one for each type of the integrators, the detection rate of false positives would increase significantly (the reported detection accuracy was only 11.1%). We would suggest *temporal gap*, *total duration* and *center distance* as the metrics for the SEQ_R integrators, and *onset distance*, *total duration* and *center distance* for the SIM_R integrators.

Chapter 5

Integration Patterns Modeling and Input Prediction

The previous analysis and user study brought the important results and new interesting findings about users' multimodal integration patterns. In this part, we will capitalize on these results and design a user model capable of adaptation to the specific integration patterns of individual users.

5.1 User Model for Input Prediction

Huang et al. [HOL06] previously used a naive variant of Bayesian Belief Network (BBN) model (a probabilistic graphical model) to predict the multimodal temporal synchronization pattern, command type and type of the next signal. These properties were modelled as output variables of BBN and all were induced from four input variables – a current signal type, current signal duration, the last integration pattern and the last command type (see Figure 5.1). The authors reported the prediction accuracy of 88% and 91% for the command type and synchronization pattern, resp., using a standard leave-one-out test.

As the reported prediction capabilities were reasonably high, we decided to use their model as a starting point and a reference. However, our recent findings (see Section 4.2) suggest other more promising characteristics of integration patterns that should provide more consistent and precise modeling

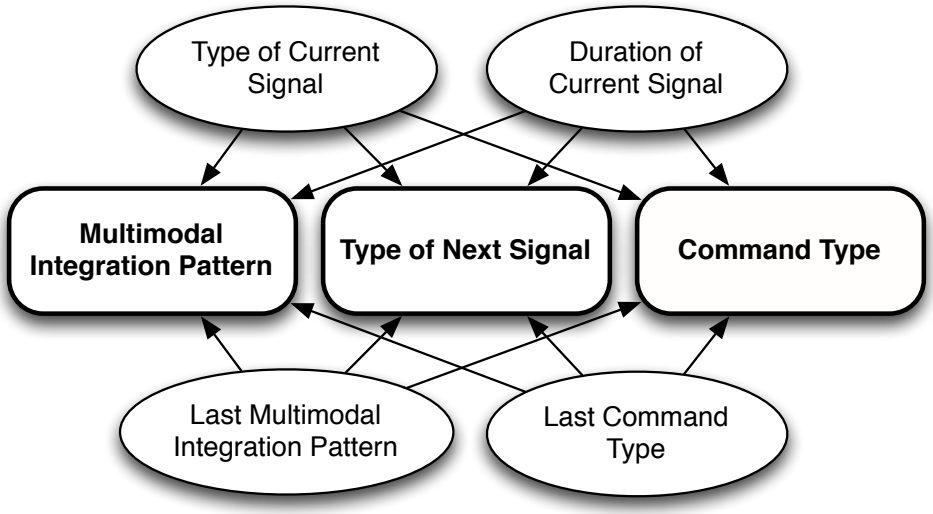


Figure 5.1: BBN model as proposed by Huang et al. (source: [HOL06])

capabilities. We took those findings into account and rebuild the model from the ground up. The designed model consists of 5 input discrete variables that all contribute to classification of 2 output variables. Their description follows:

- **MM Integration Pattern Category** (Input) – specifies user's multimodal integration pattern category as defined earlier in Section 4.3.4 (e.g. SEQ/Speech, SIM/Neutral etc.).
- **Onset Difference** (Input) – defines a temporal onset difference between the previous and the current signal. Encoding to discrete values is discussed later in this section.
- **Overlap Difference** (Input) – defines a temporal overlap between the previous and the current input signal. Encoding to discrete representations is discussed later.
- **Type of Previous Signal** (Input) – specifies a type of the previous signal (e.g. speech, etc.). If the current signal is the first in a command, *SILENCE* is used as a value.
- **Type of Current Signal** (Input) – specifies a type of the current signal (e.g. speech, gesture, etc.).
- **Command Type** (Output) – indicates a type of the current command (i.e. unimodal or multimodal).
- **Type of Next Signal** (Output) – indicates a type of the next signal (e.g. speech, gesture, etc.). If the current input signal is the last in the sequence and no other signal is expected in a command, *SILENCE* is used as a value.

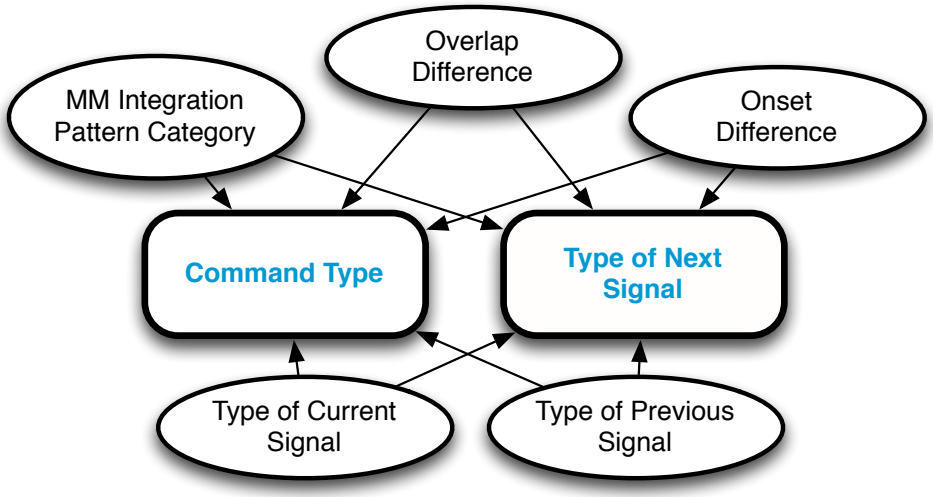


Figure 5.2: Graphical representation of our BBN prediction model

A graphical representation of our redesigned BBN classifier is depicted in Figure 5.2.

The modality dominance pattern revealed in users is represented in the model through the type of the previous and the current signal variables. The onset and overlap differences, on the other side, should reflect the $\text{SIM}_R/\text{SEQ}_R$ synchronization pattern. Finally, the multimodal integration pattern category variable allows the model to distinguish between different classes of users.

5.1.1 Variable Discretization & Optimal Training Sample Size

The model contains two continuous variables that represent time differences (the onset and overlap) between consecutive signals. Both of these variables need to be discretized in order to be used in a Naive Bayes classifier. We handled this by computing the mean μ and standard deviation σ of each variable from a training set, then calculate the difference x between the mean and the observed value v as $x = |\mu - v|$, and finaly divide results into bins. According to the 68–95–99.7 rule (see the upper part of Figure 5.3), 99.73% of the values lie within 3σ of the mean in a normal distribution. Hence, the effective division would be to separate results into two groups, one for $x \leq 3\sigma$ and other for $x > 3\sigma$. The assumption of normality comes from the results of Shapiro-Wilk tests performed during the evaluation of the user study (see Section 4.3).

For the purposes of experimentation, additional division strategies were also introduced. A total of 4 different approaches were suggested for the purpose of binning:

- $x \leq 3\sigma, x > 3\sigma$
- $x \leq 2\sigma, x > 2\sigma$
- $x \leq \sigma, x > \sigma$
- $x \leq \sigma, x \leq 2\sigma, x \leq 3\sigma, x > 3\sigma$

The first three variants differ only in a multiplier of σ. The last one introduces a more fine-grained (FG) division. An illustration of the intervals for all four proposed division strategies is depicted in Figure 5.3 (bellow the demonstration of the 68–95–99.7 rule).

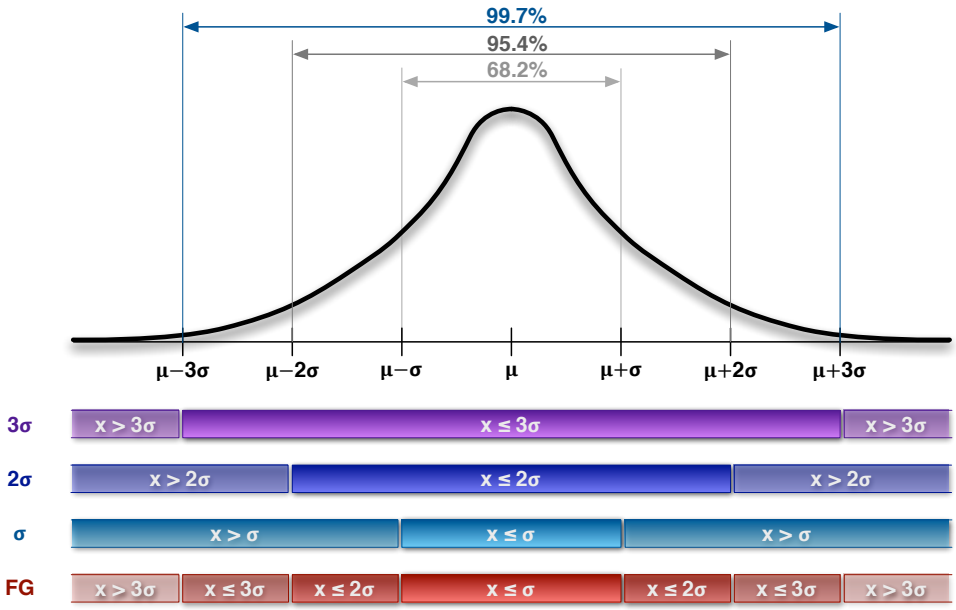


Figure 5.3: Illustration of the 68–95–99.7 rule (in the upper part). The lower part demonstrates the binning intervals for all 4 different division strategies.

An experiment was conducted to determine an ideal binning method and also to examine an optimal training sample size. The model was trained and evaulated separately on interaction data of each subject with the first N samples used as a training set and the last 35 samples as an evaluation set. Afterwards, the obtained results were averaged. The predicted variable was the next signal type. The results of the experiment is depicted in Fig. 5.4.

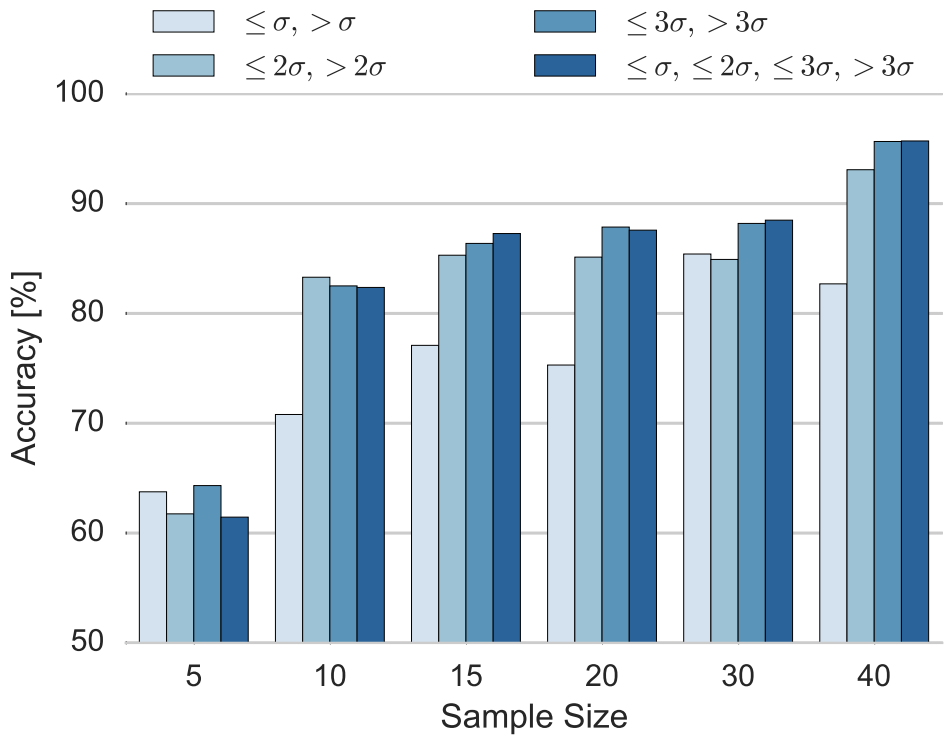


Figure 5.4: Effect of a sample size and division type on prediction accuracy of the next signal type.

The results confirmed the length of 3σ from μ as the ideal interval for the division and together with the FG division provided the best performances with equal accuracy of 94% when using a training set with 40 samples. Since the more partitioned division (i.e. FG) could offer potentially valuable advantages, as it provides more detailed temporal information, both divisions were selected for further experimentations.

From the training perspective, 40 samples offered the best learning performance, although the model performed reasonably well already from 10 samples (82% accuracy). The classification model did not show any signs of overtraining, however, there was not enough samples in the dataset to examine training sets larger than 40 samples.

5.2 Model Comparison

We built instances of the proposed and Huang's classifier in order to compared their prediction abilities to each other using a standard leave-one-out test.

Our BBN model displayed superior predictive accuracy of 99.9% and 99.3% in both tested properties, the command type and the type of the next signal, resp., and outperformed the model by Huang et al., which achieved accuracies of 74.0% and 75.5% on the same data. A comparison bar plot is shown in Fig. 5.5.

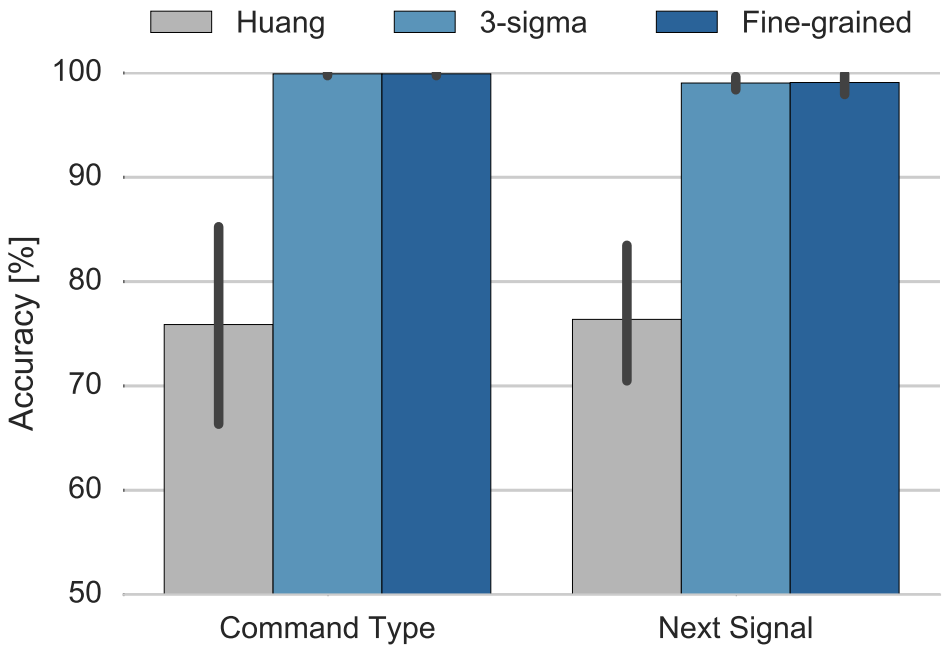


Figure 5.5: Accuracy comparison of two predicted properties between the proposed model and the one introduced by Huang et al.

The test also confirmed the performance equality of the 3σ and FG divisions in terms of their predictive accuracy.

5.3 Discussion

The remarkably high predictive accuracy of the developed model confirmed our findings according to the multimodal integration patterns and demonstrated that their proper interpretation results in more consistent and precise models.

An important property of the proposed classifier is also its short period required for training. As demonstrated, it provides accuracy over 80% already from 10 samples, which makes it a great candidate for employment in multimodal systems with real-time user adaptability.

From the perspective of continuous variables encoding, two promising divisions (3σ and FG) were discovered. While both are similar in their performance, the FG could potentially offer advantages as it provides more detailed temporal partitioning.

Chapter 6

Applications of Multimodal Integration Patterns Modeling

6.1 Applying Modeling to Improve Response Time

The previous experiments and measurements demonstrated that our BBN model provides very accurate prediction of the next signal in a sequence of inputs. We will use this feature to address the multimodal input segmentation.

The main objective is to minimize the wait periods and, as a result, gain significant improvements in response time. The plan is to benefit from the model's predictive accuracy as well as its user adaptability. Instead of relying on fixed thresholds for wait periods — as is the common solution used in multimodal systems — a segmentation procedure to be introduced will utilize user-specific temporal characteristics captured in the proposed BBN model.

6.2 Procedure of Input Segmentation

In order to provide a solution to the real-time multimodal input segmentation with low response time, we have developed a new procedure that utilizes the mentioned features of our BBN classifier. A description of an algorithm behind the procedure follows.

- Every time an end of the current input signal is detected, do recurrently:
  1. set an input value of the current signal to *SILENCE*,
  2. update the onset and overlap differences accordingly to elapsed time and encode (discretize) them,
  3. perform prediction using the BBN classifier,
  4. repeat (from the step 2) until the result of the prediction is *SILENCE* (i.e. no other input is expected) or until the arrival of another input signal

If the *SILENCE* value is predicted by the classifier, the previous signal is considered the last one in the multimodal unit, and thus the end of the segment is reached and the fusion of inputs can be eventually performed.

Note that the step 3 needs to be performed only if the temporal difference is distinctive from the previous iteration (i.e. it falls to a different interval according to the chosen division) and causes a change in the input variables.

In theory, this procedure should favor the FG division and potentially utilize its more partitioned binning to predict the end of the segment much earlier (e.g. when the difference is larger than σ but still within 2σ or 3σ).

6.3 Measurements and Results

In order to find the best result possible and achieve an improvement in the response time while preserving the high level of accuracy at the same time, two different approaches to select a training set for building a model were used for testing and evaluation. In the first approach, data from more users (or user groups) were used for the training, whereas in the second, data from a single subject were utilized for both training and testing procedures. The latter method is expected to provide better results, especially considering the response time improvement, since the model should reflect subject's behavior more precisely. The first approach, on the other hand, should be more flexible and robust if there is a small amount of training data or no data is available for the tested subject at all.

According to the recent empirical evidence and conducted user study, users can be classified into two distinctive groups (i.e. SIM_R and SEQ_R) by their

multimodal synchronization pattern. We decided to explore if there are advances in separating the subjects by the synchronization groups for the purposes of model training and evaluation.

In the first measurement, a standard leave-one-out test was used for evaluation. A classification model was built and trained for every subject. A testing dataset was built with data from the tested subject and data from all other subjects were used for the learning process. The created classifiers was then employed to evaluate response time using the procedure described earlier in this section.

Average response time measured across all subjects was 1.11 seconds (SD = .38) with prediction accuracy of 99% when the 3σ division was utilized to encode time differences. The FG division brought small improvements and provided average response time of 1.05 s (SD = .34) while maintaining the same accuracy. The response time was slightly lower for SIM_R integrators and higher for SEQ_R integrators in both cases, but the difference was only marginal (see Table 6.1).

Subject Group	3-sigma			Fine-grained		
	Resp. Time	SD	Accuracy	Resp. Time	SD	Accuracy
All	1.11 s	0.38 s	99.0%	1.05 s	0.34 s	99.0%
All (SIM)	1.08 s	0.35 s	99.3%	1.03 s	0.32 s	98.8%
All (SEQ)	1.16 s	0.43 s	98.6%	1.09 s	0.36 s	99.7%

Table 6.1: Average response time gained using a classifier trained on data from all subjects.

The second measurement was very similar to the previous one. The only difference was that subjects were separated by their dominant temporal synchronization pattern into SIM_R and SEQ_R groups. Evaluations were then performed using the classifiers trained on data from subjects within the same integration pattern group as the tested individual.

Average achieved response time is presented in Table 6.2. Using the 3σ division, it was .56 seconds (SD = .37) for subjects belonging to the SIM_R group and .93 s (SD = .31) for SEQ_R integrators with average accuracy just above 99% in both cases. The FG division introduced decrease in the response time for SEQ_R integrators to .78 s (SD = .28), which equals to a relative improvement of 16%. In this case, prediction accuracy slightly dropped to 98.4%. No difference was observed for subjects with the SIM_R integration pattern.

Subject Group	3-sigma			Fine-grained		
	Resp. Time	SD	Accuracy	Resp. Time	SD	Accuracy
SIM	0.56 s	0.37 s	99.3%	0.56 s	0.37 s	99.7%
SEQ	0.93 s	0.31 s	99.2%	0.78 s	0.28 s	98.4%

Table 6.2: Average response time gained using classifiers trained on data from SIM and SEQ integrators separately.

User-specific models were built and evaluated in the last experiment. The first 40 samples[1] from the subject's dataset were for training and the last 35 for evaluation.

The user-specific models provided average response time of .62 s (SD = .32) using the 3σ division when encoding temporal differences of onsets and overlaps. Utilization of the FG division again introduced a drop in response time to .52 s (SD = .31), which represents a relative improvement of 16% (see Table 6.3). The prediction accuracy was 95.5% equally for both. Looking at the results from a user group perspective, average response time for SIM_R integrators was .45 s (SD = .38) and .39 s (SD = .33) using the 3σ and FG division, resp. A relative improvement of the FG division over the 3σ is 13% in this case. For SEQ_R integrators, average response time was .95 s (SD = .30) with the 3σ and .78 s (SD = .27) with the FG division, which is a relative improvement of 18% in favor of the latter. The decrease in the accuracy against other training methods was mainly observed for SIM_R integrators, where it dropped to 94% in average.

Subject Group	3-sigma			Fine-grained		
	Resp. Time	SD	Accuracy	Resp. Time	SD	Accuracy
User-spec.	0.62 s	0.35 s	95.5%	0.54 s	0.31 s	95.5%
U-spec. (SIM)	0.45 s	0.38 s	94.1%	0.39 s	0.33 s	94.1%
U-spec. (SEQ)	0.95 s	0.30 s	99.0%	0.84 s	0.27 s	99.0%

Table 6.3: Average response time gained using classifiers trained on user-specific data.

A comprehensible comparison of results from all three evaluations is depicted in Fig. 6.1. The FG division utilized to encode the onset and overlap differences in the classifier model offered expected advances over the simple 3σ division in all cases. Therefore it is a preferred choice and we will focus only on model variants with this division in the following text.

[1] Provides the optimal training sample size as discussed in section 5.1.1

While the global model performed well, there are notable benefits to use more subject-adapted models. For SEQ_R integrators, only a marginal difference was observed between the group-specific and user-specific models. Both of them offered a significant drop in response time to .78 and .84 s, resp., in contrast to 1.09 s gained using the global model (a relative improvement of 28% and 23%, resp.). Even more significant improvements were observed for SIM_R integrators. Taking the global model as a reference, the group-specific model provided a relative improvement of 47% with average response time of .56 s ($SD = .37$). The user-specific classifier provided further decrease in response time to a value of .39 s ($SD = .33$) corresponding to a relative improvement of 63%.

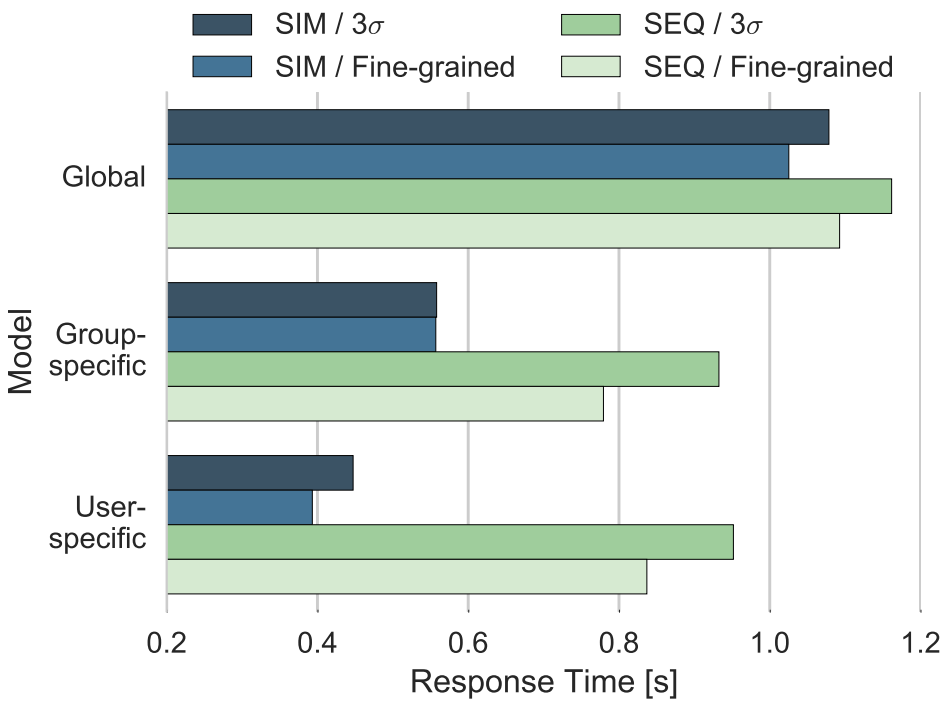


Figure 6.1: Effect of classifier model and division type on response time from perspective of different user groups

6.4 Discussion

The results confirmed our expectations about superiority of the FG over the more simple 3σ division. Therefore, the FG division is a preferred encoding variant to be implemented in user models.

The significant advancements were observed in utilization of more subject-

adapted models over a single global model. The recommendation is to use a user-specific model if sample training data are available for a given user, and a group-specific model (i.e. separately for SIM_R and SEQ_R integrators) in all other cases.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

We presented results of our user study on multimodal integration patterns in applications combining speech and gesture input. Important integration patterns described in related literature were confirmed and supplied with our own findings. Above that, we discovered a crucial shortcoming in the previously used $\text{SIM}_O/\text{SEQ}_O$ classification of the temporal synchronization pattern that causes inconsistencies across different constructions. To this end, we redefined basic conditions of the classification and gained a relative improvement of 20.7% in average consistency.

We also introduced and described a new coherent multimodal integration categorization that combines two important patterns (the $\text{SIM}_R/\text{SEQ}_R$ temporal synchronization pattern and dominant modality). The categorization seems adequately robust and general to provide a single classification offering the ability to adapt to the most significant individual differences of users during multimodal interaction.

We extended the knowledge about the individual multimodal integration patterns and encourage practitioners and researchers to consider employing them in their multimodal systems instead of ignoring them, as is the tendency at the present time. In order to successfully apply these integration patterns, designers of new generation multimodal systems should not put any presumptions on the order of involved input modes and must not constrain

the input combinations in any way considering the temporal aspects (i.e. precedence, onset, lag/overlap, etc.). Rather, the systems should be able to adapt to the individual differences in the integration patterns as described by the classification.

A new BBN classification model was designed based on the empirical evidence obtained during the user study of multimodal integration patterns. The developed model is capable of predicting the next signal in a sequence of multimodal inputs with outstanding accuracy of 99%, which is significantly superior to other results presented in the related literature.

We utilized the predictive capability and user adaptability of the classification model to address the multimodal input segmentation. To this end, the procedure of employing the classifier to segment related inputs into the multimodal (and unimodal) units was introduced. The solution provides a significant improvement in response time over the state-of-the-art approaches, while maintaining remarkably high accuracy (98–99%). To our best knowledge, the most successful approaches introduced in the related literature were able to achieve the response time between 1 and 2 seconds, at best. With our solution the response time can be improved to 0.8 s for SEQ integrators and even lowered bellow 0.5 s for SIM integrators, which represents a relative improvement of 20% and 50%, resp., at the very least. This significant decrease in response time allows a system to react more instantly on user's multimodal input with nearly real-time feedback and brings very important improvement in terms of usability, which should positively influence users' experience and satisfaction with the multimodal interactive system.

7.2 Future Work

We have demonstrated our new input segmentation procedure on the multimodal dataset with gesture-speech interaction. Nevertheless, the proposed method is general in theory and should be applicable on a broad range of other modalities as well. The decisive factor is validity of selected integration patterns (i.e. those utilized in the prediction model) for the target modality combination. In the future research, an investigation of a variety of other modality combinations focusing on an empirical proof of the integration patterns validity should be conducted. The result would provide a guidance for researchers and practitioners in the field of multimodal interactive systems indicating which combinations are appropriate for application of the introduced procedure and, conversely, which are improper.

Additionally, more participants should be involved in the future studies in order to provided more detailed statistics and to evaluate a distribution of the individual classes of users (according to the introduced categorization) in the population.

The designed and proposed user model in the form of the introduced BBN classifier (see Section 5.1) with its outstanding predictive performance encourages to be utilized beyond the input segmentation and applied to other tasks involved in the multimodal interaction processing. As an instance of such an employment would be an error detection component for fusion methods relying on machine learning techniques (e.g. in [DSL12]). We argue its utilization would effectively reduce the error rate and increase the robustness of the whole concept.

We also assume there is a possibility to use the extracted model in the security domain to detect a user or to verify its identity, since the input integration patterns exhibits unique characteristics for each subject. We would like to investigate this and also other possible applications of the classifier in our future work.

7.3 Research Contributions

The focus of this thesis is concentrated on users' multimodal integration patterns and their application to improve response time in multimodal interactive systems. The contribution of the work into this domain is summarized as follows:

- **Chapter 4**
  - An analysis of the most interesting integration patterns in terms of possible employment in the user adaptive models.
  - An introduction of a flexible multimodal framework for prototyping and development of multimodal interactive systems and its utilization to build a testing application for an empirical study.
  - Conducted tests and measurements in order to obtain empirical data of the integration patterns' characteristics.
  - An investigation of the difficulties reported regarding the accuracy of the $\text{SIM}_O/\text{SEQ}_O$ synchronization pattern.
    - A more consistent classification $\text{SIM}_R/\text{SEQ}_R$ was defined as a result of the examination.

- A new coherent categorization combining the modality precedence and the synchronization pattern was introduced providing average consistency of 95.5%.

- **Chapter 5**
  - Designed a new BBN classification model for a highly accurate multimodal input prediction.
    - The model is capable of predictive accuracy of 99%.
    - Proposed and examined different variants for encoding of continuous input variables (3σ and FG superior to others).
    - An optimal training sample size was assessed (optimal results for 40 samples; reasonable accuracy (82%) already from 10 samples).

- **Chapter 6**
  - Introduced a new procedure for multimodal input segmentation that employs the designed prediction model.
  - Performed tests and measurements to evaluate an impact of the different training sample selection strategies on the resulted accuracy and response time.
    - The user-specific models provided the best performance (0.8 s for SEQ_R and 0.5 s for SIM_R integrators with accuracy of 99%).
    - The group-specific models offered slightly higher average response time (i.e. worse), but they are convenient when there are no sample data for a specific user.
  - The proposed solution to the input segmentation provides a significant relative improvement in response time in comparison with the best approaches introduced in the related literature — at least 20% for SEQ_R and 50% for SIM_R integrators, resp.

The aforementioned contributions achieved throughout the research related to this work resulted into several articles and papers published in international journals and conference proceedings. The related publications are listed bellow and associated with the specific chapter of this thesis:

- Results related to **chapter 4** are included in:
  - Roman Hák, Jakub Doležal and Tomáš Zeman, *Manitou: A Multimodal Interaction Platform*, Proceedings of 2012 5th Joint IFIP Wireless and Mobile Networking Conference (WMNC), IEEE, 2012, pp. 79–87.

- Roman Hák and Tomáš Zeman, *Manitou: An Open Framework for Multimodal Interaction*, Proceedings of the 19th International Conference on Distributed Multimedia Systems (DMS), Knowledge System Institute Graduate School, 2013, pp. 75–78.
- Roman Hák and Tomáš Zeman, *Consistent Categorization of Multimodal Integration Patterns During Human-Computer Interaction*, Journal on Multimodal User Interfaces, 2017, *in print*.

- Results related to **chapter 5** are included in:
  - Roman Hák and Tomáš Zeman, *Improving Response Time through Multimodal Integration Pattern Modeling*, in 2016 IEEE International Symposium on Multimedia (ISM), IEEE Computer Soc., 2016, pp. 419–424.
- Results related to **chapter 6** are included in:
  - Roman Hák and Tomáš Zeman, *Improving Response Time through Multimodal Integration Pattern Modeling,* in 2016 IEEE International Symposium on Multimedia (ISM), IEEE Computer Soc., 2016, pp. 419–424.

References

[BJ08] Srinivas Bangalore and Michael Johnston, *Robust Gesture Processing for Multimodal Interaction*, Proceedings of the 10th international conference on Multimodal interfaces - ICMI '08 (New York, New York, USA), ACM Press, 2008, pp. 225–232.

[BJ09] _____, *Robust Understanding in Multimodal Interfaces*, Computational Linguistics **35** (2009), no. 3, 345–397.

[BL12] Mark Billinghurst and Minkyung Lee, *Multimodal Interfaces for Augmented Reality*, Expanding the Frontiers of Visual Analytics and Visualization (2012), 449–465.

[Bol80] Richard A. Bolt, *"Put-that-there": Voice and Gesture at the Graphics Interface*, Proc. of the 7th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '80, vol. 32, ACM Press, November 1980, pp. 262–270.

[CJM+97a] Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow, *QuickSet: Multimodal Interaction for Distributed Applications*, Proc. of the fifth ACM International Conference on Multimedia - MULTIMEDIA '97, ACM Press, 1997, pp. 31–40.

[CJM+97b] _____, *QuickSet: Multimodal Interaction for Simulation Set-up and Control*, Proceedings of the fifth conference on Applied natural language processing - ANLC '97 (Morristown, NJ, USA), Association for Computational Linguistics, 1997, pp. 20–24.

[CKB+15] Philip R. Cohen, Edward C Kaiser, M Cecelia Buchanan, Scott Lind, Michael J Corrigan, and R Matthews Wesson, *Sketch-Thru-Plan*, Communications of the ACM **58** (2015), no. 4, 56–65.

[DLI10] Bruno Dumas, Denis Lalanne, and Rolf Ingold, *Description languages for multimodal interaction: a set of guidelines and its illustration with SMUIML*, Journal on Multimodal User Interfaces **3** (2010), no. 3, 237–247.

[DSL12] Bruno Dumas, Beat Signer, and Denis Lalanne, *Fusion in multimodal interactive systems: An HMM-Based Algorithm for User-Induced Adaptation*, Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems - EICS '12, ACM Press, 2012, pp. 15–24.

[EJ12] Patrick Ehlen and Michael Johnston, *Multimodal Interaction Patterns in Mobile Local Search*, Proc. of the 2012 ACM International Conference on Intelligent User Interfaces - IUI '12, 2012, pp. 21–24.

[GA04] Anurag Gupta and Tasos Anastasakos, *Dynamic Time Windows for Multimodal Input Fusion*, 8th International Conference on Spoken Language Processing - ICSLP, 2004, pp. 1009–1012.

[HDKC+06] D. Huggins-Daines, Mohit Kumar, Arthur Chan, A.W. Black, Mosur Ravishankar, and A.I. Rudnicky, *Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices*, Proc. of IEEE International Conference on Acoustics Speech and Signal Processing, 2006, pp. 185–188.

[HDZ12] Roman Hak, Jakub Dolezal, and Tomas Zeman, *Manitou: A Multimodal Interaction Platform*, 2012 5th Joint IFIP Wireless and Mobile Networking Conference (WMNC), IEEE, 2012, pp. 60–63.

[HO06] Xiao Huang and Sharon Oviatt, *Toward Adaptive Information Fusion in Multimodal Systems*, Machine Learning for Multimodal Interaction (Steve Renals and Samy Bengio, eds.), Lecture Notes in Computer Science, vol. 3869, Springer Berlin Heidelberg, 2006, pp. 15–27.

[HOL06] Xiao Huang, Sharon Oviatt, and Rebecca Lunsford, *Combining User Modeling and Machine Learning to Predict Users' Multimodal Integration Patterns*, Machine Learning for Multimodal Interaction (Steve Renals, Samy Bengio, and Jonathan G. Fiscus, eds.), Lecture Notes in Computer Science, vol. 4299, Springer Berlin Heidelberg, 2006, pp. 50–62.

[HPSM11] Ellen C. Haas, Krishna S. Pillalamarri, Chris C. Stachowiak, and Gardner McCullough, *Temporal Binding of Multimodal*

Controls for Dynamic Map Displays, Proceedings of the 13th International Conference on Multimodal Interfaces - ICMI '11, ACM Press, 2011, p. 409.

[HZ13] Roman Hak and Tomas Zeman, *Manitou: An Open Framework for Multimodal Interaction*, Proceedings of the 19th International Conference on Distributed Multimedia Systems (Skikie, IL, USA), Knowledge Systems Institute Graduate School, 2013, pp. 75–78.

[HZ17] _____, *Consistent Categorization of Multimodal Integration Patterns During Human-Computer Interaction*, Journal on Multimodal User Interfaces (2017), *in print*.

[JB00] Michael Johnston and Srinivas Bangalore, *Finite-state Multimodal Parsing and Understanding*, Proceedings of the 18th conference on Computational linguistics - COLINGS 2000 (Morristown, NJ, USA), Association for Computational Linguistics, 2000, pp. 369–375.

[JB05] _____, *Finite-state Multimodal Integration and Understanding*, Natural Language Engineering **11** (2005), no. 2, 159–187.

[JBV+02] Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor, *MATCH: An Architecture for Multimodal Dialogue Systems*, Proc. of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, no. July, 2002, pp. 376–383.

[JCM+97] Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James a. Pittman, and Ira Smith, *Unification-based Multimodal Integration*, Proceedings of the 35th annual meeting on Association for Computational Linguistics - ACL '97 (Morristown, NJ, USA), Association for Computational Linguistics, 1997, pp. 281–288.

[Joh98] Michael Johnston, *Unification-based Multimodal Parsing*, Proceedings of the 36th annual meeting on Association for Computational Linguistics - ACL '98 (Morristown, NJ, USA), vol. 1, Association for Computational Linguistics, 1998, pp. 624–630.

[KB06] Edward C Kaiser and Paulo Barthelmess, *Edge-splitting in a Cumulative Multimodal System, for a No-Wait Temporal Threshold on Information Fusion, Combined with an Under-Specified Display*, Ninth International Conference on Spoken Language Processing - ICSLP, 2006.

[LBB+13] Minkyung Lee, Mark Billinghurst, Woonhyuk Baek, Richard Green, and Woontack Woo, *A usability study of multimodal*

input in an augmented reality environment, Virtual Reality **17** (2013), no. 4, 293–305.

[Lew12] James R. Lewis, *Usability Testing*, Handbook of Human Factors and Ergonomics, John Wiley & Sons, Inc., mar 2012, pp. 1267–1312.

[LNP+09] Denis Lalanne, Laurence Nigay, Philippe Palanque, Peter Robinson, Jean Vanderdonckt, and Jean-François Ladry, *Fusion Engines for Multimodal Input: A Survey*, Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09 (New York, New York, USA), ACM Press, 2009, p. 153.

[MKM+14] Madoka Miki, Norihide Kitaoka, Chiyomi Miyajima, Takanori Nishino, and Kazuya Takeda, *Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech*, EURASIP Journal on Audio, Speech, and Music Processing **2014** (2014), no. 1, 1–7.

[OCL04] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford, *When Do We Interact Multimodally?*, Proc. of the 6th International Conference on Multimodal Interfaces - ICMI '04, ACM Press, 2004, pp. 129–136.

[OCT+03] Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson, and Lesley Carmichael, *Toward a Theory of Organized Multimodal Integration Patterns during Human-Computer Interaction*, Proc. of the 5th International Conference on Multimodal Interfaces - ICMI '03, ACM Press, 2003, pp. 44–51.

[ODK97] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn, *Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction*, Proc. of the SIGCHI Conference on Human Factors in Computing Systems - CHI '97, ACM Press, 1997, pp. 415–422.

[OLC05] Sharon Oviatt, Rebecca Lunsford, and Rachel Coulston, *Individual Differences in Multimodal Integration Patterns: What Are They and Why Do They Exist?*, Proc. of the SIGCHI Conference on Human Factors in Computing Systems - CHI '05, ACM Press, 2005, pp. 241–249.

[Ovi99a] Sharon Oviatt, *Mutual Disambiguation of Recognition Errors in a Multimodal Architecture*, Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '99 (Pittsburgh, PA, USA), no. May, ACM Press, 1999, pp. 576–583.

[Ovi99b] _____, *Ten Myths of Multimodal Interaction*, Communications of the ACM **42** (1999), no. 11, 74–81.

[Ovi03a] _____, *Advances in Robust Multimodal Interface Design*, IEEE Computer Graphics and Applications **23** (2003), no. 5, 62–68.

[Ovi03b] _____, *User-Centered Modeling and Evaluation of Multimodal Interfaces*, Proceedings of the IEEE **91** (2003), no. 9, 1457–1468.

[RMM+04] Leah M. Reeves, Jean-Claude Martin, Michael McTear, TV Raman, Kay M. Stanney, Hui Su, Qian Ying Wang, Jennifer Lai, James A. Larson, Sharon Oviatt, T. S. Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, and Ben Kraal, *Guidelines for Multimodal User Interface Design*, Communications of the ACM **47** (2004), no. 1, 57.

[SHS+14] Felix Schüssel, Frank Honold, Miriam Schmidt, Nikola Bubalo, Anke Huckauf, and Michael Weber, *Multimodal Interaction History and its use in Error Detection and Recovery*, Proc. of the 16th International Conference on Multimodal Interaction - ICMI '14, ACM Press, 2014, pp. 164–171.

[SN10] Marcos Serrano and Laurence Nigay, *A wizard of oz component-based approach for rapidly prototyping and testing input multimodal interfaces*, Journal on Multimodal User Interfaces **3** (2010), no. 3, 215–225.

[SPH98] R. Sharma, V.I. Pavlovic, and T.S. Huang, *Toward Multimodal Human-Computer Interface*, Proceedings of the IEEE **86** (1998), no. 5, 853–869.

[VW96] Minh Tue Vo and Cindy Wood, *Building an application framework for speech and pen input integration in multimodal learning interfaces*, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 6, IEEE, 1996, pp. 3545–3548.

[WWL07] Jacob O Wobbrock, Andrew D Wilson, and Yang Li, *Gestures without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes*, Proc. of the 20th annual ACM Symposium on User Interface Software and Technology - UIST '07, ACM Press, 2007, pp. 159–169.

[XGO02] Benfang Xiao, Cynthia Girand, and Sharon Oviatt, *Multimodal Integration Patterns in Children*, Proc. of International Conference on Spoken Language Processing, 2002, pp. 629–632.

[XO03] Benfang Xiao and Sharon Oviatt, *Modeling Multimodal Integration Patterns and Performance in Seniors : Toward Adaptive*

Processing of Individual Differences, Proc. of the 5th International Conference on Multimodal Interfaces - ICMI '03, 2003, pp. 256–272.

List of Publications

Publications related to the topic of this thesis

Papers in journals with impact factor

- Roman Hák and Tomáš Zeman, *Consistent Categorization of Multimodal Integration Patterns During Human-Computer Interaction*, Journal on Multimodal User Interfaces, 2017, *in print.*
  - *authorship:* Hák – 50%, Zeman – 50%

Conference papers listed in WoS

- Roman Hák and Tomáš Zeman, *Improving Response Time through Multimodal Integration Pattern Modeling*, Proceedings of 2016 IEEE International Symposium on Multimedia (ISM), IEEE Computer Soc., 2016, pp. 419–424.
  - *authorship:* Hák – 50%, Zeman – 50%

- Roman Hák and Tomáš Zeman, *Evaluation of Multimodal Fusion Methods Using Integration Patterns Modeling*, Proceedings of 38th International Conference on Telecommunications and Signal Processing, IEEE, 2015, pp. 377–381.
  - *authorship:* Hák – 50%, Zeman – 50%
- Roman Hák, Jakub Doležal and Tomáš Zeman, *Manitou: A Multimodal Interaction Platform*, Proceedings of 2012 5th Joint IFIP Wireless and Mobile Networking Conference (WMNC), IEEE, 2012, pp. 79–87.
  - *authorship:* Hák – 33.3%, Doležal – 33.3%, Zeman – 33.3%

Other papers

- Roman Hák and Tomáš Zeman, *Manitou: An Open Framework for Multimodal Interaction*, Proceedings of the 19th International Conference on Distributed Multimedia Systems (DMS), Knowledge System Institute Graduate School, 2013, pp. 75–78.
  - *authorship:* Hák – 50%, Zeman – 50%
- Roman Hák and Tomáš Zeman, *Testing and Evaluation of Multimodal Fusion Engines*, Proceedings of 15th International Conference on Research in Telecommunication Technologies, Slovak University of Technology in Bratislava, 2013, pp. 23–28.
  - *authorship:* Hák – 50%, Zeman – 50%

Other publications (not related to the thesis)

Journal papers

- Jiří Danihelka, Roman Hák, Lukáš Kencl and Jiří Žára, *3D Talking-Head Interface to Voice-Interactive Services on Mobile Phones*, International Journal of Mobile Human Computer Interaction. 2011, 3(2), pp. 50–64.
  - *authorship:* Danihelka – 40%, Hák – 30%, Kencl – 20%, Žára – 10%

Book chapters

- Jiří Danihelka, Lukáš Kencl, Roman Hák and Jiří Žára, *3D Talking-Head Interface to Voice-Interactive Services on Mobile Phones*, Developments in Technologies for Human-Centric Mobile Computing and Applications, IGI Global, 2012, pp. 130–144. ISBN 9781466620681.
  - *authorship:* Danihelka – 25%, Kencl – 25%, Hák – 25%, Žára – 25%

Other papers

- Roman Hák, Jakub Doležal and Tomáš Zeman, *ManiTutor: multimodální e-learningová platforma*, Sborník příspěvků z konference a soutěže eLearning 2012, Gaudeamus, 2012, pp. 32–37.
  - *authorship:* Hák – 33.3%, Doležal – 33.3%, Zeman – 33.3%
- Roman Hák, Jakub Doležal and Tomáš Zeman, *An Extensible E-Learning Platform Combining Speech and Graphical User Interfaces*, Proceedings of Information and Communication Technology in Education 2012, Ostravská univerzita v Ostravě, 2012, pp. 79–87.
  - *authorship:* Hák – 33.3%, Doležal – 33.3%, Zeman – 33.3%
- Roman Hák, *Multimodální webový prohlížeč Manitou*, Sborník příspěvků z konference a soutěže eLearning 2011, Gaudeamus, 2011, 59–62.
  - *authorship:* Hák – 100%
- Roman Hák, Jakub Doležal and Tomáš Zeman, *Multimodální interakce v oblasti e-learningu*, Sborník příspěvků z konference a soutěže eLearning 2011, Gaudeamus, 2011, pp. 138–143.
  - *authorship:* Hák – 33.3%, Doležal – 33.3%, Zeman – 33.3%
- Jakub Doležal, Tomáš Zeman and Roman Hák, *Voice-Driven Applications: Architecture Proposal and Performance Evaluation*, Proceedings of 2011 18th International Conference on Systems, University of Sarajevo, 2011, pp. 329–332.
  - *authorship:* Doležal – 33.3%, Zeman – 33.3%, Hák – 33.3%
- Zdeněk Bečvář and Roman Hák, *GPS Communicator - Tool for Analysis of NMEA Protocol*, Access server, 2011, 9(201102), pp. 1–5.
  - *authorship:* Bečvář – 50%, Hák – 50%

- Jiří Danihelka, Roman Hák, Lukáš Kencl and Jiří Žára, *3D Talking-Head Interface to Voice-Interactive Services on Mobile Phones* , Proceedings of Speech in Mobile and Pervasive Environments. Speech in Mobile and Pervasive Environments, 2010, pp. 1–8. (**Best Paper Award**)
  - *authorship:* Danihelka – 40%, Hák – 30%, Kencl – 20%, Žára – 10%