Multi-modal Response Generation

WONG Ka Ho

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Systems Engineering and Engineering Management

Supervised by

Professor Helen M. Meng

October 2005

Abstract of thesis entitled:

Multi-modal
Response Generation
Submitted by WONG Ka Ho
for the degree of Master of Philosophy
at The Chinese University of Hong Kong in October 2005

This thesis describes the development of a Chinese multi-modal response generation system on the Hong Kong tourism domain. We aim to enhance the user's experience through expressive multi-modal responses. In order to understand the human's natural strategies in multi-modal communication, we collected a multi-modal dialogue corpus. We extracted parts of the corpus and an expert re-designed the responses using multiple modalities with efficiency considerations. We studied the responses and show that certain modalities combinations are more preferable to others. Deictic term usages differ according to different patterns. The expert tended to use redundant information instead of deictic term for some key concept-values. Responses with two dialog acts, INFORM and SUGGEST, in ASK_ATTRACTION task goal, include more multi-modal events that reflect the complex content associated with these two dialog acts. Similar to other multi-modal systems, the information type has strong effect on the modality selection.

We have also implemented a text-to-audiovisual speech (TTVS) system to increase the computer expressive power. The TTVS system supports audiovisual speech as one of the modalities. We have applied a linear blending technique to animate a three-dimensional talking head. The talking head can provide lip movement and expressions (e.g. smile and worry). Experiment showed that the TTVS system can increase the user perception.

We designed several principles and heuristics for selecting suitable modalities in multi-modal response generation. We considered the modality preferences, preciseness, conciseness, foci and visual space constraints during the design of heuristic rules. The heuristics the TTVS system have been integratede in an end-to-end multi-modal response generation component for a dialog system in the Hong Kong tourism domain.

摘要

這篇論文集中探討中文多模態表達生成系統(multi-modal response generation)於香港旅遊資訊上的研發。我們的目的在於提高使用者對系統反應的理解程度(user perception)。我們選用了多模態表達(multi-modal response)來提高電腦系統的表現力(expressive power)以達到我們的目的。爲了了解人類在多模態表達中的行爲，我們搜集了一個多模態對話語料庫。我們把一部份對話抽了出來，並由一名專家利用多模態方法重新設計出最有效的表達作回應。我們分析了不同的表達，並顯示出某些模態組合比較常用。指示語(deictic term)的運用在多模態表達並不同於多模態輸入模式(multi-modal input pattern)。專家傾向使用冗餘信息(redundant information)多過需要使用者解構的指示語來表達某些重要的資料。告知(INFORM)及建議(SUGGEST)這兩個在詢問景點(ASK_ATTRACTION)目的(task goal)中的意向(dialog act)表達會更常使用多模態。這反映了兩個意向中的複雜信息。同其他多模態系統類似，資訊類型(information type)對模態選擇有很大的影響。

文字轉視聽系統提供視聽語音爲其中一種表達模態。因此，我們實行文字轉視聽系統(text-to-audiovisual speech system)來提高表現力。我們使用了線性混色(linear blending)技術來制作三維臉部動畫。臉部動畫能提供口唇移動及表情(expression)(例如：笑及擔心)。在測試中，文字轉視聽系統能提高使用者在聽覺上的知覺。

我們設計了多項原則及人手自訂了啓發式規則(heuristic rules)來選擇不同模態作回應。我們在編寫啓發式規則時，考慮到模態偏好(modality preference)、明確(preciseness)、簡潔(conciseness)、重點(focus)及視覺空間限制(visual space constraint)等因素。我們整合了模態選擇及文字轉視聽系統，而整合了的系統能於香港旅遊資訊領域中示範多模態的表達方式。

Acknowledgements

I would like to sincerely thank my supervisor, Professor Helen Meng, for her guidance throughout this research project. Her comments and opinion are invaluable to my research, paper and thesis writing. I also thank Helen for providing a lot of training and many chances to explore the outside. I can still remember the phone call at night to have a long discussion on my ICASSP paper. The acceptance of the paper provided me the opportunities to participate in an international conference that broadened my view. Moreover, she provides us with excellent computing resources so that we can work smoothly.

I also wish to thank the members of my thesis committee, Professor Helen Meng, Professor Chun-Hung Cheng and Professor Wai Lam from the Department of Systems Engineering and Engineering Management in the Chinese University of Hong Kong, Professor Hong-Va Leong from the Hong Kong Polytechnic University, for their time, effort and valuable advice.

As a member of the Human-Computer Communications Laboratory, I would like to thank all my lab mates. Thanks to Kon-Fan, Henry and Sam for handling many equipment purchasing issue. Thanks to Henry and Ida for coordinating the move of the laboratory in last year, especially Henry who was just joined us for a few days. I want to thank Cheryl, Henry, Ida and Michael for their share of TA workload; Tiffany and Devon for their support in the development of the text-to-audiovisual (TTVS) system; Tony for his support on machine translation in the TTVS system; Cheryl, Henry, Ida, Jessica and

Michael for their help in M3 data collection. Thanks to Ada, Chat, Edmond, Homa, Julia, Ka-Fai, Sunny, Winnie for sharing joy with me. Besides, I would like to thank Aggie and Ida for their endless support and encouragement. Moreover, I need to thank Albert from the Department of Computer Science and Engineering in the Chinese University of Hong Kong, as my partner in the TTVS system. He brought me a successful to publish a paper and acquired a chance to visit Canada.

Arthur from the Department of Electronic Engineering, Esther, Iris, Mandy, Monica, Mr. Leung and technical staff from our Department are very nice and I am appreciative of their support in these years. Most importantly, I give my most sincerely thanks to Ida for checking my writing, discussing research idea, providing technical support, taking administrative work and keeping me smile in her greatest effort.

Finally, I would like to express my very deep gratitude to my parents, brother and sister for their caring and support so I can focus on my research work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Multi-modal and Multi-media

Modality and medium can be interpreted with different meanings in different systems and tasks. Oviatt [1] uses the term "modality" to refer to a form of input and the term "medium" for output only. In Dix [2], a multi-modal system is developed to take advantage of the multi-sensory nature of humans. Humans can use auditory or tactile to improve the interactive nature of the system. On the other hand, multi-media system uses more than one media to communicate. Textual, graphical, iconic and others are interpreted as different media.

As there should have a consistent definition for multi-modal and multi-media system, we need to define the terms 'modality' and 'medium'. In Bernsen [3], graphic, acoustic and haptic are three medium where a medium is a physical realization of information. Bordegoni [4] defined modality as a "*mechanism of encoding information for presentation to humans or machines in a physically realized form*" (source: [4]). In this thesis, we use the definition of medium from Bernsen and interpret 'modality' as 'encoding mechanism' so as to provide a more detailed classification. For example, map and photo are two different modalities in the graphical medium.

1.2 Overview

We are in the computer age and almost interact with computer system everyday, from calculator, ATM machine to personal computer. The interaction channels with computers can be tactile (e.g. touch screen in ATM machine), audio (e.g. speech in telephone conversation system), etc. The task can be as simple as calculation (e.g. calculator) or complex as flight reservation [5]. More powerful computation capacity allowed us to perform more and more various tasks with computer. Each task may have its own set of commands. One possible solution to enter different commands are typing on keyboard. Entering commands by keyboard is not user-friendly enough and difficult for people with little or no experience of computers. Researcher started to develop speech interface because speech input is more natural communication channel in human and computer knowledge is not required (e.g. the actual command word to be input through a computer keyboard.

Recently, computer devices become more and more portable. The diversity of usage context (e.g. in room, on the street, in car, etc.) causes people to seek for more effective way of communication rather than speech only. On the other hand, computer devices are also able to deliver more multi-media information such as map, which is necessary for multi-modal interaction. People try to consider the possibility of multi-modal communication on human-computer interaction.

Human has experienced multi-modal interaction with human a long time. For example, when a tourist asks for the direction of path, we reply with speech and use a finger to point the direction. Human maximizes their expressive power of different information with integrate using of speech, hearing, vision, gestures and others. For example, locative information may be expressed by speech complement with pointing/circling on a map. Multi-modal also offers a stable performance in mobile computing [6]. On input side, users can switch from speech to handwriting in noisy environment. On output side, computer can switch from graphic to speech in eye-busy environment. Therefore, computer can also increase its expressive power with selecting suitable modalities.

Many multi-modal systems were developed to realize multi-modal interaction. For example, users can manipulate entities on map with speech or pen-based gesture in the Quickset system [7]. When we travel in New York, the MATCH system [8] can assist us to navigate the city. We can give a gaze to the computer in project Oxygen [9] to tell her that you are speaking to her in future. In Europe, SmartKom [10] was a mixed-initiative dialogue system which allows both user and system to influence the dialogue flow. She has combined spoken dialogs, graphical interfaces and natural gestural interactions together. August Strindberg is an author of the nineteenth century but also the animated agent in the August system [11]. August runs on a tourism kiosk which presents information using synthetic speech, facial expressions and head movements.

Some development effort has been devoted to encourage the inter-change ability of multi-modal information. Multi-modal Interaction Activity [12] seeks to extend the Web ability. Users can dynamically select the most suitable modality of interaction. VoiceXML Form's XHTML+Voice 1.1 (commonly refered as X+V) [13] combined the VoiceXML with XHTML to provide Web clients the ability to support visual and spoken interaction. Speech Application Language Tags (SALT) [14] enables telephony-enabled and also multi-modal to access different information devices. Extensible Multi-Modal Annotation (EMMA) [15] is a markup language to annotate multi-modal input at various interpretations at different system levels. In other words, it is used to provide semantic annotation following the chain of multi-modal interaction, from the low level input to various processing stage and to the multi-modal interaction manager.

Multi-modal Response Generation (MMRG) is a critical component for the overall usability and perceived intelligence of the multi-modal dialogue system (MMDS). Multi-modal dialogue systems are similar to the spoken dialogue systems (SDSs). Both of them need to generate response which is coherent, cooperative and tailored to the user's preference. MMDSs have more advantage than SDSs or other uni-modal system. On the input side, MMDS provides flexible use of input modality to convey different types of information, adapt different environments, input various tasks, fit for individual user preference and ability, etc [16]. On the output side, MMDS uses combination of modalities to overcome the diversity of user such as hear impaired, different usage contexts like noisy, improve response conciseness including the use of map to indicate location, make the response to be more informative by photo or video, etc. However, MMRG is not an easy task and more challenging than textual/spoken language generation. This is not only handle textual/spoken content but also the selection of modality and coordination of them. The presentation of multi-modal response is not as simple as just show some texts. It requires a graphical user interface, and other advanced techniques such as text-to-speech synthesizer and text-to-audiovisual speech system.

1.3 Thesis Goal

The goal of this work is to develop a multi-modal response generation system with as focuses on the expressive power maximization and the guidelines of modality selection. We maximize the expressive power by developing a text-to-audiovisual speech (TTVS) [17] system. We attempt to understand human behavior in multi-modal response communication and based on our understanding, design the guidelines of modality selection in response generation. The domain selected is Hong Kong tourism domain. The long-term goal of this work is to integrate with a multi-modal understanding component to become a multi-modal tour guide of Hong Kong.

In order to understand the human behavior of multi-modal response, a Wizard-of-Oz setup was used to collect multi-modal dialogues. A communication expert designed responses on some collected dialogues. We studied the relation among modality, deictic term, intention and information type. The definition of deictic term is[1], "*the function of pointing or specifying from the perspective of a participant in an act of speech or writing; aspects of a communication whose interpretation depends on knowledge of the context in which the communication occurs.* " Some intentions and information types are more preferrable than the use of multi-modal. Modality selection of one modality is depended on some of the other modalities to be a combination.

We developed a text-to-audiovisual speech (TTVS) system to provide verbal communication and also non-verbal, non-vocal communication. The TTVS system received spoken content, prosody parameters and body action parameters to generate a lip-synchronized facial animation. We studied the relationship between Cantonese syllable and visemes label mapping. The definition of viseme is[2], "*A viseme is a generic facial image that can be used to describe a particular sound. A viseme is the visual equivalent to a phoneme (unit of sound in spoken language).*" A blending technique was applied to achieve a real-time facial animation. For the lip-synchronization, a linear interpolation calculates the visual parameters corresponding the transition of sub-syllable in speech.

We focus on modality selection in MMRG. We observed several responses which required the use of multi-modal. We use heuristic rules to perform

[1] WordNet from http://www.answers.com/topic/deixis

[2] http://whatis.techtarget.com/definition/0,,sid9_gci213308,00.html

modality selection for the responses. Our goal is to devise a set of selection principles which is possible for applying on different context. A preliminary system is developed to support MMRG.

1.4 Thesis Outline

This thesis is organized as follows: Chapter 2 describes the MMRG structure, previous work in TTVS system and modality selection. Chapter 3 limits our information domain. Chapter 4 discusses the data collection process and foundings so as to support multi-modal response generation. Chapter 5 presents the details of text-to-audiovisual speech system. Chapter 6 introduces our attempt on modality selection and details of the preliminary system. Finally, conclusions and future work are provided in Chapter 7.

Chapter 2

Background

This thesis explores the modality selection and the use of text-to-audiovisual speech (TTVS) [17] in multi-modal response generation (MMRG). MMRG refers to the use of different modalities to generate a concise and informative response. MMRG includes many tasks and requires TTVS system to present the generated response. Multi-modal fission [10] is one of the major tasks in MMRG. One of the challenges in MMRG is how to select a set of modalities to convey a message. We first study the major task in MMRG, i.e. multi-modal fission, to understand the sub-tasks of it and is described in Section 2.1. Multi-modal fission means using combinations of modalities to present a message.

Since there are many different modalities, we want to know when, which and how different modalities should be selected in natural multi-modal communication. Therefore, we introduce how people collect multi-modal corpus to support the understand of human behaviour in Section 2.2.

During the study, we have found that many multi-modal dialogue systems (MMDSs) include the use of a talking head or an animated agent [11]. They are one kind of modalities of MMDS and provide a non-verbal and non-vocal communication in MMDS. Therefore, we will discuss the talking head generation system, namely TTVS system in Section 2.3. Different approaches in modality selection will also be presented in Section 2.4.

2.1 Multi-modal Fission

Foster [18] has stated in his paper that "*fission is the process of realising an abstract message through output on some combinations of the available channels*" (source: [18]). Foster [18] classified multi-modal fission into three categories. Figure 2.1 shows the one of the potential architecture of MMRG. The user request and response intention will be given by multi-modal understanding and dialogue manager respectively. Modality fission selects and organizes the response content. The text-to-audiovisual speech system drives a talking head to speak out the content. The system needs to co-operate with text-to-speech synthesizer and facial animation generator. Other information such as map and photograph will be shown on graphical user interface. Finally, user receives the response. We will discuss the tasks in modality fission components. The first one is content selection and structuring as follow:

- Content Selection and Structuring
  Multi-modal dialogue system (MMDS) needs to provide specific information relevant to the user's request. It needs to determine cooperative intention and related semantic content. Research effort has been devoted to the development or use of different theories to capture the intention exchange in dialogue. Seneff [5] uses the rule-based approach to determine suitable intention in flight reservations domain. A set of ordered rules was included in dialogue control and determined what to say. The rules development required a lot of linguistic knowledge and man-power. Meng [19] used the transition rules to determine the response intention. The transition rules mapped the user intention to response intention where the rules were automatically devised from dialogue corpus. It required

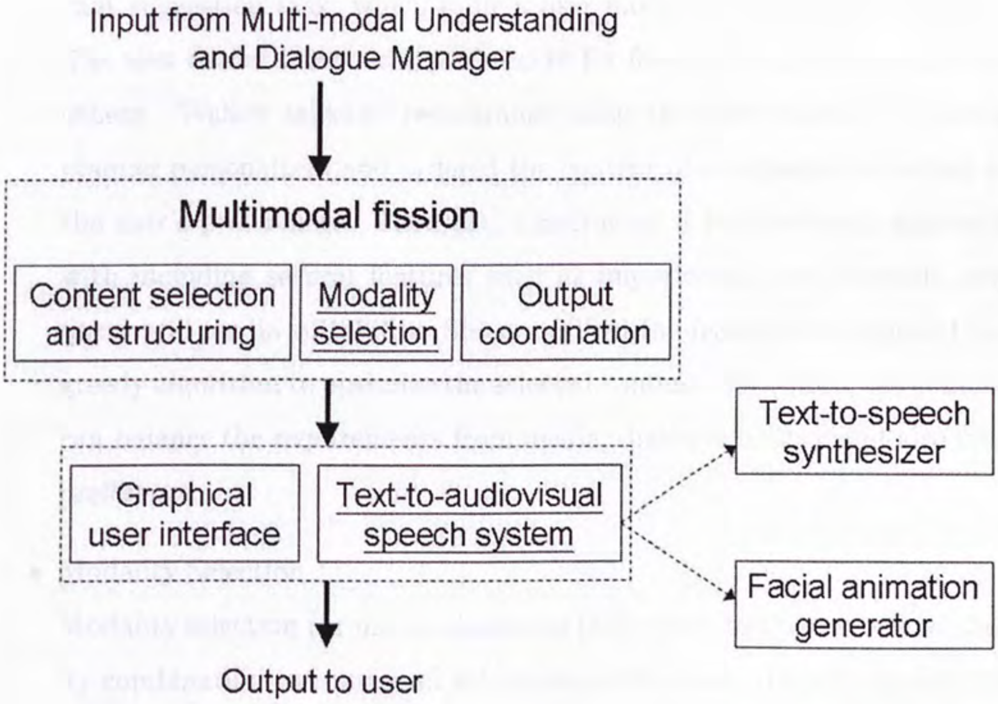

Input from Multi-modal Understanding and Dialogue Manager
Multimodal fission
Content selection and structuring
Modality selection
Output coordination
Graphical user interface
Text-to-audiovisual speech system
Text-to-speech synthesizer
Facial animation generator
Output to user

Figure 2.1: A graphical illustration of the tasks and flow in MMRG. We will briefly discuss most of the tasks. The tasks we are focused on are underlined and their details will be discussed in Sections 2.3 and 2.4.

enough data to training.

There are many considerations in content selection, including user preference, user query and dialog history. Different researchers tried to consider different set of factors. Walker [20] included a city navigation guide in New York City and also tailored the suggestion to user on the restaurant suggestion task, which built a user model for restaurant selection. The user model shows user preferences for food quality, service cost and others. Walker selected restaurants using the user model. A speech planner personalized and ordered the content of a response according to the user's preferences. Zhou [21] constructed a feature-based approach with including several features such as importance, user interest, user query and media-suitability. She quantified the features and applied the greedy algorithm to optimize the selected content. Therefore, the content can balance the requirements from media, dialogue history and also user preference.

- Modality Selection

  Modality selection (or media allocation [21]) refers to the use of a modality combination to convey all information effectively [18][22]. André [22] provided a summary about the modality selection: "*Given a set of data and a set of media, find a media combination that conveys all data effectively in a given situation*" (source: [22]). To perform modality selection, researchers tried to study and use different sources of information such as information type (e.g. location or abstract action [23]), multi-modal interaction properties (e.g. redundancy or complementarity [24]), modality characteristic (e.g. permanent or transient [25]), available resources (e.g. screen size on mobile phone [21]) and user characteristic (e.g. visual-impaired, elderly or interests [22]). Researchers analyze the characteristics and their effects when they design an approach to select a set of modalities to generate a multi-modal response. Modality selection is a new research problem which does not occur in spoken dialogue system. We will present several approaches for modality selection in Section 2.4.

- Output Coordination

  When a response includes more than one modalities, we need to coordinate the use of deictic terms, physical layout and temporal relation [18]. Deictic term provides a linkage as cross-modal reference such as '我建議去哩度[pointing to the photo of the Western Market]'.

Physical layout coordination, refers to the arrangement of visual objects in proper positions. For example, we have to put the name of a photograph closes to its corresponding photograph. User may not be able to match the photograph with its name when they are far apart on the display.

Temporal coordination means that two or more modalities should occur in a coherent time. An example is shown in Table 2.1 to demonstrate the importance of temporal coordination. If the occur time of two pointings (one is point to the Peak tram and the other one is to the Peak) are swapped in Table 2.1, then the user will be confused with the location of taking the Peak tram and the location of the Peak.

User 1:	我可以點樣上去太平山?
System 1:	你可以在這裏[pointing to the Peak tram]乘山頂纜車到太平山 [pointing to the Peak].

Table 2.1: An example about the importance of temporal coordination.

2.2 Multi-modal Data collection

2.2.1 Collection Time

In order to understand how human input and output with different modalities, researchers simulated MMDS and collected a set of multi-modal corpora. There are different ways to collect data in different stages. Typically, a large multi-modal data collection in early stage of the development was performed. Multi-modal data includes recording from different modalities (e.g. speech and body gesture) and relationship (e.g. temporal and spatial) among them. SmartKom [28] collected 1,500 GByte multi-modal data. Rapp [29] argued that collecting data at early stages of the development could not reflect the change of setup of the system. The collected data did not match the developing system. It is possible to make the collected data to be useless. Therefore, Rapp [29] collected data in several times to reflect different stages of development.

2.2.2 Annotation and Tools

Annotation is a challenge in multi-modal corpus. W3C [30] provided Extensible MultiModal Annotation (EMMA) specification [15] for annotating different multi-modal input. SmartKom [28] considered how to annotate different modalities using consistent file format (BAS Partitur Format). MMAX [31] was an annotation tool for multi-modal corpus, which annotates data into XML format. However, Garg [32] evaluated different annotation tools and decided to use different annotation tools for different parts (e.g. using MMAX for making labels but using PERL scripts to convert data between different formats).

2.2.3 Knowledge of Multi-modal Using

Much research work has been devoted to explore modality using knowledge from corpus. On the input side, Oviatt [33] found input pattern was related to the task difficulty. Users tend to use multiple modalities in very difficult task such as "*Place a maintenance shop near the intersection of I-405 and Hwy 30 just east of Good Samaritan*" (source: [33]). When the dialogue turn was related to new context, users were also likely to use more multi-modal. For the output side, Kruijff-Korbayová [34] discovered how human present the search results about a list of MP3. Human presented the results by displaying a screen outputs and describing what is shown. The comments about same screen output (e.g. the using of table) from different people was varied. Some people think that it was useful but some may not.

2.3 Text-to-audiovisual Speech System

Modalities can be classified into four classes as shown in Table 2.2 [26]. Human uses the combination of different modalities to communication. We use speech in public speaking and text in writing. We use non-verbal and vocal communication such as higher volume to emphasis message. Non-verbal and non-vocal communication such as lip-movement help user to recognize speech output. Facial expression allows system to provide positive feedback by smile.

SmartKom [10], August [11] and COMIC [27] are aware of the use of non-verbal, non-vocal communication. They provided a talking head or an animated agent like the Smartakus in Smarktom to support the present of lip-movement and emotion [35]. Talking head system can be applied on different applications such as language learning [36]. Talking head is generated by a TTVS system. As it is important and useful to use talking head in MMDS, we illustrate several approaches about TTVS system.

	Verbal (symbolic)	Non-verbal (non-symbolic)
Vocal	Spoken communication, etc.	Different prosody, sound effect, etc.
Non-vocal	Textual communication, etc.	Lip movement, facial expression, graphic, mouse gesture, etc

Table 2.2: Communication matrix of verbal/non-verbal versus vocal/non-vocal behaviors. Examples of the each type of communication are shown.

Figure 2.1 shows that TTVS system cooperates text-to-speech synthesizer and facial animation generator. TTS synthesizer generates speech and different prosodies. Facial animation generator generates facial expressions such as smile and worry. TTVS system focused on synchronization between the output of TTS synthesizer and facial animation generator.

2.3.1 Different Approaches to Generate a Talking Heading

A talking head can be generated from two different kinds of input: text or speech. Nakamura [37] worked on audio-to-visual speech system, which is based on acoustic signal to generate talking head. We focus on text input in our work which is a TTVS system. Figure 2.2 illustrates the classification of current talking head approach [38]. Different approaches for generating talking head will be discussed as follows.

- Model-based approach

  Geometric modeling and bio-mechanical modeling are the modeling methods in model-based approach. Geometric modeling addresses a talking head model with a set of three-dimensional coordinates (i.e. vertices). It provides an easy way to translate, zoom and rotate the talking head. The talking head model can be adapted to be a specific person easily. Although the talking head using geometric model is less realistic when

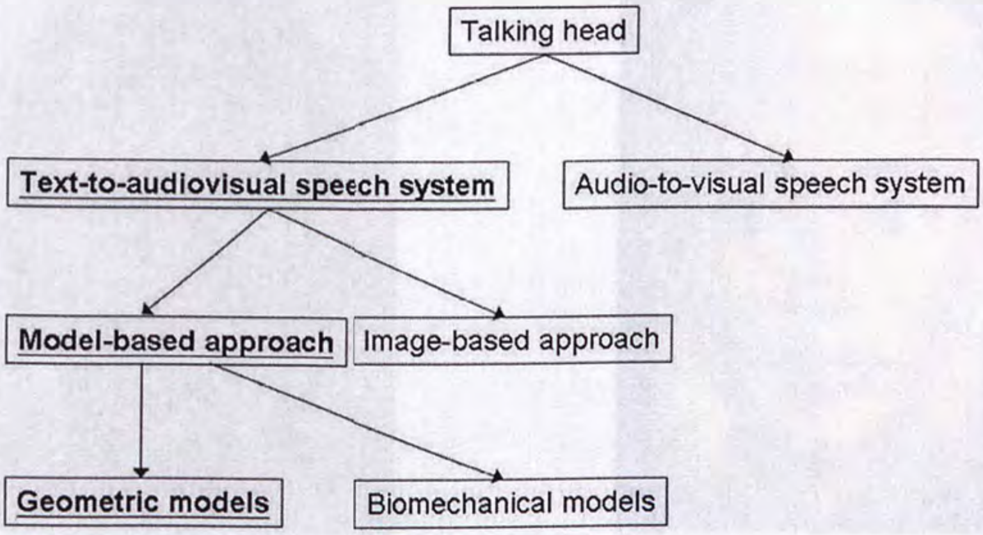


Figure 2.2: Talking head system can be classified into different approaches. The approach we used is underlined.

compared with the one of image-based approach (Figure 2.3 shows a comparison of two approaches), the high flexibility makes it to be a common approach in TTVS system. Moreover, geometric model fits to the use of digital entertainment, virtual character and language learning. SYNFACE [40] used this approach on talking face telephone.

Geometric modeling does not include the physiological mechanisms in a head. Bio-mechanical modeling [41] tries to include the muscle movement mechanisms and simulates the real talking head. However, the complexity in this approach is very high and requires extensive data collection and calculation.

- Image-based approach

  In image-based approach, the talking head is represented by two-dimensional photograph. It provides a high photo-realistic talking head. TTVS sys-

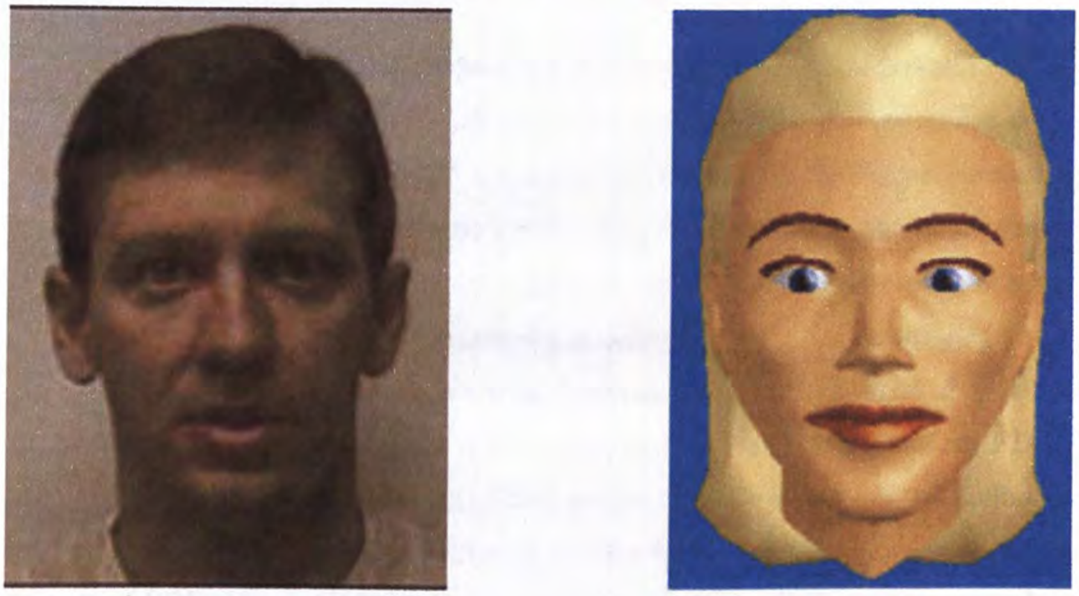


Figure 2.3: Pictures of two talking heads from image-based approach (left) and model-based approach (right). The left picture is borrowed from MikeTalk [17] and the right one is borrowed from SYNFACE [40].

tem using image-based approach is similar to concatenation TTS system. The TTVS system concatenates several pre-recorded images. The challenge of this approach is the transition between two images. Ezzat [17] developed a solution which is using morphing technique between two viseme images. Bregler [39] designed the video rewrite approach. Video rewrite is similar to concatenative speech synthesizers but video rewrite concatenated the proper sequence of visemes instead of phonemes.

2.3.2 Sub-tasks in Animating a Talking Head

There are a few sub-tasks to create and animate a talking head. We discuss the sub-tasks of animating a talking head briefly as follows:

- Facial definition

In MPEG-4 standard[1], a face should be defined by Face Definition Parameters (FDP) [42]. FDP is a set of parameters to describe a face model. LUCIA's [43] 3D talking head is based on MPEG-4 standard and includes co-articulation model. The co-articulation model captures the effect from current face's shape to the before or after face's shape. Although FDP is an international standard, Bailly [44] mentioned that precisely define a face need more parameters and the number of FDP is not enough. Moreover, it is very complex to generate a face from FDP. Therefore, we do not adopt FDP to our model. A face model includes a lot of vertices. The number of vertices is much more than the number of FDP. Our system directly included all vertices which are necessary for generation.

- Facial animation

When we have a face model, we need to consider how to animate the model. Animation is composed with a series of frame. We need to calculate the coordinate of each vertex of face in next frame based on current frame.

There are three possible ways to animate a face model. The first way is pre-defined all movement of each vertex associated with each parameter for each frame. However, this solution may lost the flexibility of animating different characters. Second is deformation. Radial basis functions (RBF) can be used in facial animation [45]. We need to consider the performance of solution. This is because the required standard frame rate is quite high (around 30 frames per second) in computer animation. We only have very short time (around 0.03 second) to finish the calculation of a frame. This solution requires complex calculation to estimate all vertices movement in animation. Third way is to pre-defined a set of key

[1] MPEG-4 Standard, http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm

frames. If the key frame is known, the problem has become a morphing problem. Many morphing techniques (e.g. linear interpolation) can be used in the calculation of the intermediate frames. The third solution is simple when compare to the second one and has enough flexibility when compare to the first solution. Therefore, we adopt the third way in our TTVS system.

- Visual parameters generation

  When we can produce a particular face model and understand the way of animating the face model with visual parameters values, the next problem is defining visual parameters to control the animations. There are many different definition of parameters. In MPEG-4, there are Facial Animation Parameters (FAPs) [46]. The aim of FAP is not to define a face model. It aims to control the animation synthetic face model [47].

  Some other set of parameters focus on shape of lips. Motion unit is set of vertices coordinates. A MU represents the face shape when subject is pronouncing an English phoneme. Hong [48] collected a video corpus with index of English phonemes to learn motion unit (MU). In [49], the visual feature vectors represent the shape of lips. Audio feature vectors have been mapped to visual feature vectors. In our TTVS system, we use a set of weights to control the animation and linear interpolation on parameter generation.

2.4 Modality Selection

We will discuss different approaches in modality selection as one of the tasks in MMRG.

2.4.1 Rules-based approach

Many systems use rules to select modalities. COMET (COordinated Multimedia Explanation Testbed) [23] distinguished six different types of information such as locations and physical attributes (e.g. size and shape of an object). For different information types, COMET specified certain modalities on them. For example, a rule said that use graphics alone for locations and physical attributes. In abstraction actions and relations among actions, COMET would use the text only but use graphics and text for simple and complex actions. The rules are from user feedback for the experiment with hand-design display.

Arens [25] studied the modality characteristics (transient, visual, easy to capture user attention and others) and information characteristic (two dimensional, urgency, volume and others). Then, Arens described some possible rules such as not using a transient modality when the information amount is large, using the modality could capture user attention for urgent information.

Rules-based approach requires the knowledge of how human uses multimodal as output. The collected data provided us the necessary information. Moreover, rule-based approach is the simplest one. Therefore, we will use this approach as our starting point toward a multi-modal dialogue system.

2.4.2 Plan-based approach

SmartKom [10] uses the techniques in WIP [50] (Wissensbasierte Prasentationsplanung, Knowledge-based Presentation Planning) and PPP (Personalized Plan-based Presenter) [51] for modality fission. The modality selection process is a plan-based approach. The dialog manager provides the presentation goal which is defined as modality-free representation markup language (M3L) [10]. SmartKom [10] used presentation strategies to decompose the presentation goal into many sub-tasks. Modality using is pre-defined in the sub-tasks planning. For example, the presentation goal is showing a location of an attraction. It can be divided into two sub-tasks: show the location of the attraction and find an indicator to point the location of that attraction. We need a map to achieve the first sub-task while the second one can be achieved by mouse gesture.

Plan-base approach provides a detailed control of presentation. However, the designing of plan is challenging when the task is more and more complex. The management of different plans will be another problem.

2.4.3 Feature-based approach

Zhou [21] performed modality selection during content selection. She called the process to be media allocation. She defined a set of quantified feature. Some of quantified features are:

- Media-suitability: The fitness of a media to express the data. The media-suitability modeling was based on media properties and data semantic categories. For example, numerical information should be presented by text rather than speech.
- Media-capability: Media-specific designers modeled the capability of certain media. For example, media-specific designers could assign that graphics did not capable of expressing spatial distance (set the corresponding media-capability values to be lower) due to technical limitation.
- Time cost: The feature measured the time needed to present a concept using speech.
- Unit space cost: Unit space cost was the amount of pixel needed to convey a concept on screen.

Zhou's work also considered other features such as user interest and interaction history. After quantified all features, all features contributed an overall desirability and a greedy algorithm was applied to maximize the desirability. The optimization process selected the content and also the media. This approach ensured the selected content to be presentable in a given media. The optimization-based approach balances the different requirements.

The advantages of feature-based approach is that many research techniques can be applied on the quantified features. We can cluster responses and observe the co-relationship between response and modality. We can apply data mining on the quantified features. However, this approach requires quantification of all features. It is an exhaustive task to quantify all features in the world.

2.4.4 Corpus-based approach

Corpus-based appraoch uses the statistic results in modality selection. COMIC [27] generates deictic gesture, using a simulated mouse pointer, to locate objects on screen. Foster [27] collected a set of multi-modal corpus. She considered two features (the first reference or not, was deictic term or not) of a speech reference. A weighted random choice to determine to use a deictic gesture or not based on the probability about having a deictic gesture when given the two feature values. Similar technique was applied to determine the type of deictic gesture (pointing, circling or moving the image).

2.5 Summary

In this chapter, we have given an overview of the tasks in multi-modal response generation. They are content selection, modality selection and output coordination. Then, we have presented one of the supporting system, which is TTVS system. We have introduced several approaches on TTVS system. We have presented several multi-modal data collection work. Some explored knowledge was presented as examples of multi-modal using knowledge. For example, more multi-modal inputs are occurred when the task is more difficult.

Finally, we have discussed approaches of modality selection.

Chapter 3

Information Domain

We have chosen tourist information domain as our information domain with a focus on the Hong Kong Tourism Board webpage —DiscoverHongKong[1]. This is because there is an increasing contribution of tourism towards Hong Kong's economy and it is of local and public interests. In this chapter, we will describe the tourist information domain and the way we organize the information to suit our task.

3.1 Multi-media Information

A tourist can discover a place in different angles, including shopping, accomodation, touring, gourmet, heritage, events, conventions, meeting, incentives, etc. We choose to focus on the information of touring. This is because touring information domain covers a wide range of information, including tour, attraction, activity, transportation, event, fare, etc. Different media of information have been provided in the Hong Kong Tourism Board website so as to fit the needs of different tourists. Multimedia information includes maps, photographs, animation, text, video, audio and hyperlink are used. Figure 3.1 shows a capture of a url with different multimedia information. The pho-

[1] DiscoverHongKong, http://www.discoverhongkong.com

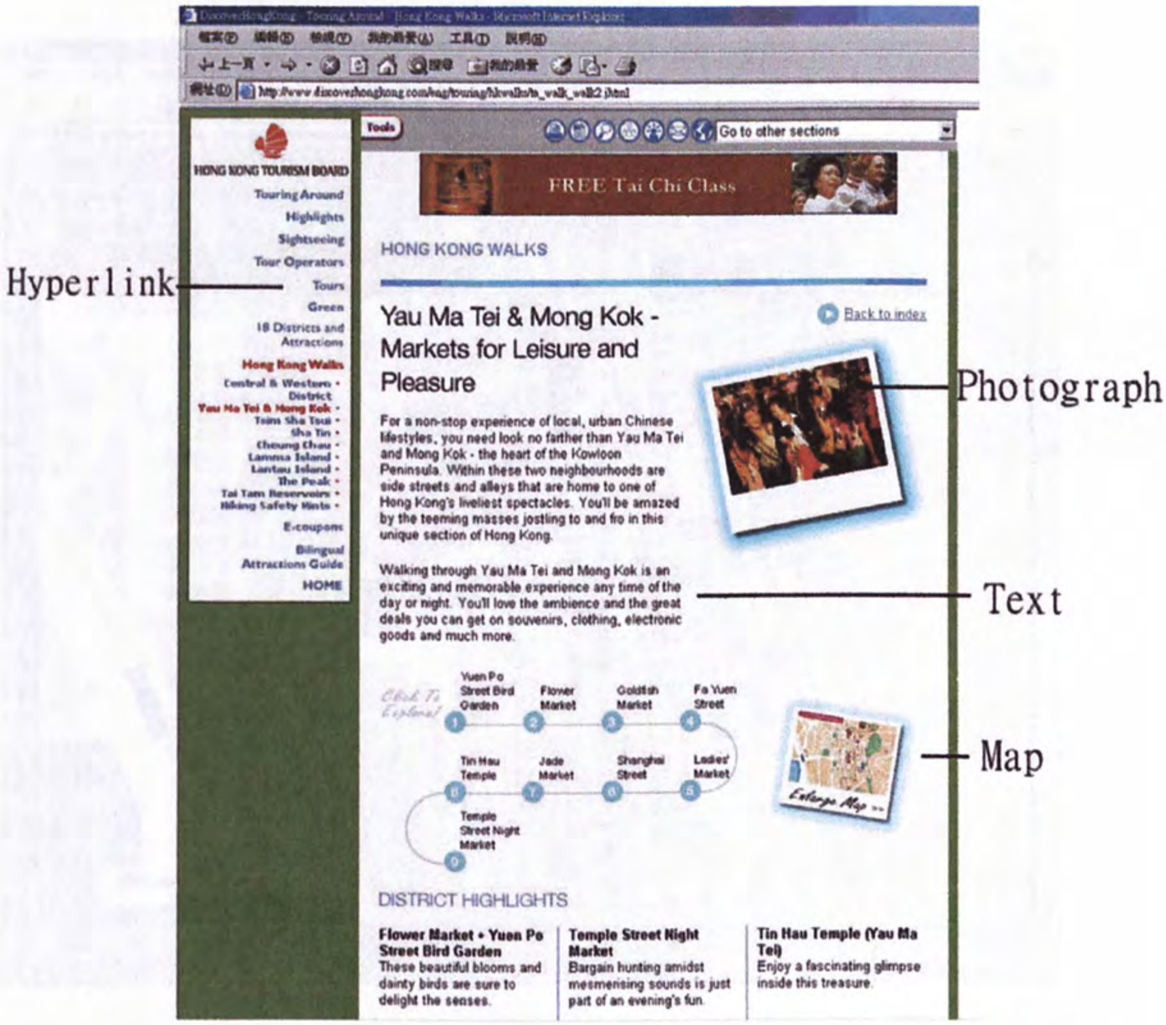
tograph shows the appearance of the streets and markets. Descriptions on different attractions are shown in text. User can click on the hyperlink of a map and get a map which is opened in a new web browser. User can get the location or transportation information by looking at the map opened. Figure 3.2 shows an example of the enlarged map in the new web browser. The map contains information of many landmarks, markets and a recommended walking route (shown in orange line).

Multiple modalities are selected to use to present the multi-media information in the Hong Kong Tourism Board website. Table 3.1 shows a mapping between different modalities and media of information presented. Map and photograph may not always be embedded into a url. Therefore, we consider them as an independent modality. A url mainly contains textual information. Since our goal is to build a multi-modal dialogue system, we expect that speech will be one of the modalities supported. Table 3.1 only shows the modalities that must be supported by the tourist information domain. We expect more modalities will be discovered after data collection and anaylysis, which will be mentioned later.

3.2 Task Goals, Dialog Acts, Concepts and Information Type

3.2.1 Task Goals and Dialog Acts

Task goal (TG) is a domain-dependent information that shows the domain-specific goal of a communication. We study the structure and information of Hong Kong Tourism Board website and define seven task goals. Appendix C.1 shows the seven task goals with their definition and examples. Task goals, ASK_INFO and ASK_SUGGEST, are some kinds of catch-all task goals. When no other task goals can be fitted, ASK_INFO or ASK_SUGGEST will be our



http://www.discoverhongkong.com/eng/touring/hkwalks/ta_walk_walk2.jhtml

HONG KONG WALKS

Yau Ma Tei & Mong Kok - Markets for Leisure and Pleasure

For a non-stop experience of local, urban Chinese lifestyles, you need look no farther than Yau Ma Tei and Mong Kok - the heart of the Kowloon Peninsula. Within these two neighbourhoods are side streets and alleys that are home to one of Hong Kong's liveliest spectacles. You'll be amazed by the teeming masses jostling to and fro in this unique section of Hong Kong.

Walking through Yau Ma Tei and Mong Kok is an exciting and memorable experience any time of the day or night. You'll love the ambience and the great deals you can get on souvenirs, clothing, electronic goods and much more.

DISTRICT HIGHLIGHTS

Flower Market • Yuen Po Street Bird Garden
These beautiful blooms and dainty birds are sure to delight the senses.

Temple Street Night Market
Bargain hunting amidst mesmerising sounds is just part of an evening's fun.

Tin Hau Temple (Yau Ma Tei)
Enjoy a fascinating glimpse inside this treasure.

Figure 3.1: A screen capture of the Hong Kong Toursim Board website. Different media of information are shown on the same page.

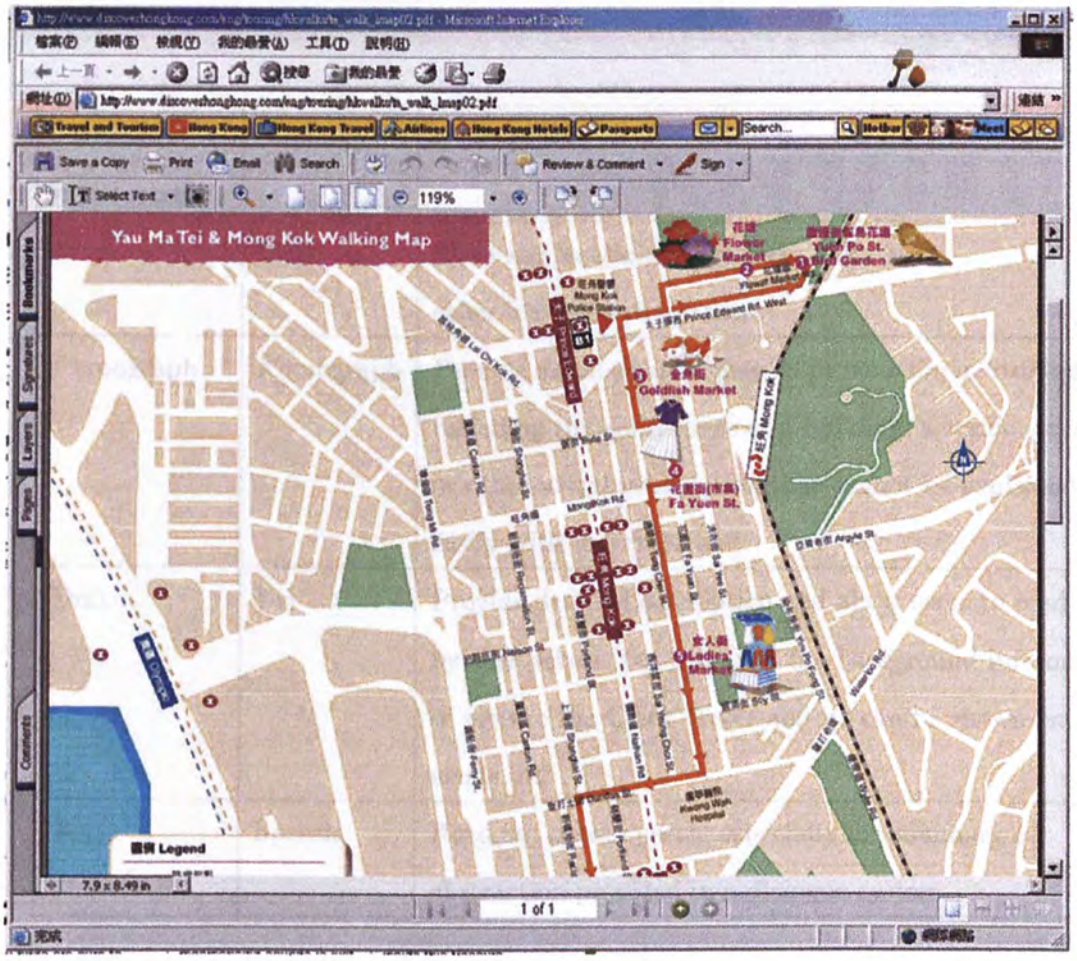


Figure 3.2: A new web browser contains an enlarged map. User can find the landmarks, markets, street names and recommended walking route on the map.

Associated multi-medium	Modality	Description
Map	Map	Show the location information, address or other locative information.
Photograph	Photograph	Provide an alternative may to present the information, including the appearance of a building, a scenic view or other visual information that cannot be explained easily with speech or text.
Text	Url	Provide detailed description and allow users to read it by themselves. For example, the background information of picture the Clock Tower and its can be shown on a single url.
Text	Speech	Read out using speech, especially for summary or abstraction provided in a dialogue system.

Table 3.1: The modalities required to present the tourist information domain multimedia information

final choice depending on the content.

Dialog act (DA) expresses the primary communicative intention of a communication. We have made reference to VERBMOBIL-2 [53] and Yip [54] to find out the dialog acts that are suitable for our domain. We have selected and adopted some of the dialog acts in them to our information domain. Some new dialog acts such as REQUEST_PREFERENCE and GIVE_PREFERENCE are introduced to fit our domain. Detailed definitions and examples (Table 3.2 of the task goals, user and system dialog acts are in Appendixes C, D and E. There are eighteen dialog acts for user requests and twenty dialog acts for system responses.

ASK_ATTRACTION	‘有，你可以去香港歷史博物館.’
ASK_FEE	‘你需要成人票還是小童票?’
ASK_INFO	‘香港有十八區.’
ROUTE_SEEKING	‘你想塔咩野類型的交通工具?’
ASK_SUGGEST	‘我建議你行山頂先．這是最短路線 ..’
ASK_TOUR	‘這是大嶼山旅行團的資料.’
RESERVATION	‘我幫你訂左 三張船飛啦.’

Table 3.2: Examples of seven task goals.

3.2.2 Concepts and Information Type

A task goal includes many concepts. Since the variation of concepts can be very large, we have only focused on one of the task goal, i.e. ASK_ATTRACTION. This is because ASK_ATTRACTION is the most frequently occurred task goal in the multi-modal dialog corpus collected. Table 3.3 shows the concepts and their possible values. We have manually annotated them in our collected corpus because we need to explore the possible words used and the variations of each concept.

LOCATE_DISTRICT	（位於）中西區
ATTRACTION_DESCRIPTION	地道的，最大的
BUILDING	禮賓府
AREA	時代廣場
VISIT_DAY	星期日

Table 3.3: Examples of concepts and their possible value in the ASK_ATTRACTION task goal.

We need another classification so as to help the modality selection. We focus on the format of content rather than the semantic meaning. We have classified the concepts into six information types [56] as shown in Table 3.4. Appendix F shows all relationships among different concepts and information types. Appendix G shows the concepts and their values.

3.3 User's Task and Scenario

Every user is requested to plan a three-day itinerary in Hong Kong. They can plan the itinerary according to their interests and the dialogue will end when the planning is done. The information in focus is the Hong Kong Tourism Board website — DiscoverHongKong. Table 3.5 shows a segment of the dialogue between user and system. The example shows the necessary usage of multi-modal response (in square bracket). Table 3.5 illustrates how user and system use multi-modal input and response respectively. System used a map to indicate the location of four open markets (System 1) and then pointing to the suggested route on map rather than described it in speech (System 2). User pointed out some interested attractions in User 3 and the system highlighted the two museums that are open at night as response (System 4).

Information type	Definition	Examples
Locative	Place of attraction	Address, location of attraction
Descriptive	Description of attraction	History of a historic attraction
Physical object	A tangible and visible object	Building, Museum
Abstract	A concept not associated with any specific instance	Class label of attraction
Nominal	Named entity, which is not belong to other information type	A scenic area
Temporal	Time related information	Opening hours

Table 3.4: The definitions and examples of information types.

All dialogue turns in Table 3.5 include the use of speech (i.e. contain spoken content). Therefore, we believe that speech is a major modality in dialogue. The commonly use of speech also shows that it is important for us to increase the user perception on speech information. Possible information related includes lip-movement, expression and head movements. We try to use a talking head system to provide lip-movement and to increase the user perception. Details of the talking head system will be discussed in Chapter 5.

3.4 Chapter Summary

We have described the multi-media information provided our information domain, i.e. the Hong Kong Tourism Board website. The rich multi-media information provided leads to the necessary of multi-modal response. We show the definitions and examples on task goals, dialog acts and concepts. The con-

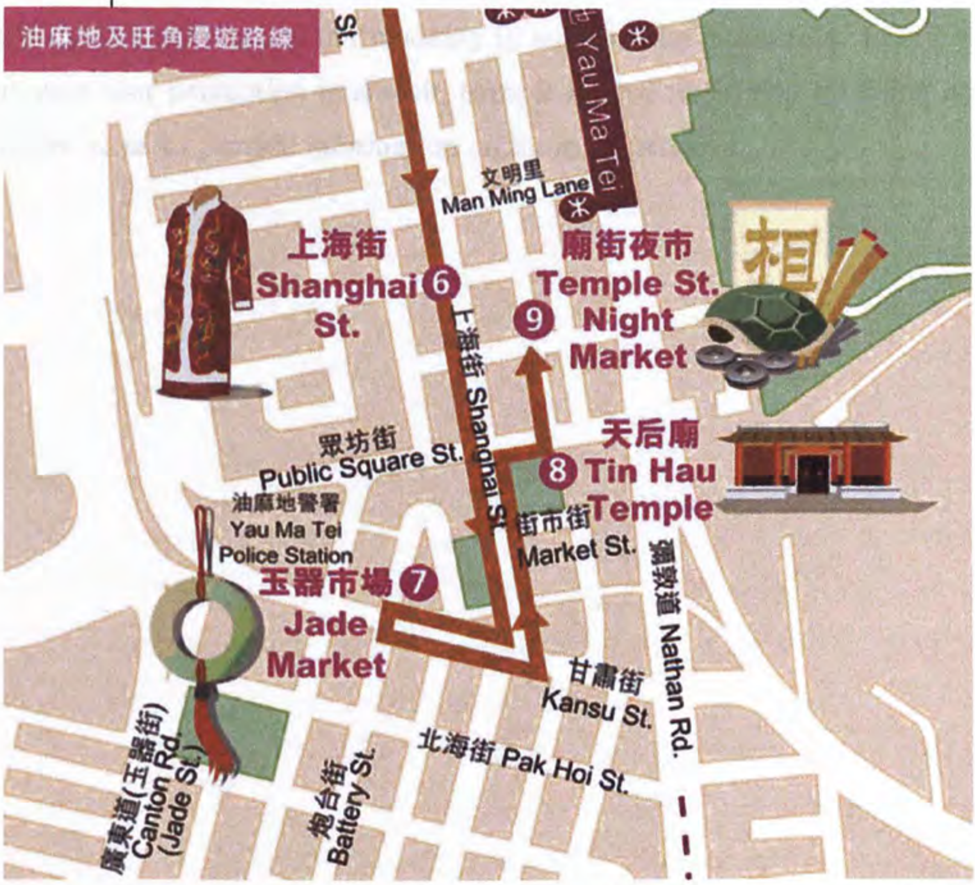
User 1:	我可以係邊度搵到露天市場(open market)?
System 1:	這裡顯示了旺角(Mongkok)的多個露天市場 [the map of Yau Ma Tei and Mongkok]).
	 油麻地及旺角漫遊路線 Yau Ma Tei 文明里 Man Ming Lane 上海街 Shanghai St. 6 廟街夜市 Temple St. Night Market 9 上海街 Shanghai St. 眾坊街 Public Square St. 天后廟 Tin Hau Temple 8 油麻地警署 Yau Ma Tei Police Station 街市街 Market St. 玉器市場 Jade Market 7 彌敦道 Nathan Rd. 甘肅街 Kansu St. 北海街 Pak Hoi St. 廣東道(玉器街) Canton Rd. (Jade St.) 炮台街 Battery St. 相
User 2:	我可以點先順路行哂全部?
System 2:	你可以沿著地圖上的橙色線走 [point to the route on the map].
User 3:	我想下晝先去喱度[point to 廟街 (the Temple Market)] 然後再去喱度 [point to 玉器市場 (the Jade Market)]
System 3:	玉器市場只開到六點鐘．不如你先去玉器市場跟住去廟街
User 4:	行完都夜啦，但我仲想去博物館睇下
System 4:	[Highlight two museum names on screen] 這兩間博物館夜晚都開放

Table 3.5: A segment of scenario collected, in which user tries to plan a day of itinerary in Hong Kong.

(Figure below System 1) Part of the Yau Ma Tei and Mongkok map showed in System 1. The orange route is a suggested walking path provided by the website.

cepts are further grouped into information type so as to help the process the modality selection. A scenario of the tourist information domain are shown to show the use of multi-modalities in user requests and system responses. The scenario also shows the effective use of multiple modalites. The scenario indicated that speech is a major modality in multi-modal interaction. In order to increase user perception in speech, there is a need to develop a talking head system so as to provide information on lip-movements.

Chapter 4

Multi-modal Response Data Collection

Appropriate use of multiple modalities can result in effective presentation. The use of multiple modalities can increase conciseness, be more informative and easier to follow. For example, address can be presented with speech and pointing gesture. A suggestion complemented with a photo can make the message more informative. Detailed description can be presented in a URL so that user can read it in the way he/she likes. We need to use different modalities to fit the requirement of different responses. Therefore, we need to understand human behavior on multi-modal responses so as to select suitable modalities. In order to achieve the goal, we perform a multi-modal dialogue data collection and analysis.

M3 stands for mobile, multi-biometric and multi-modal. The M3 corpus is a collection of mobile, multi-biometric and multi-modal data which is used to support the research in multi-biometrics and multi-modal human-computer communication systems. The M3 corpus is consisted of several components: multi-modal dialogue data, biometrics data (e.g. fingerprints, passport photo and personal information), audio-visual data and speech data. Since this thesis is focused on the development of multi-modal response generation, we focus on the multi-modal dialogue data in this chapter.

This chapter presents the way we collect the multi-modal responses so as to explore how human use different modalities and their behavior in natural human-computer communication. We used the Wizard-of-Oz (WOZ) setup to collect multi-modal dialogues on tourist information domain in Hong Kong. A data optimization process is then performed. An expert is found to design the collected responses into multi-modal responses according to the collected dialogue structure. The offline design phase allowed the expert to plan an effective multi-modal response.

We have designed a set of annotation — M3 Markup Language (M3ML) to annotate the collected multi-modal responses. We have made reference to the EMMA: Extensible MultiModal Annotation markup language [15] for the design process. Design of M3ML is based on the EMMA structure, attributes and values. We have also extended EMMA to support the definition of spatial coordination.

We observe the usage patterns including the possible modalities and their combinations. We have also analyzed their relationships with intention (i.e. task goal and dialog act) and information types (e.g. locative information) so as to find out the modality preference.

4.1 Data Collection Setup

The multi-modal dialogue data collection is used to analyze how users and wizards use different modalities so as to maximum their expressive power. For example, user can point or circle a location on map instead of speaking the location name out. We applied a methodology for data collection which is known as the Wizard-Of-Oz (WOZ) setup. WOZ setup means a person (the 'wizard') plays part or whole role of the system we target to develop. Users interact with the user interface of a program which has partially or not yet been developed. The wizard controls the output of the user interface in the same way of a system. Users can talk to the wizard at any time and any way they like. Users feel that they are playing with a real system.

We setup a WOZ setup in a sound-proof recording room for the multi-modal dialogue data collection. User communicates with a simulated system (the human wizard) through instant messaging software. Instant messaging is the ability to exchange messages and share applications with a chosen friend who is connected to the Internet. The wizard can see the user's screen through the instant messaging software so he/she knows the user's current focus. The interaction is video-taped so as to support our observation on the usage patterns of different modalities. Details of the collection will be discussed in the following sections.

4.1.1 Multi-modal Input Setup

We have put many different devices in both the user's side and the wizard's side so as to maximize user's and wizard's expressive powers. Availability of different devices is used to support different modalities. Table 4.1 summarizes the supporting modalities, devices and their descriptions. User is free to switch among different modalities whenever he/she thinks it is suitable. We show the setup and close up of the screen on the user side in Figures 4.1 and 4.2.

4.1.2 Multi-modal Output Setup

Since the wizard is omniscient computer, the setup and the devices provided are slightly different from the user side. Table 4.2 lists the devices provided and the corresponding supporting modality on the output side. A shared browser allows wizard to monitor the user action on screen. Since we do not want that user know the super-computer is simulated by human, we do not put a

Modality	Device(s)	Description
Speech	Microphone	Deliver user's speech request to system
Text	MSN ® Messenger message box [1]	Deliver user's textual request to system
Emoticons	MSN ® Messenger message box	Express his/her emotion with emoticons
Pointing	Mouse	Point on an image or a map location for indication
Circling	Mouse	Move the mouse around an image or a map location for indication
Highlight	Mouse	Select some textual information
Hypertext	Mouse	Click and go to a web page to indicate his/her own interest preference
Gesture	Web cam	Express his/herself with body motion or facial expression

Table 4.1: Input modalities and related devices provided for the input side.

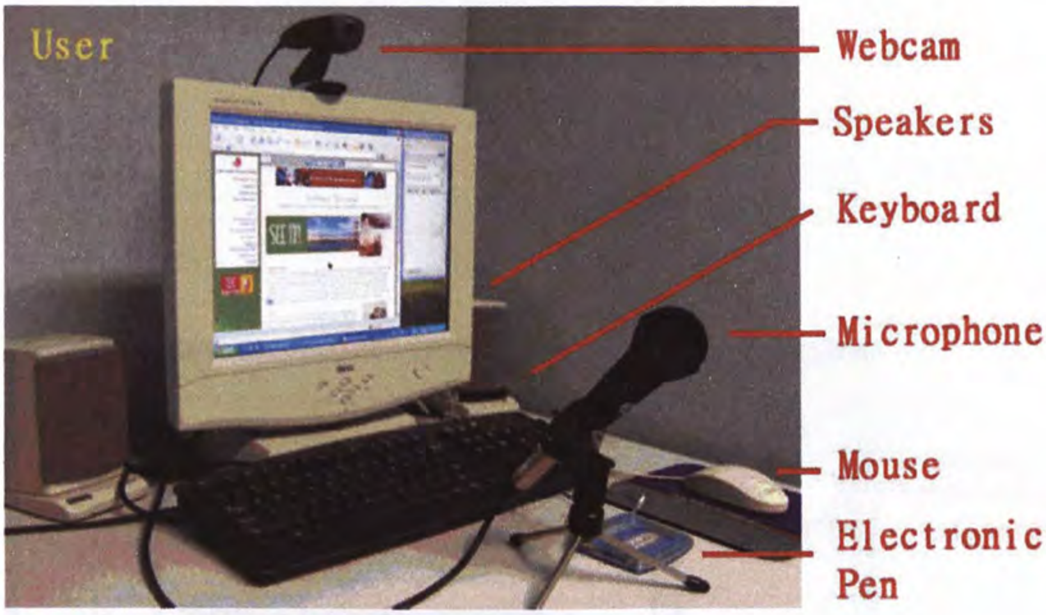


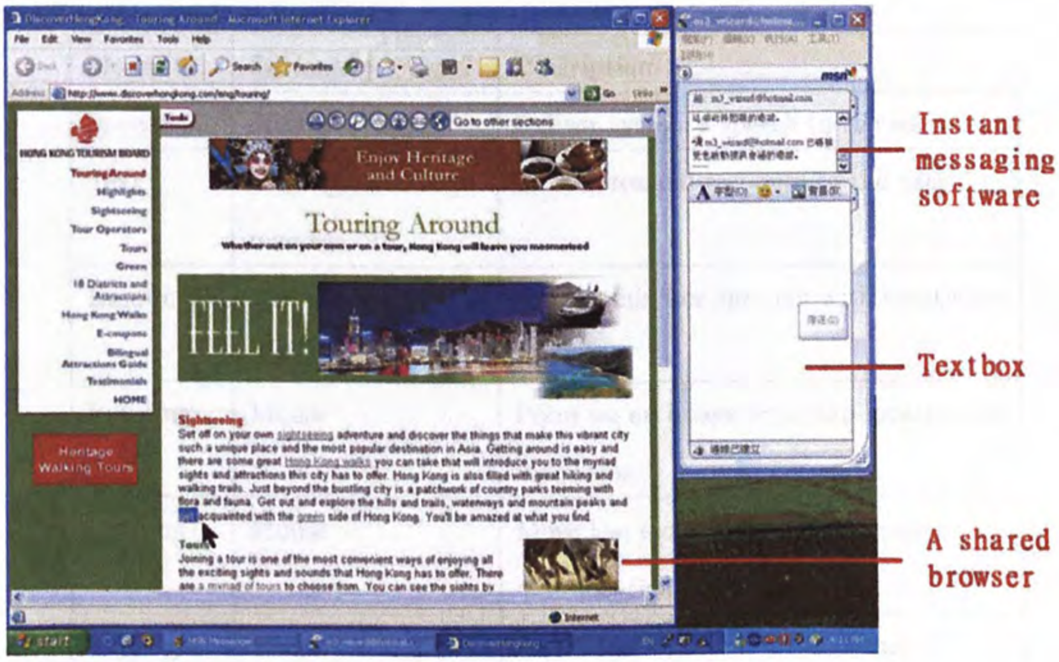
Figure 4.1: WOZ setup of the user side in the recording room for data collection.

web cam on the wizard side. Figures 4.3 and 4.4 show the setup and close up of the screen on the wizard side in the same room. Due to the limitation of physical space, user and wizard are separated with a notice board in the same room. The separation is used to prevent users to discover that they are communicating with a human (i.e human-human interaction) instead of a system (i.e. human-computer interaction).

4.2 Procedure

4.2.1 Precaution

Before the data collection starting, a short briefing session is given to each of the users. They are briefed with tasks that they requested to do and the information domain focused in the data collection. Users are requested to be a tourist who will visit Hong Kong for three days. They need to plan a three-day itinerary according to his/her interests. The 'system' which can reply to



Instant messaging software
Textbox
A shared browser
DiscoverHongKong - Touring Around - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites
Address http://www.discoverhongkong.com/eng/touring/
Go
Go to other sections
HONG KONG TOURISM BOARD
Touring Around
Highlights
Sightseeing
Tour Operators
Tours
Green
18 Districts and Attractions
Hong Kong Walks
E-coupons
Bilingual Attractions Guide
Testimonials
HOME
Heritage Walking Tours
Enjoy Heritage and Culture
FEEL IT!
Internet
start

Touring Around

Whether out on your own or on a tour, Hong Kong will leave you mesmerised

Sightseeing

Set off on your own sightseeing adventure and discover the things that make this vibrant city such a unique place and the most popular destination in Asia. Getting around is easy and there are some great Hong Kong walks you can take that will introduce you to the myriad sights and attractions this city has to offer. Hong Kong is also filled with great hiking and walking trails. Just beyond the bustling city is a patchwork of country parks teeming with flora and fauna. Get out and explore the hills and trails, waterways and mountain peaks and get acquainted with the green side of Hong Kong. You'll be amazed at what you find.

Tours

Joining a tour is one of the most convenient ways of enjoying all the exciting sights and sounds that Hong Kong has to offer. There are a myriad of tours to choose from. You can see the sights by

Figure 4.2: A capture of the user's screen. Wizard can see user's action on the shared browser and show information to the user using the same browser. User and wizard can communicate on the text box provided by the instant messaging software.

Modality	Device(s)	Description
Speech	Speaker	Deliver system's speech to the user
Text	MSN ® Messenger message box	Deliver textual response to the user
Emoticons	MSN ® Messenger message box	Express his/her emotion with emoticons
Pointing	Mouse	Point on an image or a map location for indication
Circling	Mouse	Move the mouse around an image or a map location for indication
Highlight	Mouse	Select some textual information
Hypertext	Web browser	Share a web browser, show a web page or send an URL to user

Table 4.2: Output modalities and the corresponding devices provided on the wizard side.

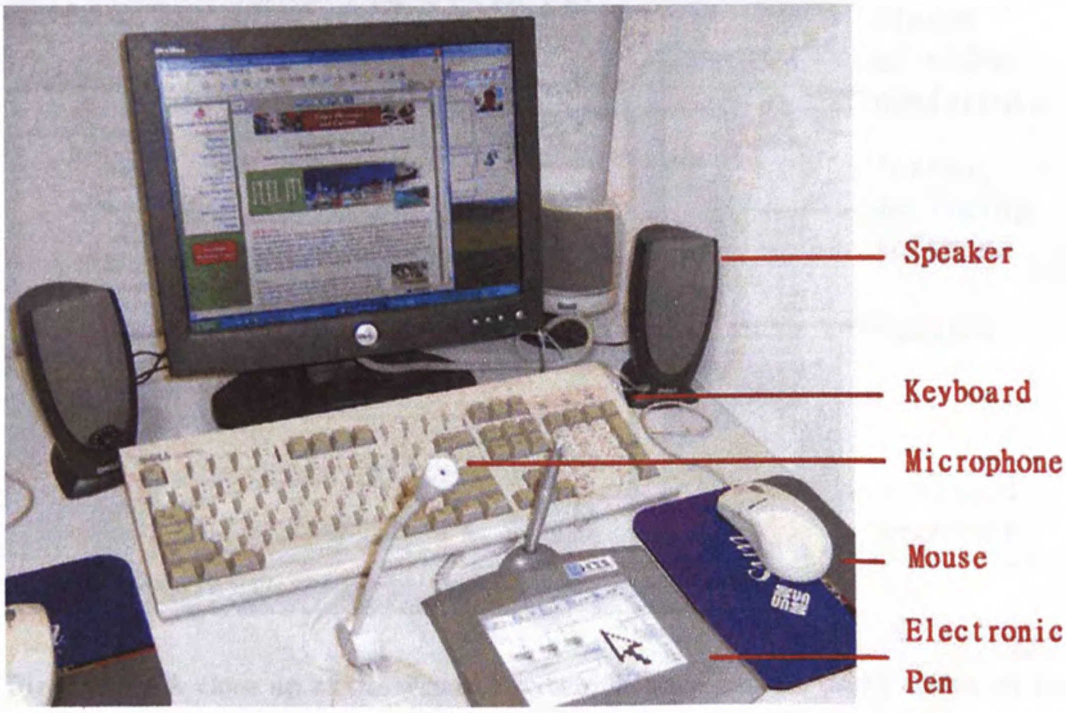
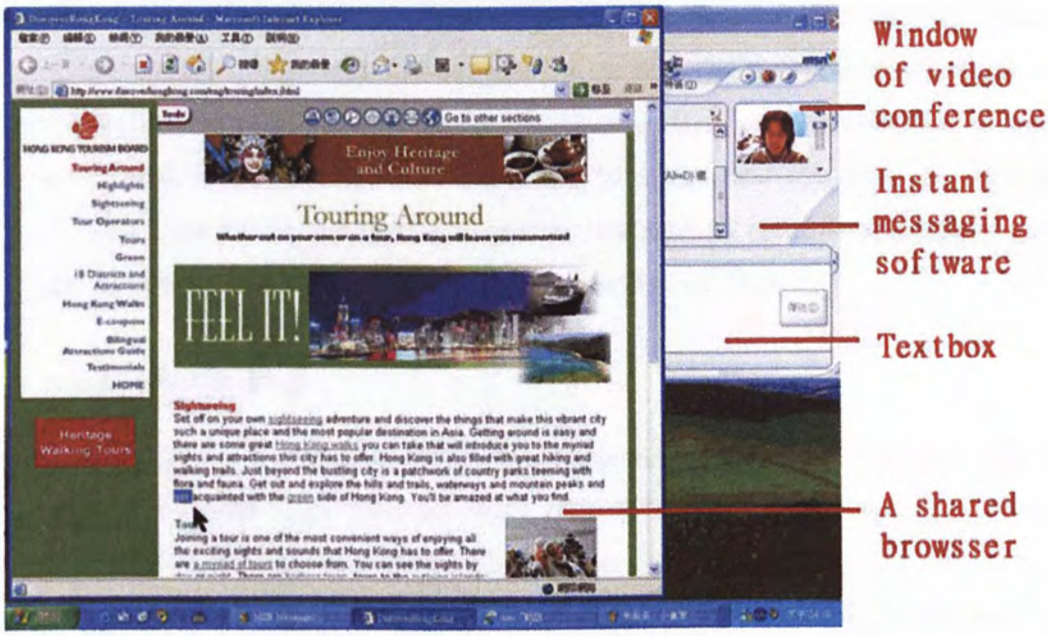


Figure 4.3: WOZ setup of the wizard side in a recording room for data collection.



Sightseeing

Set off on your own sightseeing adventure and discover the things that make this vibrant city such a unique place and the most popular destination in Asia. Getting around is easy and there are some great Hong Kong walks you can take that will introduce you to the myriad sights and attractions this city has to offer. Hong Kong is also filled with great hiking and walking trails. Just beyond the bustling city is a patchwork of country parks teeming with flora and fauna. Get out and explore the hills and trails, waterways and mountain peaks and get acquainted with the green side of Hong Kong. You'll be amazed at what you find.

Tours

Joining a tour is one of the most convenient ways of enjoying all the exciting sights and sounds that Hong Kong has to offer. There are a myriad of tours to choose from. You can see the sights by

Figure 4.4: A close up of the wizard's screen. Wizard can see user's action on the shared browser. Wizard can also see user's facial expression and gesture with the small screen of camera in the top left corner.

the user with the knowledge obtained from the Hong Kong Tourism Board website only. Any information outside the website will be considered as not available. A self-browsing period will be given to all the users so that they can have a basic idea on the information provided and think of some criteria for their inquiries before the process start.

A demonstration of the use of different devices and modalities will also be given so as to ensure users and wizards understand the controls and functions of the simulated system. For example, a control right should be obtained by the wizard from the user's shared browser so that user can see the actions performed by the wizard. Wizard needs to release the control right back to the users after each response so users can continue to make their inquiries and let the wizard see their actions on the same shared browser.

4.2.2 Recording

Since we do not have real time gesture capturing softwares and speech decoding system, we have used a digital video camera to video-taped the interactions. We have conntected user's computer to an external monitor (therefore, two monitors in total for user's computer) so as to minimize the disturbance to users. We have put a digital video camera that is facing to the external monitor of the user's machine. We also put another web cam in front of the users to capture their facial expressions during communication. Dialogues in video tapes are then digitalized into MPEG-2 format for later reviewing. Figure 4.5 shows the data collection setup.

4.2.3 Data Size and Type

There are fifty-seven subjects who have participated in the data collection in total. Each subject recorded one dialog to three dialogs (Table 4.3). Each record includes the video for user's screen action, audio from user request and

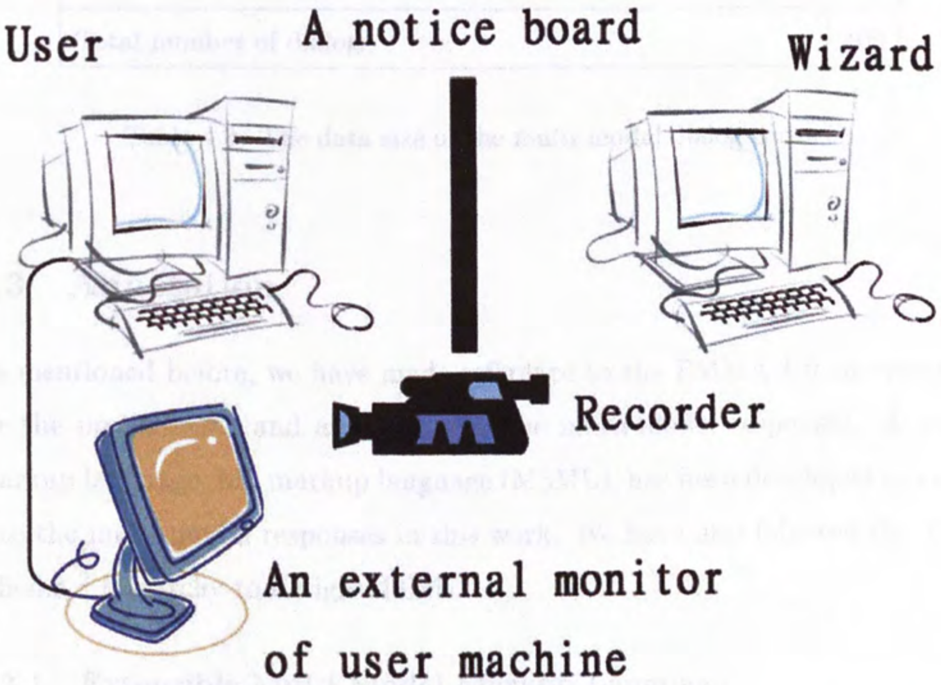


Figure 4.5: The data collection setup in a sound-proof recording room.

wizard response. The wizard response action also is included into the video for user's screen action.

Number of subjects who have recorded exactly one dialog	23
Number of subjects who have recorded exactly two dialogs	25
Number of subjects who have recorded exactly three dialogs	9
Total number of subjects	57
Total number of dialogs	100

Table 4.3: The data size of the multi-modal dialog corpus.

4.3 Annotation

As mentioned before, we have made reference to the EMMA 1.0 specification for the presentation and annotation of the multi-modal responses. A set of markup language, M3 markup language (M3ML), has been developed to annotate the multi-modal responses in this work. We have also followed the XML schema [2] hierarchy to design M3ML.

4.3.1 Extensible Multi-Modal Markup Language

EMMA annotates a request/response with many interpretation elements. It defines interpretation as "*The emma:interpretation element holds a single interpretation represented in application specific markup.*". There are three relationships defined among interpretation and two of them (*group* and *sequence*) are related to the response side. *Group* is defined as "*the emma:group element is used to indicate that the contained interpretations are related in some manner*" (source: [15]) in EMMA. *Sequence* is defined as "*the emma:sequence is a*

[2] http:/W3C XML Schema, http://www.w3.org/XML/Schema

container for one or more interpretation or interpretation container elements and denotes that these are sequential in time" (source: [15]).

Time information is one of the important informatioon in multi-modal dialogue system. EMMA introduces a set of time attributes. The start time attribute can be absolute time or relative time value. The relative time can be used to indicate the temporal coordination between two interpretations. For example, the start time of one interpretation can be annotated on how many milliseconds after the start time of another interpretation. All the time information is required to store in a very high precision. For example, the value of duration attribute must be stored in the unit of milliseconds. Figure 4.6 shows the set of time attributes in red boxes.

4.3.2 Mobile, Multi-biometric and Multi-modal Annotation

Multi-modal response generation includes temporal and spatial coordinations. Since EMMA only provides annotation on time information, we have extended and applied the idea of time attributes (i.e. annotation of time information) to spatial information in M3ML. We have defined a set of relative coordinates attributes. For example, a point was shown in the centre of map to indicate the location of a building. The *coordinate-ref-uri* attribute indicating the coordinates of pointing action is corresponded (in relative coordinates) to the map coordinates and the pointing coordinates are defined by *x-offset* and *y-offset*. This is shown in blue boxes in Figure 4.7, which is in the presentation format of XML schema.

The high precision (in millisecond) cannot be easily achieved in most of the situations. For example, a typical digital video recorder has only around twenty-five to thirty frames per second which means the precision is around 0.03 per frame only. We have defined a set of semantic values in order to substitute the precision defined in EMMA. We can annotate the action time

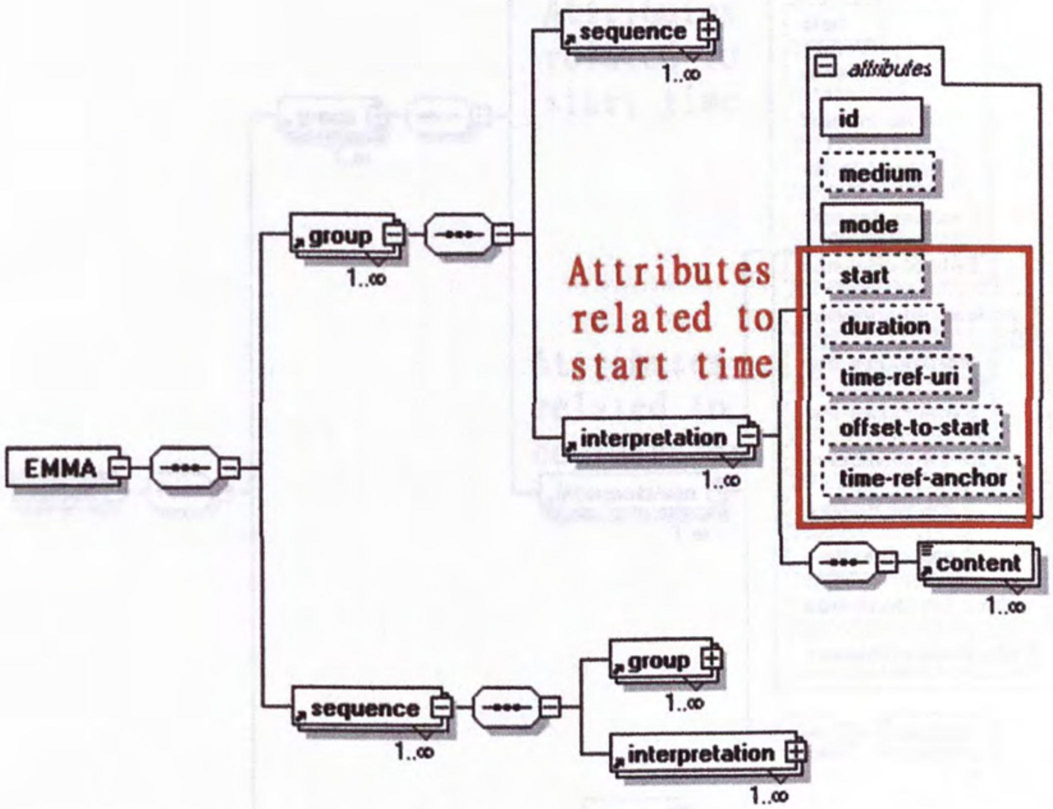


Figure 4.6: Extracted EMMA schema that we have referred to. The red square highlights the attributes related to timestamp.

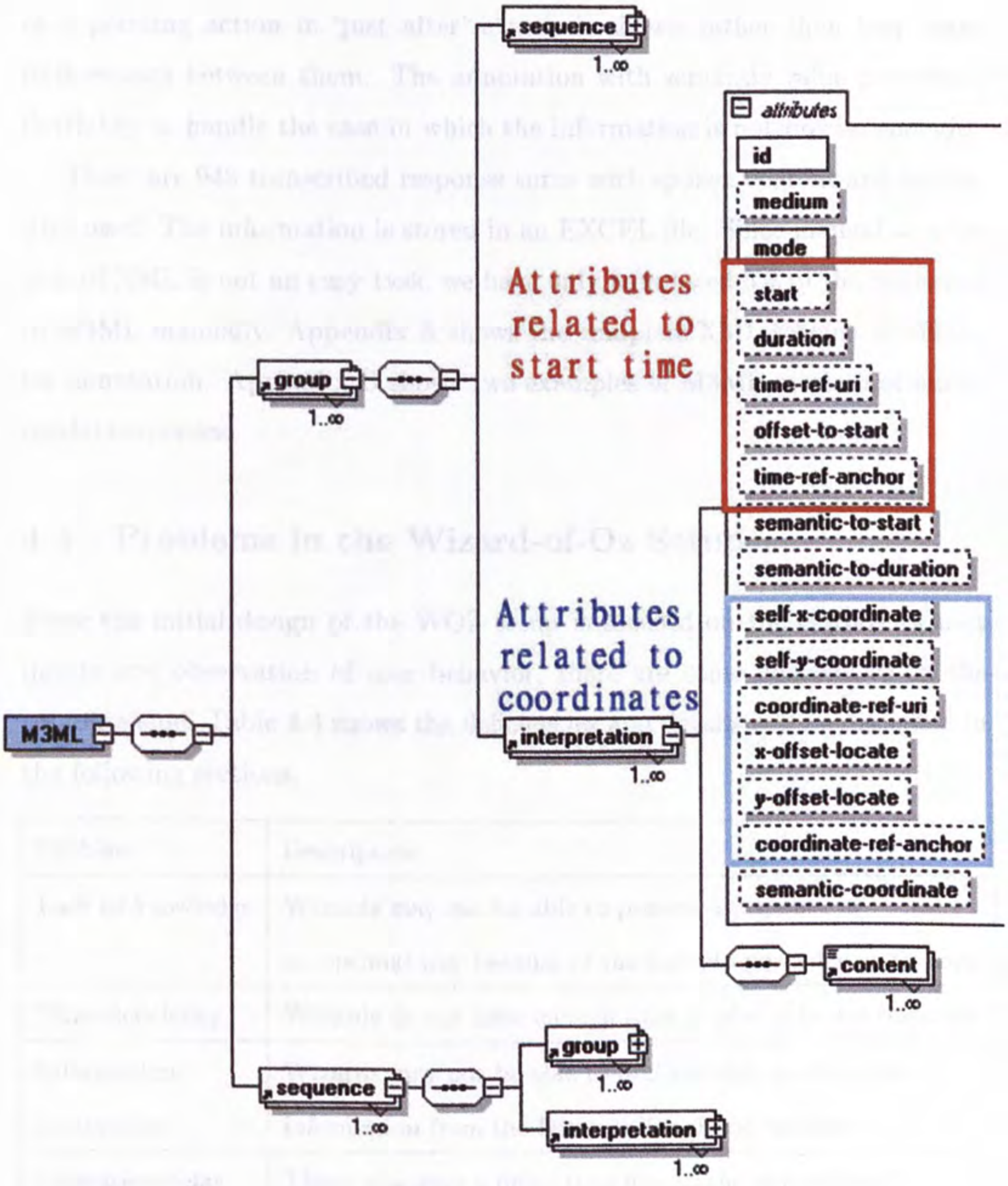


Figure 4.7: M3ML schema. The red box highlights the attributes related to timestamp that are borrowed from EMMA. We adopted the idea and add a set of attributes to spatial information as highlighted in blue box.

of a pointing action in 'just after' a map is shown rather than how many milliseconds between them. The annotation with semantic value provides a flexibility to handle the case in which the information is not precise enough.

There are 948 transcribed response turns with spoken content and modalities used. The information is stored in an EXCEL file. Since manual annotation of XML is not an easy task, we have only annotated 5% of the responses in M3ML manually. Appendix A shows the complete XML schema of M3ML for annotation. Appendix B shows two examples of M3ML markup of multimodal responses.

4.4 Problems in the Wizard-of-Oz Setup

Since the initial design of the WOZ setup is focused on the capture of user inputs and observation of user behavior, there are some deficiencies on the response side. Table 4.4 shows the deficiencies and details will be discussed in the following sections.

Problem	Description
Lack of knowledge	Wizards may not be able to present a response in an optimal way because of the lack of knowledge or training
Time deficiency	Wizards do not have enough time to plan a better response
Information availability	Wizards may not be able to find suitable multi-media information from the DiscoverHongKong website
Operation delay	There is always a delay time due to the operation of user interface or loading of information
Lack of modalities	Wizards cannot use their facial expression due to the lack of web cam on the wizard side

Table 4.4: Examples on the deficiency of the WOZ setup in response generation.

4.4.1 Lack of Knowledge

In the multi-modal dialogue collection, wizards are students from the Faculty of Engineering who is not an expert in communication. Table 4.5 shows some multi-modal responses from multi-modal dialogue data collection which have room for improvements.

User 1:	請問中環有咩地方值得去睇？
System 1:	不如去蘇豪 (Soho, an attraction in Hong Kong).
User 2:	有無資料介紹？
System 2:	蘇豪係一個新的美食區，包括些利街，伊近利街，卑利街，士丹頓街及奧卑利街一帶，小街道之間聚集了不少具有異國風情的特色餐廳，咖啡室及酒吧.

Table 4.5: An example on response which needs improvement. For example, we can provide more information on an attraction to the user in System 1 by showing its picture. Wizard should provide a URL with the introduction of the attraction rather than speak the description out. This can prevent the loss of spoken information during delivery.

4.4.2 Time Deficiency

The interactions are performed in real-time during the multi-modal dialogue data collection process. The response planning time is limited in real-time situation. In a wizard turn, the wizard is not only concerned with his/her tasks as shown in Table 4.6 but also observe the user's facial expression. If the user is impatience, wizard needs to provide some feedback before the completion of the content retrieval and response planning. Wizard has to finish a few tasks in a limited time because user does not want to wait so long. As a result, wizard may not be able to plan the best multi-modal response.

Task	Description
Understand the request	Wizard needs to know the content of the request
Resolve any deictic term	When user uses multi-modal input, some deictic term (e.g.‘喱度’) may be used. It is referred to another modality and need to be resolved.
Information retrieval	Wizard retrieves textual and multi-media information (e.g. map) from the Internet. Retrieved information is used for the planning and output of a multi-modal response.
Response planning	Wizard plans the way to present the retrieved information.

Table 4.6: The tasks of a wizard in a dialogue turn.

4.4.3 Information Availability

Wizard always needs to retrieve multi-media information from the DiscoverHongKong website. The website provides a wide range of domain information, including tour, attraction, activity, transportation, event, time, fare, etc. However, the website cannot cover all kinds of multi-media information in the tourist information domain. Sometimes, the wizard has to retrieve information from the Internet using search engine. It takes some time to finish the retrieval task. The lack of multi-media information forces the wizard to use uni-modal response although multi-modal response is better in some cases. Table 4.7 shows a sample interaction with speech only. Figure 4.8 shows the same interaction with the support of multi-media information (e.g. photo).

User 1:	係海洋公園裏面有邊個關於海洋的好地方呀?
System 1:	你可以去太平洋海岸參觀?
User 2:	太平洋海岸係點架?
System 2:	太平洋海岸好似加洲海岸，有很多海豹海獅睇.

Table 4.7: An example of uni-modal response. In System 2, the use of a photo would be better. However, the lack of the relevant photo in the DiscoverHongKong website forces the wizard to use speech only.

4.4.4 Operation Delay

The WOZ setup also limits the behaviour of wizard. The instant messaging software is a powerful tool for multi-modal expression. Wizard needs extra work to operate such a powerful tool. For example, when wizard needs to show a URL, he/she needs to input the URL to the web browser. When wizard wants to show another URL, a series of clicking actions on hyper-links or type in the new URL is needed. Both kinds of operation affect the fluency of a multi-modal response. Whenever a website is opened, the long loading time will also cause a delay in the response.

4.4.5 Lack of Modalities

Since the human wizard is hidden to simulate a super-computer, we did not put a web cam on the wizard side. The wizard cannot use his/her facial expression as one of the response modaliies as we did not put a webcam on the wizard's side.

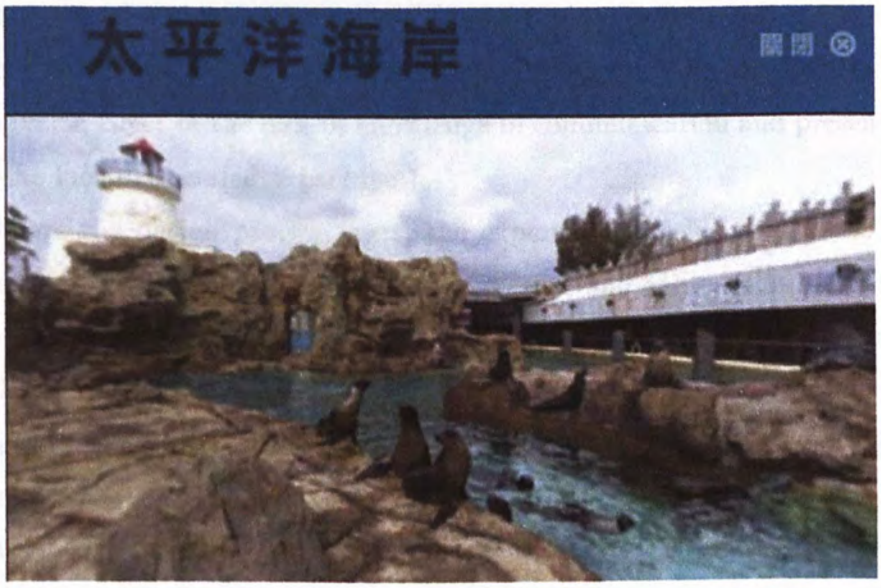
User 2:	太平洋海岸係點架?
System 2:	這就是太平洋海岸. [Show a picture]
	

Table 4.8: An example of multi-modal response for Table 4.7. Wizard uses speech and photograph to introduce the attraction.

(Figure) The picture showed in the turn of System 2. This photograph is found from the Ocean Park website through a series of searching steps. It is out of our focused website and requires some time to retrieve. Wizard always does not have enough time to retrieve related photograph.

4.5 Data Optimization

We perform data optimization of the multi-modal responses by an offline designing process. We designed the multi-modal response based on the response content extracted from the collected multi-modal dialogue.

4.5.1 Precaution

We have hired an expert from the School of Journalism and Communication to help us to design the responses with multiple modalities. This can reduce the adverse effect of the lack of knowledge in communication and presentation (i.e. the lack of knowledge problem).

Before the design process start, the expert was briefed with some background information on multi-modal interaction and visited the multi-modal dialogue data collection. We put emphasis on natural interaction rather than simulation of a 'real' system response. We encouraged the expert to fully utilize the multi-modal information available including multi-media, facial expression, vocal expression, etc.

We extract response content from the responses collected by the WOZ setup. The expert studied each dialogue and response content extracted and then design the best response. The expert is provided with the extracted content instead of the original response. This is used to prevent from the learning effect of the presentation, which means the presentation wordings and styles from the expert may be affected by the collected one. Table 4.9 shows an example of the information provided to the expert.

4.5.2 Procedures

Optimization is an offline process, which provides the expert with enough planning time to retrieve information and plan the response. In the design, we did not request the expert to retrieve the actual multi-media file. For example, the

User 1:	請問中環有咩地方值得去睇？
System 1:	[Attraction suggestion: 蘇豪]
User 2:	蘇豪係點架?
System 2:	[Attraction description: 一個新的美食區，包括些利街，伊近利街，卑利街，士丹頓街及奧卑利街一帶，小街道之間聚集了不少具有異國餐廳的特色餐廳，咖啡室及酒吧.]

Table 4.9: The information provided to the expert. The expert can use any modality to present the information. For example, the expert can use photo and pointing action to present the attraction in System 1. The expert can also improve the response in System 2 by searching and showing a URL. Certainly, the expert cannot suggest another attraction. Otherwise, the dialogue structure will be changed.

floor-plan of a shopping mall cannot be found on the DiscoverHongKong website. When the expert needs to use the floor-plan for a multi-modal response of locative information, the expert can assume the floor-plan is exist. Therefore, the process can be prevented from the problem of information availability mentioned before. Certainly, the expert needs to specify what is the content he/she expected for record.

After the offline design phase, we simulated and recorded the designed interaction. There was another person who acts as 'a simulated user'. The expert faced to the user so that the expert was able to use facial expressions such as a puzzling face to indicate unclear user request or she is unable to satisfy the user request. We wanted to minimize any adverse effect on presentation fluency due to the operation delay. Therefore, the expert sketched any graphic on paper and point on it if necessary. The approach could improve the fluency of response and allow the expert to focus on his/her presentation rather to control a software. A digital video recorder was used to take a record for the expert presentation. Figure 4.8 shows the recording setup.

4.5.3 Data Size in Expert Design Responses

We randomly selected 31 dialogues and re-designed them by an expert (Table 4.10). There are 948 response turns in total. Table 4.11 shows some information about the expert designed responses.

Total number of dialogs	31
Total number of subects	26
Number of subjects have exactly one dialog	21
Number of subjects have exactly two dialogs	5

Table 4.10: The data size of re-design multi-modal dialog corpus.

Number of response turns with task goal:	
ASK_ATTRACTION	333
ROUTE_SEEKING	256
ASK_INFO	162
ASK_SUGGEST	82
ASK_TOUR	49
ASK_FEE	42
RESERVATION	24
Total	948

Table 4.11: The distribution of different task goals in the expert designed responses.

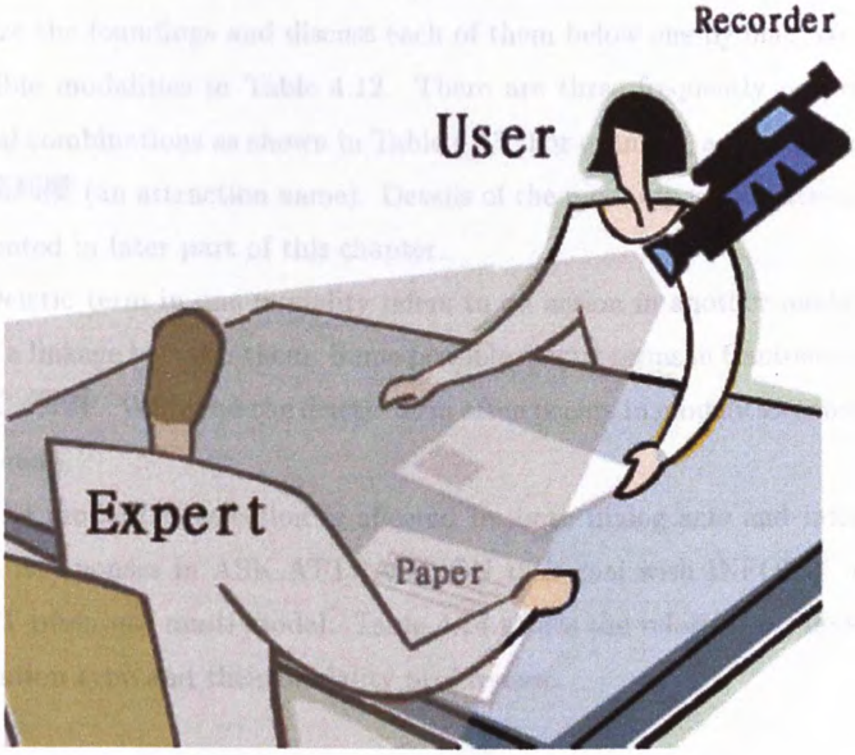


Figure 4.8: The recording setup for designed multi-modal response. The expert faces the simulated user so as to use facial expression. The expert uses multi-modal response with paper so he/she can focus on presentation rather than software operation. For example, the expert presents a locative information by sketching a map very roughly on a paper and point on the sketched map. A video recorder records all actions for the expert for further study on the temporal relationship and other.

4.6 Analysis and Discussion

We have analyzed the expert designed responses on the modality used, modality combination, deictic term, dialog act and information type. We first summarize the foundings and discuss each of them below one by one. We list the possible modalities in Table 4.12. There are three frequently occurs multimodal combinations as shown in Table 4.13. For example, a concept-value can be 美利樓 (an attraction name). Details of the modality combinations will be presented in later part of this chapter.

Deictic term in one modality refers to an action in another modality and form a linkage between them. Some possible deictic terms in Cantonese include ‘哩度’ ‘哩個’. We found the deictic term often occurs in modalities combination with map.

The modalities selection is affected by both dialog acts and information type. Responses in ASK_ATTRACTION task goal with INFORM or SUGGEST often use multi-modal. Table 4.14 shows the relationship between information type and their modality preferences.

4.6.1 Multi-modal Usage

We study the occurrence of modalities that have been used in the response of tourist information domain. Table 4.15 shows the distribution of multi-modal response and uni-modal responses. Uni-modal response includes the use of speech only. They include greeting (e.g. ‘welcome’), agreement (e.g. ‘Yes’, ‘OK’), close (e.g.‘Bye’), etc.

4.6.2 Modality Combination

After study the availability of different modalities, we have to study the use of multiple modalities in a response. Different combinations of modalities can be used for different concept-values and a response may include multiple concept-

Modality	Description
Speech	Read out using speech, especially for summary or abstraction provided in a dialogue system.
Text	Spoken content can be shown on text box so user can read if he/she misses the speech.
Facial animation	The synchronized lip movement with speech can increase the user perception. The combination of facial animation and speech is a kind of audio-visual speech.
URL	Provide detailed description and allow users to read it by themselves. For example, the background information of the Clock Tower can describe with an URL.
Photograph	Provide an alternative may to present the information, including the appearance of a building, a scenic view or other visual information that cannot be explained easily with speech or text.
Map	Show the location information, address or other locative information.
Pointing	Pointing is a kind of mouse gesture and always be used with other modalities. Wizard used pointing to indicate some locations on screen.

Modality	Description
Circling	Circling is another type of mouse gestures. It is similar to pointing but circling is always used to indicate an area on screen. There are other possible mouse gestures like stroke which cannot be found from the collected responses. We mainly discover pointing and circling in the collected response.
Highlight	Textual information on a URL can be highlighted so as to emphasize the content or indicate the focus of a URL.
Blinking dot	Blinking dot is used to capture the users' attention to a point on screen.
Text	Text is used to show some

Table 4.12: The possible modalities in WOZ.

Modality Combination	Percentage
Speech and photo	31.3%
Speech and URL	21%
Speech, map and pointing	19.5%
Total number of concept-value using multi-modal:	486 times

Table 4.13: The top three most preferable modality combinations in multi-modal responses. Speech is always come with facial animation so we skip in the table.

Information Type	Modality Preference
Locative information	Map
Descriptive information	URL
Physical object information	Photo

Table 4.14: The modality preference of three information types. The other information types prefer to use speech only.

Total number of response turns:	948
Uni-modal responses:	533 (56.2%)
Multi-modal responses:	415 (43.8%)

Table 4.15: The distribution of uni-modal and multi-modal responses designed by the expert.

values. For example, wizard can suggest two attractions in one response (i.e. two attraction concept-values) or suggest an attraction, its location and opening hours (one attraction, one location and one opening hours concept-values).

From the collected response turns, the expert presented 486 concept-values in multi-modal. The rates of occurrences of different multi-modal combination within 486 concept-values are shown in Table 4.16. Since speech is always present as one of the modalities in corpus, there are 128 possible combinations. Table 4.16 only show the combinations that can be found in our selected response data.

Although there are many possible modality combinations, several combinations are more preferred by the expert. Speech with URL or photo, speech with map and pointing (italicized in Table 4.16) are the top three most preferred modalities combinations.

Multi-modal combinations			
Total number of concept-values using multi-modal: 486			
	URL	Photo	Map
Speech	*21.0%*	*31.3%*	5.6%
Speech, pointing	9.7%	0.8%	*19.5%*
Speech, circling	2.3%	0.6%	2.5%
Speech, highlight	5.1%	0%	0%
Speech, blinking dot	0%	0%	1.6%

Table 4.16: The percentages and the occurrences of the modality combinations in re-designed multi-modal responses. For example, 21.0% is a result of percentage of occurrence of the combination of speech and URL among all concept-value using multi-modal (486 times). Combination of speech and speech with transition modeling are shown in rows. Transition modality means that the content (e.g. the point for pointing) only appear for a short time. However, URL, photo and map always stay on screen after display.

4.6.3 Deictic term

Deictic terms sometimes come with spoken content and can be used as an indication of the present of another modality. Therefore, we analyze which combinations of modality will contain a deictic term present. There are 29.4% (143/486) of concept-values using multi-modal include deictic terms. The occurrence is not very high because the expert often used the attraction name, location name or other named entities as a redundant information (see Figure 4.17) to make the message becomes more clear. It is different from multi-modal input, which usually includes deictic term(s). [57] Table 4.18 shows different modality combinations correspond to the use of deictic term.

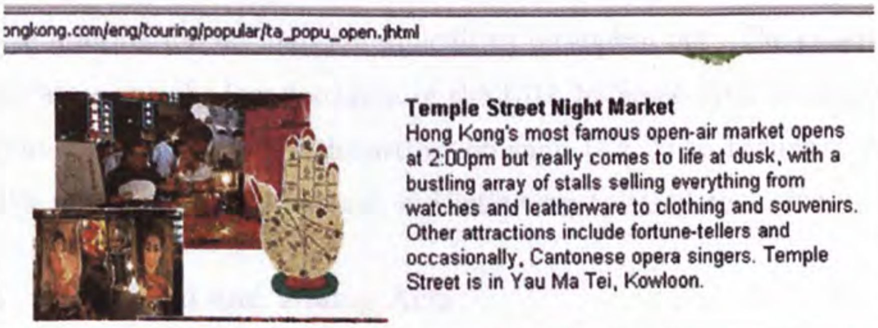
<table>
<tr><td>User 1:</td><td>油麻地有咩好玩？</td></tr>
<tr><td>System 1:</td><td>不如去廟街(the Temple Street)
[show the URL of the Temple Street])</td></tr>
<tr><td colspan="2">
ongkong.com/eng/touring/popular/ta_popu_open.jhtml
Temple Street Night Market
Hong Kong's most famous open-air market opens at 2:00pm but really comes to life at dusk, with a bustling array of stalls selling everything from watches and leatherware to clothing and souvenirs. Other attractions include fortune-tellers and occasionally, Cantonese opera singers. Temple Street is in Yau Ma Tei, Kowloon.</td></tr>
</table>

Table 4.17: The expert used the name of the street as a redundant information. (Figure) The URL of the Temple Street wich is a redundant information with the spoken content of 'the Temple Street' as shown in System 1.

Combination with map has a higher occurrence of deictic terms. When we study Table 4.18, the fourth row (italicized) shows that there is a higher occurrences of deictic terms in pointing with map than pointing with URL.

The usage of deictic terms in multi-modal response			
	URL	Photo	Map
Speech	9.1% (44)	6.4% (31)	2.3% (11)
Speech, pointing	*1.4% (7)*	*0.4% (2)*	*4.9% (24)*
Speech, circling	0.0% (0)	0.6% (3)	2.5% (12)
Speech, Highlight	1.2% (6)	0.0% (0)	0.0% (0)
Speech, Blinking dot	0.0% (0)	0.0% (0)	0.6% (3)

Table 4.18: The percentages and the occurrences of deictic terms with different modality combinations. For example, 44 times of speech and URL combination include deictic terms. The total number of occurrences is 486.

Circling with map (the fifth row) also has a higher occurrence of deictic terms than circling with URL. There are only a few deictic terms used when pointing or circling are shown together with a URL. This is because the address or locative information on map are difficult to be spoken out. The expert tried to indicate a specific location/area of the URL by name even using pointing or circling. The expert used the attraction name (e.g. 'The Temple street') or the title of a specific part instead of deictic term to avoid confusion.

4.6.4 Task Goal and Dialog Acts

Further analysis of the relationship between dialog acts and modality in ASK_ATTRACTION task goal shows that two of the dialog acts tend to include more multi-modal responses. ASK_ATTRACTION has the highest frequency of occurrences (333 response turns, 35%). Different dialog acts in the ASK_ATTRACTION tend to have their own preferences of modalities (refer to Table 4.19). Since some dialog acts in ASK_ATTRACTION are seldom occurred, we focus on the dialog acts which have occurrence more than 5%.

The filtered dialog acts consist of 84.4% of responses in ASK_ATTRACTION.

Two dialog acts, INFORM and SUGGEST, in ASK_ATTRACTION tend to use more multi-modal responses. Combinations of speech with photo or speech with URL are preferred. The other dialog acts such as REQUEST_COMMENT prefer to use speech only. A possible reason is that SUGGEST and INFORM includes more concept-values (e.g. SUGGEST always includes some attraction names). REQUEST_COMMENT and other dialog acts include fewer concept-values. For example, REQUEST_COMMENT is often in the form of '你鍾唔鍾意個建議' in speech.

Task goal: ASK_ATTRACTION	Occurrences of a modality (rank in decending order)		
Dialog act	First (%)	Second (%)	Third (%)
INFORM (121 responses)	Photo (34.7%)	URL (22.3%)	Speech only (15.7%)
SUGGEST (75 responses)	URL (29.3%)	Speech only (28.0%)	Photo (25.3%)
REQUEST_COMMENT (43 responses)	Speech only (81.4%)	URL (7.0%)	Map (7.0%)
FEEDBACK_POSITIVE (24 responses)	Speech only (75.0%)	URL (20.8%)	Highlight (4.2%)
REQUEST_INFO (18 responses)	Speech only (100.0%)	NA (0.0%)	NA (0.0%)

Table 4.19: Occurrences of modalities with different dialog acts in ranked order. The expert chose to use more multi-modal responses in the dialog acts of INFORM and SUGGEST. Simple uni-modal responses are used in the dialog act of FEEDBACK_POSITIVE.

4.6.5 Information Type

We focus on the relationship between information type and modality in ASK_ATTRACTION task goal with SUGGEST dialog act. Although INFORM is the most frequently occurred dialog act in ASK_ATTRACTION task goal, it is a catch-all dialog act and the variety in the content is very high. It is very difficult to analyze and more data is needed to observe the pattern of it. The pattern of SUGGEST is relatively more consistent and can be used for further study. We refer the six information types [56] in Chapter 3, Table 3.4. Table 4.20 shows the three most preferable combinations of modality of each type of information.

For each of the information type, we have to find out which combination of modalities is more suitable for each of them. Speech only is always the first priority of consideration. However, locative information, physical object information and descriptive information are less prefer to use uni-modal spoken responses. For locative information, it can be a specific point like an address or an area like as an district. For a specific point, the use of map and pointing is preferred. For an area, expert chooses to speak the district rather than pointing to the district on a large map (e.g. the full map of Hong Kong).

The expert chooses to display a URL (sometimes is photo) for long descriptive information such as history. While the expert prefers to use a simple spoken response only for some short description. We have found that the average length of descriptive information using multi-modal in term of Chinese characters is around 4.76 and the average length for the short spoken response is around 3.9 characters. Therefore, we use 4 as a threshold for the distinguish between long and short description.

Visual information like URL and photo are more suitable to introduce a physical object information which mainly is buildings. Expert chooses to show a single URL or photo to present physical object information. Table 4.21 shows

Information type	Occurrences of a modality (rank in decending order)		
	First (%)	Second (%)	Third (%)
Locative information (29 tokens)	Speech only (44.8%)	Map (31.0%)	Pointing (27.6%)
Descriptive information (143 tokens)	Speech only (55.2%)	URL (28.7%)	Photo (16.1%)
Physical object information (94 tokens)	Speech only (57.4%)	Photo (20.2%)	URL (19.1%)
Abstract information (24 tokens)	Speech only (70.8%)	URL (16.7%)	Photo (12.5%)
Nominal information (252 tokens)	Speech only (73.8%)	Photo (14.7%)	URL (9.9%)
Temporal information (16 tokens)	Speech only (100.0%)	NA (0.0%)	NA (0.0%)

Table 4.20: This Table shows the top three frequencies occur modality of different information type. We consider responses for ASK_ATTRACTION task goal with SUGGEST dialog act.

the example of the use of photo for physical object information.

User 1:	香港係唔係有個建築物叫會展，請問佢係點樣的？
System 1:	呢張係香港會議展覽中心的相片，請看 [Show the photo of Hong Kong Convention and Exhibition Centre]

Table 4.21: The expert used photo to describe a building.

4.7 Chapter Summary

In this chapter, we present the details of the multi-modal dialogue data collection in M3 corpus that use to support the research effort in MMRG. We improved multi-modal responses using expert knowledge. In order to annotate the multi-modal response, we have designed a convention of markup, namely M3 markup language (M3ML) with reference to the EMMA. We have identified possible modalities and their combination. We also studied several semantic-to-modality relation. In deictic term, the expert likes to use redundant information rather than deictic term in general. However, the expert used deictic term on speech with URL and any combination with map. For dialog acts in ASK_ATTRACTION, if the dialog act includes simple content like responses '好唔好', uni-modal is preferred. Information type is also affected the choice of modality. Using map for locative information is preferred. URL can be used to present descriptive information like history of attraction while photo is for physical object.

Chapter 5

Text-to-Audiovisual Speech System

Text-to-audiovisual speech (TTVS) [17] system is a useful and important tool for multi-modal presentation. It can offer a multi-modal presentation of dynamic information, e.g. for news and weather information reporting, for applications in edutainment, for personified dialogue systems, or as an aid for the hearing-impaired [40] where the simulated lip movements can help the user decipher the spoken message. The talking face can also convey non-verbal communicative signals, such as emotions [58].

This chapter describes our attempt to develop a TTVS system to support multi-modal presentation. The TTVS system accepts the Chinese textual input and generates synthetic speech with a talking face. For the speech generation, we use the syllable-based concatenative text-to-speech (TTS) synthesizer, CU VOCAL [59]. CU VOCAL supports different prosody and different speaking rate. In facial animation, we use a simple blending process based on weight morphing to achieve a high resolution animation of talking head. The TTVS system supports different expressions including agreement using head nod, disagreement using head shake and the emotion states that are smile and worry.

We drive the talking head animation parameters from speech parameters using linear-interpolation so as to integrate the speech and facial animation.

We will illustrate the basic unit of speech and facial animation in Section 3.1. We will then describe the face model definition and facial animation of our first TTVS system, LinLin, in Section 3.2. LinLin can be a virtual performer. We will talk about the integration between speech and facial animation in Section 3.3. We will discuss the user perception experiments in Section 3.4. The experiments evaluated how LinLin can improve the speech recognition. We will describe our second TTVS system, Windy, to support multi-modal presentation in Section 3.5.

5.1 Phonemes and Visemes

A phoneme is the smallest unit of sound in spoken language. A viseme is a visual equivalent to a phoneme. The definition of viseme is [1], "*A generic facial image that can be used to describe a particular sound.*"

CU VOCAL is a syllable-based concatenative text-to-speech (TTS) synthesizer for Cantonese, a major dialect of Chinese predominant in Hong Kong, South China and many overseas Chinese communities [60]. Cantonese is monosyllabic in nature (like Chinese) and the dialect has a rich tonal structure with between six to nine tones. Coarticulatory effects in CU VOCAL are captured in terms of distinctive features. The TTS engine also uses right tonal context for unit selection. Figure 5.1 illustrates typical input and output for CU VOCAL.

Chinese does not have explicit word delimiters and a word may contain one or more characters. Hence the input Chinese character string is tokenized into Chinese words by a greedy algorithm with reference to a lexicon and the word pronunciations are looked up from a dictionary. For example, in

[1] http://whatis.techtarget.com/definition/0,,sid9_gci213308,00.html

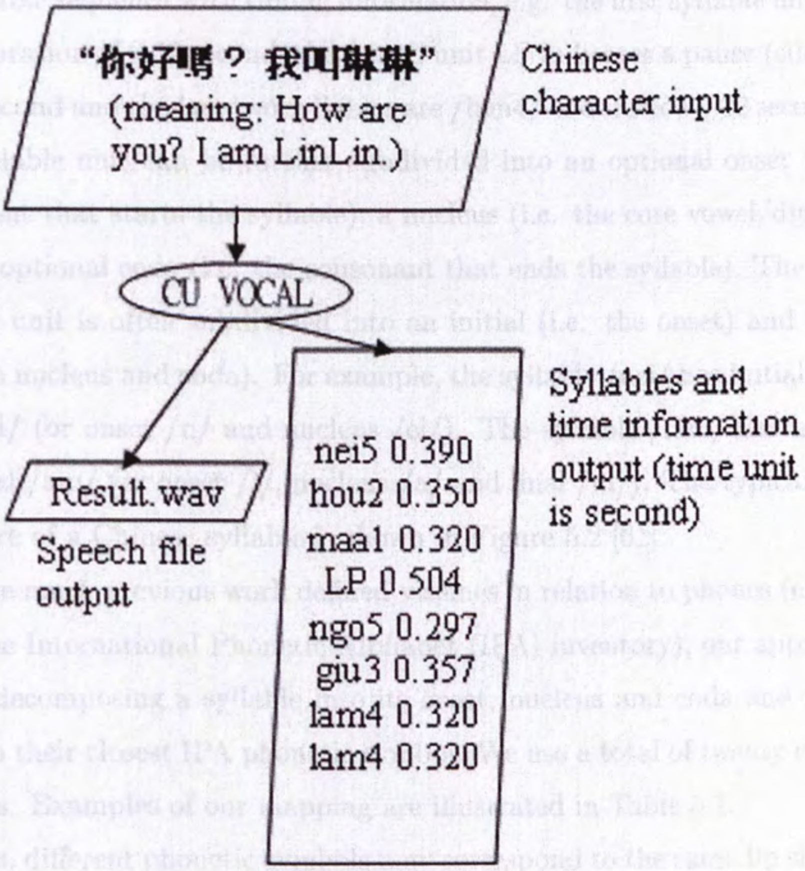


Figure 5.1: The sample input and output of CU VOCAL.

Figure 5.1, the first character in the input text string, "你 "(meaning: you), is pronounced as /nei5/ (i.e. the syllable is /nei/ with tone five. The syllable inventory adopted in CU VOCAL follows the Linguistic Society of Hong Kong (LSHK)[2] convention. CU VOCAL generates the synthetic speech output (in .wav format). The TTS engine has also been extended to explicitly generate the syllable sequence with timing information, e.g. the first syllable unit /nei5/ has a duration of 0.39 second, the fourth unit LP indicates a pause (silence) for 0.504 second and the last two syllables are /lam4/ of duration 0.32 second each. The syllable unit can be further subdivided into an optional onset (i.e. the consonant that starts the syllable), a nucleus (i.e. the core vowel/diphthong) and an optional coda (i.e. the consonant that ends the syllable). The Chinese syllable unit is often subdivided into an initial (i.e. the onset) and the final (i.e. the nucleus and coda). For example, the syllable /nei/ has initial /n/ and final /ei/ (or onset /n/ and nucleus /ei/). The syllable /lam/ has initial /l/ and final /am/ (or onset /l/, nucleus /a/ and final /m/). The typical syllable structure of a Chinese syllable is shown in Figure 5.2 [61].

Since much previous work defined visemes in relation to phones (e.g. those from the International Phonetic Alphabet (IPA) inventory), our approach involves decomposing a syllable into its onset, nucleus and coda and mapping these to their closest IPA phonetic symbol. We use a total of twenty eight IPA symbols. Examples of our mapping are illustrated in Table 5.1.

Since different phonetic symbols may correspond to the same lip shape, the twenty eight symbols are mapped to only sixteen visemes in total. Examples of mappings from symbols to visemes are provided in Table 5.2. Figures 5.3 also illustrate the viseme models of /b/ and /a/ respectively.

[2] Linguistic Society of Hong Kong, http://cpctp2.cityu.edu.hk/lshk

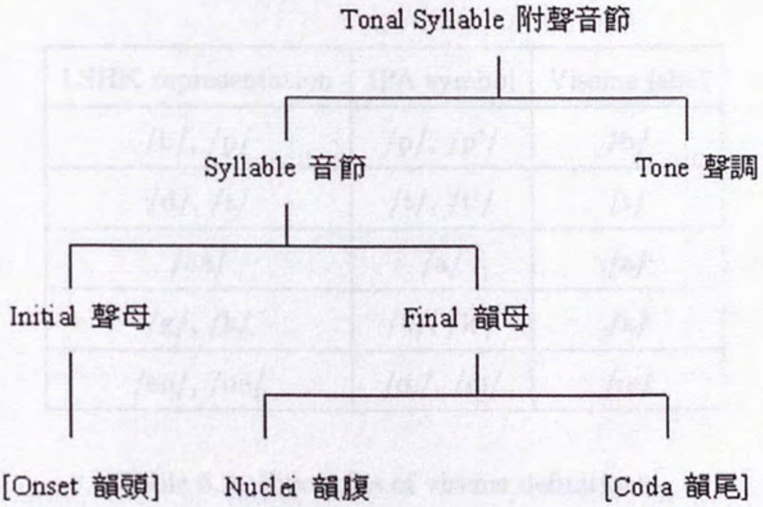


Figure 5.2: Typical structure of a Chinese syllable. The components in a pair of square brackets are optional consonants in a Chinese syllable.

LSHK representation	Corresponding IPA symbol
/aa/	/a/
/b/	/p/
/d/	/t/
/e/	/e/
/g/	/k/
/k/	/k'/
/o/	/p'/

Table 5.1: Sample mappings from LSHK to IPA syllable.

LSHK representation	IPA symbol	Viseme label
/b/, /p/	/p/, /p'/	/b/
/d/, /t/	/t/, /t'/	/t/
/aa/	/a/	/a/
/g/, /k/	/k/, /k'/	/k/
/eo/, /oe/	/œ/, /œ/	/œ/

Table 5.2: Examples of viseme definitions.

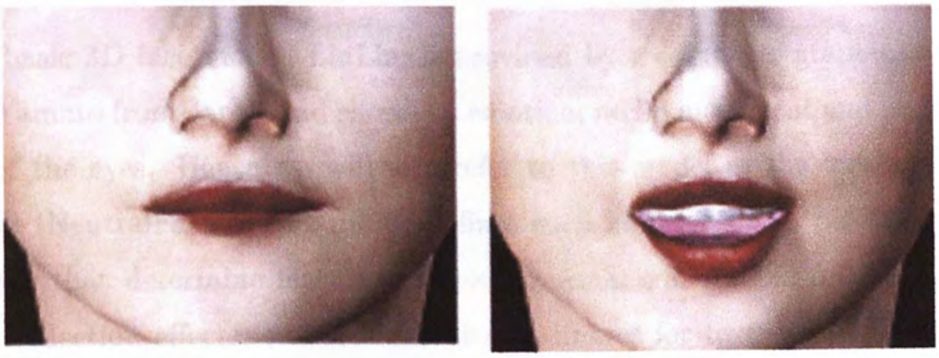
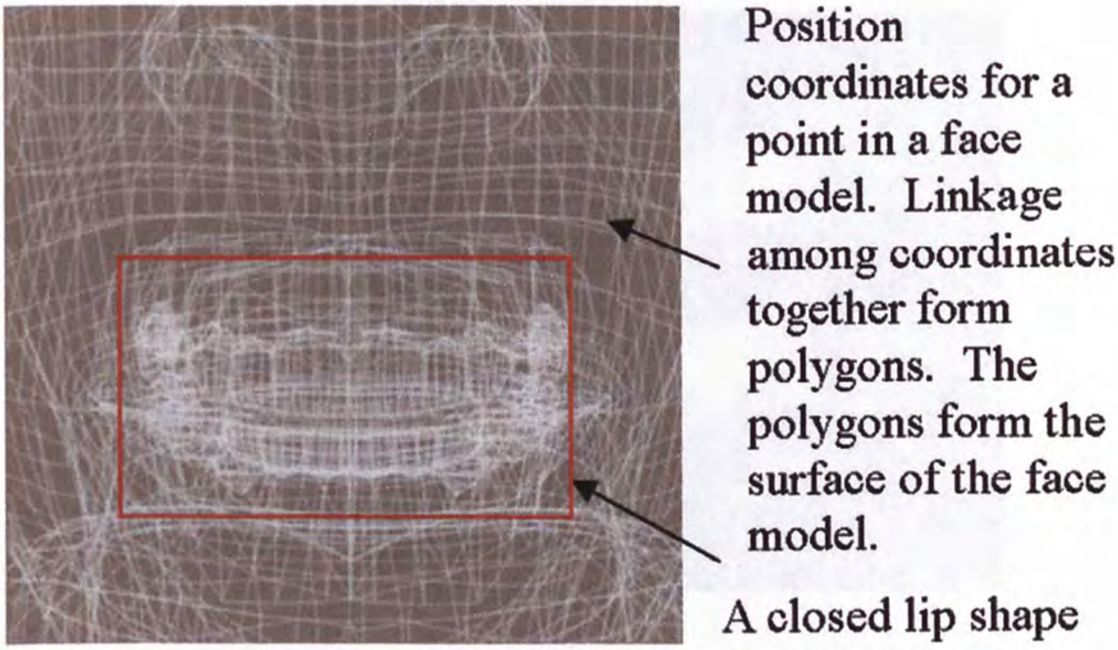


Figure 5.3: Static viseme models for the IPA symbols /b/ (left) and /a/ (right).



Position coordinates for a point in a face model. Linkage among coordinates together form polygons. The polygons form the surface of the face model.

A closed lip shape

Figure 5.4: The position coordinates and their linkages.

5.2 Three-dimensional Facial Animation

5.2.1 Three-dimensional (3D) Face Model

The basic 3D face model, LinLin, is provided by a computer graphics artist, Mr. Yamato from Japan and shows no emotion, no lip movement and no blinking of the eyes. Hence we will also refer to this model as the 'neutral' face model (NeutralFace). This model defines such information as position coordinates that determine feature positions, normal coordinates for computing light reflection effects, texture and its coordinates for texture mapping, and other information. Light reflection and texture help create a more realistic appearance of the face model. The position coordinates are linked together to form a network of polygons in order to determine the shapes of facial features (Figure 5.4).

Therefore we can define a 3D model in terms of a sequence of position

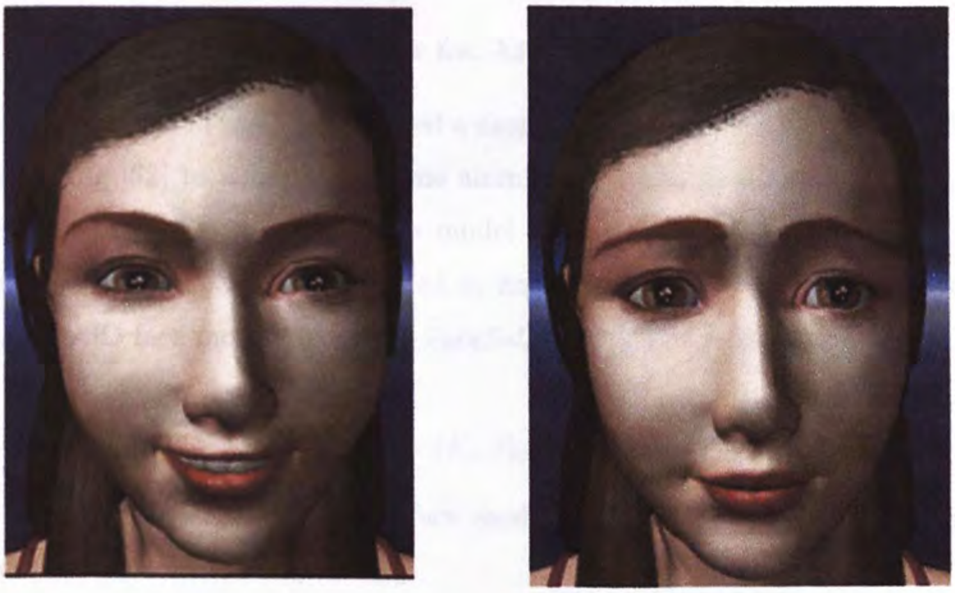


Figure 5.5: Static emotion models of "smile" (left) and "worry" (right).

coordinates (also known as vertices in a 3D object) as Equation 5.1,

$$F_j = (x_{0j}, y_{0j}, z_{0j}, x_{1j}, y_{1j}, ..., x_{mj}, y_{mj}, z_{mj}) \tag{5.1}$$

where F_j is a face model j, m is the number of position coordinates in a face model, x_{kj}, y_{kj}, z_{kj} are the k^{th} coordinate for x-axis, y-axis and z-axis in face model j

Modify the neutral face model by a 3D character design software tool, Poser ®5 can form the sixteen target viseme models and target emotion models (SmileFace and WorryFace). We have modified the models manually with reference to [62]. Examples are shown in Figures 5.3 and 5.5. Target models are static models that correspond to the visemes and emotions. Hence our TTVS system stores nineteen models in all (one neutral, sixteen viseme models, one smile model and one face model).

5.2.2 The Blending Process for Animation

We have designed and implemented a simple blending process based on weighted morphing [62] to achieve real-time animation. Animation focuses on the lip shapes and emotions. Each face model is represented by a vector in a 3 × m-dimensional space (m is defined in Equation 3.1 known as the *FaceSpace*. All the 3D face models form the *FaceSet*, as indicated in Equation 5.2.

$$FS = \{F_1, F_2, ..., F_n\} \tag{5.2}$$

where *FS* is a *FaceSet*, F_j is a face model $j \in$ *FaceSpace* and n is the total number of face models (n = 19).

Facial animation can be viewed as migration from one point in the *FaceSpace* to another point. A common approach to generate a new face model (*NewFace*) is generation by a linear combination of the face models in the *FaceSet* by the use of *blending weights* (see Equation 5.3). Each blending weight controls the dominance of its corresponding face model (in the FaceSet) in the NewFace model.

$$NewFace = \sum_{i=1}^{n} a_i F_i and \sum_{i=1}^{n} a_i = 1 \tag{5.3}$$

where a_i is the blending weight of each face model in the face set

In this work, we incorporate some modifications of the above method. Animation is achieved by the use of *deformation vectors* (DV) derived from the target viseme/emotion models. For example the DV for viseme label /b/ is defined in Equation 5.4 and the DV for smile is defined in Equation 5.5.

$$DV_{viseme_b} = Viseme\_b\_Face - NeutralFace \tag{5.4}$$

$$DV_{Smile} = SmileFace - NeutralFace \tag{5.5}$$

Different DVs can be linearly combined with the neutral face model (*NeutralFace*) to form a new face model (*NewFace*), as illustrated in Equation 5.6.

$$NewFace = NeutralFace + \sum_{i=1}^{n} a_i DV_i \tag{5.6}$$

where DV_i is the DV for the i^{th} face model in the *FaceSet* and a_i is the blending weight for face model i.

If we want the virtual character to smile more intensely, we simply increase the blending weight for DV_{Smile}. The blending weight can be used to control the facial expressions (smile and worry) and different lip shapes. The smile affects mainly the lip shape and worry affects the eyebrow.

5.2.3 Connectivity between Visemes

The transition between two phonemes in synthesized speech corresponds to the transition between two visemes in facial animation. Smooth transition is achieved by controlling the weights in the blending technique. We will elaborate on this point by means of an example. Consider TTVS system for the Chinese word "中間"(meaning: center) pronounced as /zung/ /gan/ in LSHK syllables. For a given syllable, we reference the CU VOCAL syllable corpus to get the average duration among the occurring instances. For example, the syllable /zung/ averages 0.33 second in duration. We also reference the corpus to get the average fraction of the syllable's duration that is occupied by its initial and final respectively. For example, the syllable /zung/ has the initial /z/ and final /ung/. The initial /z/ takes up about a quarter of the syllable's duration on average, while the remaining three quarters is taken up by the final /ung/. The final can be further subdivided into the nucleus /u/ and coda /ng/. For the sake of simplicity, we assume the nucleus and coda for the final /ung/ have equal average durations. Hence about 0.5 of the average duration of /zung/ is occupied by the syllable onset /z/, about 0.375 by the syllable

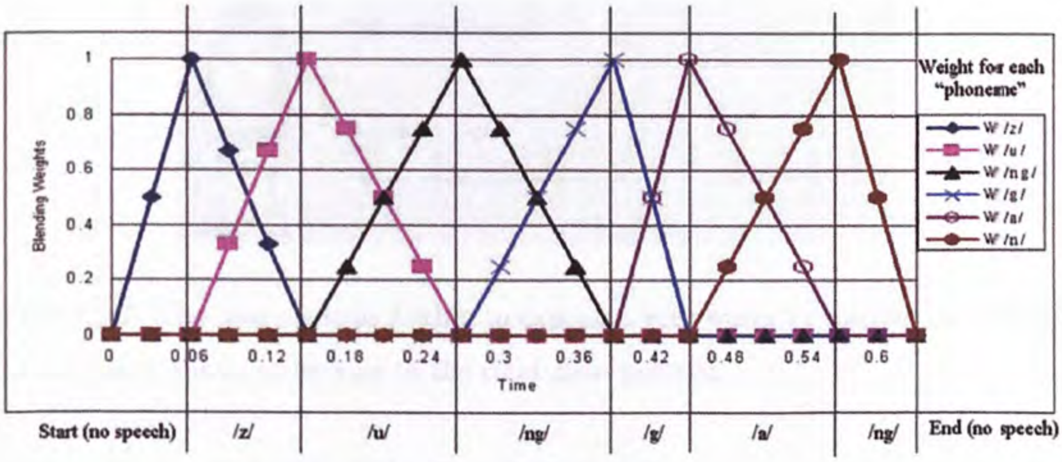


Figure 5.6: Variation of blending weights over time for three-dimensional animation. The blending weight of a viseme is zero before the start of animation or after the end of animation. During the animation, it is changed using a linear interpolation.

nucleus /u/ and the remaining fraction of 0.375 by the syllable coda /ng/. In order to use this information for facial animation, we locate the visemes that correspond to the IPA symbols /z/, /u/ and /ng/ respectively. Since these are static viseme models, we need to determine the blending weights that correspond to these visemes for 3D animation. A linear interpolation is used as shown in Figure 5.6. Each viseme starts with a unity weight at its start instant, and linearly decreases to zero weight at its end point. This defines the variation of the blending weights over time and our system demonstrates that this achieves a realistic and smooth facial animation effect.

The variation of blending weights for emotion face models (see Figure 5.5) used in a similar way for 3D face rendering, as compared with the viseme weights. These are defined manually by the user by means of a slider rule in our system's interface (see Figure 5.7). The overall control flow for our TTVS system is depicted in Figure 5.8.

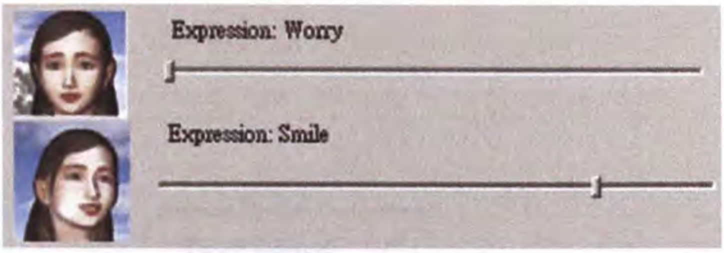


Figure 5.7: The user controls LinLin to express a very worry by moving the marker in the lower one of slider rule to the right most position.

5.3 User Perception Experiments

We ran user perception experiments with twelve randomly generated seven-digit strings. For each digit string we generate either (i) an audio recording of the synthesized speech in a noisy (cafeteria) environment; or (ii) a video file that augments the noisy synthesized audio with a talking face. Our tests involve sixteen Cantonese-speaking subjects. Each subject is presented with the twelve audio/video files and asked to write down the digit string that was spoken. Subjects have no prior knowledge of the lengths of the digit strings. Equations 3.7 and 3.8 show the calculation of total error rates and accuracy. Table 5.3 shows the experimental results in terms of substitution (S), deletion (D), insertion (I) errors and accuracy.

$$\begin{aligned} Total\ error\ rates(\%) &= Substitution\ error\ rates(\%) \\ &+ Deletion\ error\ rates(\%) + Insertion\ error\ rates(\%) \end{aligned} \tag{5.7}$$

$$Accuracy(\%) = 1 - Total\ error\ rates(\%) \tag{5.8}$$

The digits '5' and '2' are pronounced in Cantonese as /ng3/ and /yi6/ respectively. These are often misrecognized due to their low energies. Furthermore, their visemes look similar —both have a slightly open lip shape.

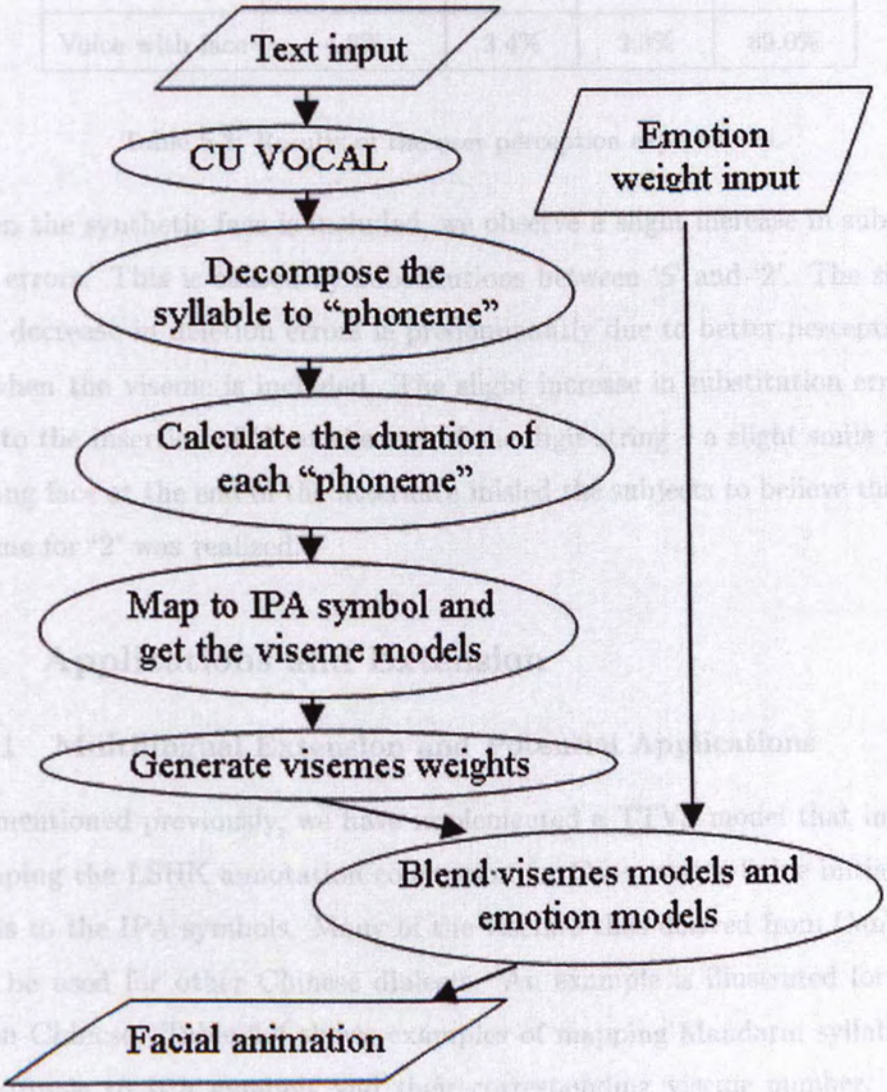

Text input
CU VOCAL
Emotion weight input
Decompose the syllable to "phoneme"
Calculate the duration of each "phoneme"
Map to IPA symbol and get the viseme models
Generate visemes weights
Blend visemes models and emotion models
Facial animation

Figure 5.8: The overall flow about text to audiovisual speech.

	Error rates (%)			
	Substitution	Deletion	Insertion	Accuracy
Voice only	3.1%	14.7%	1.6%	80.6%
Voice with face	4.8%	3.4%	2.8%	89.0%

Table 5.3: Results of the user perception experiments.

When the synthetic face is included, we observe a slight increase in substitution errors. This is caused by substitutions between '5' and '2'. The significant decrease in deletion errors is predominantly due to better perception of '5' when the viseme is included. The slight increase in substitution errors is due to the insertion of '2' at the end of the digit string – a slight smile in the talking face at the end of the utterance misled the subjects to believe that the viseme for '2' was realized.

5.4 Applications and Extension

5.4.1 Multilingual Extension and Potential Applications

As mentioned previously, we have implemented a TTVS model that involves mapping the LSHK annotation convention for Cantonese syllable initials and finals to the IPA symbols. Many of the visemes thus derived from Cantonese can be used for other Chinese dialects. An example is illustrated for Mandarin Chinese. Table 5.4 shows examples of mapping Mandarin syllable initials/finals to IPA symbols and their corresponding viseme number. Analogous mapping processes can be applied to other languages to extend the facial animation in our TTVS system to other languages. As a demonstration, we have generated the IPA phonetic sequences from the lyrics of four songs (in Cantonese, Mandarin, Japanese and English respectively) and then derive the viseme sequence (with timing information) for real-time facial animation.

Hence in addition to TTVS in Cantonese, the virtual character LinLin can sing in the four different languages. The TTVS system supports some special effects such as snowing abd moving background so as to perform better during singing. Users can interact with LinLin and have a try on several potential applications (including news reporter, air-ticket booking and English-to-Chinese machine translation) through the user interface as shown in Figures 5.9 and 5.10. A video demonstration of the potential applications and different languages is available at http://www.se.cuhk.edu.hk/TTVS.

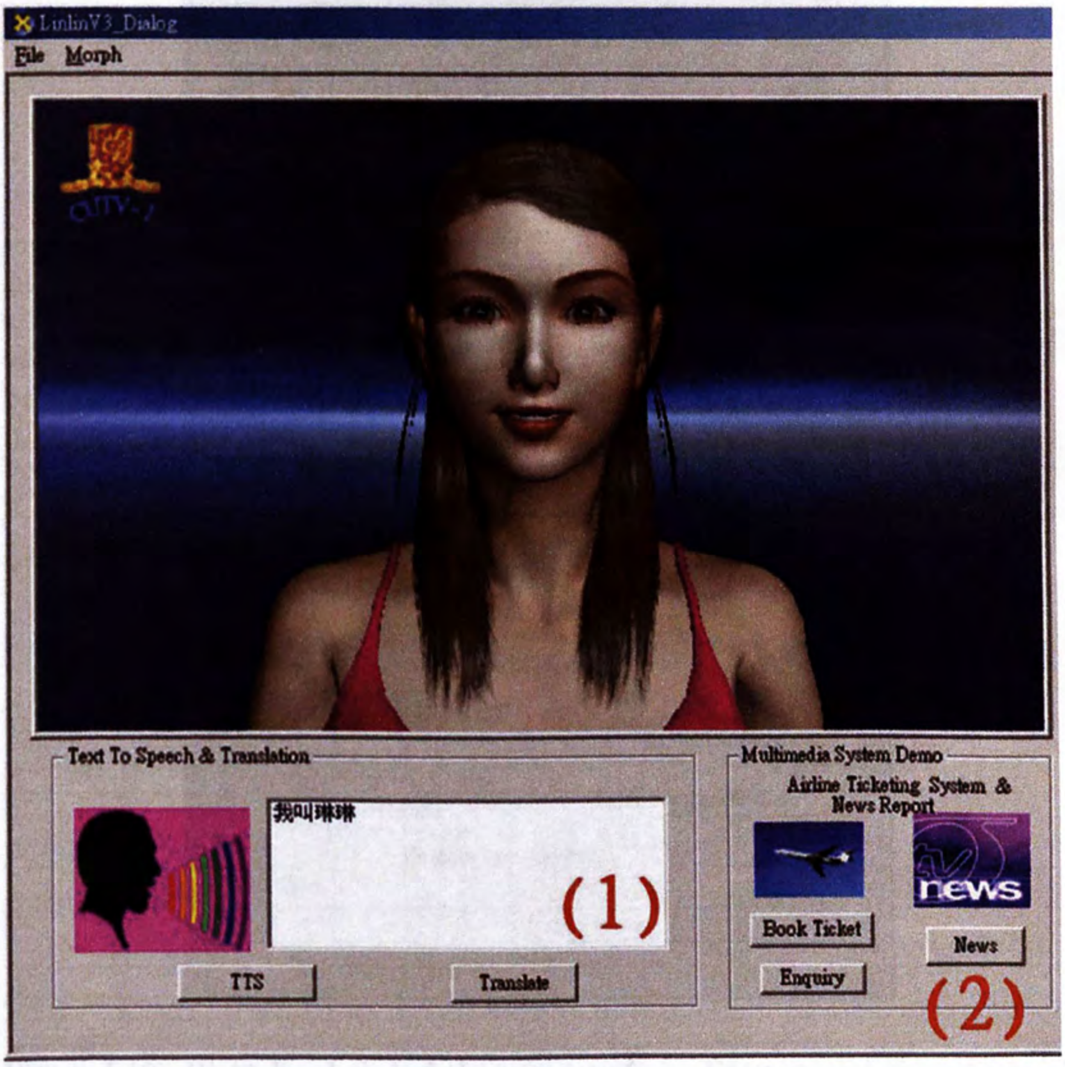
Mandarin pronunciation symbol	IPA symbol	Viseme label
/b/, /p/, /m/	/p/, /p'/	/b/
/d/, /t/	/t/, /t'/	/t/
/g/, /k/	/k/, /k'/	/k/

Table 5.4: Example mappings between IPA and Mandarin sub-syllable structures.

5.5 Talking Head in Multi-modal Dialogue System

The requirement of a TTVS system in multi-modal dialogue system is different from a virtual performer. For example, LinLin used the graphical processing unit (GPU)[3], to support some special effects such as snowing and the moving background. The GPU requires special display card so the choice of system platform is limited to the computer with such graphic card (i.e. mainly desktop computer). Since the multi-modal dialogue system may be used in different devices such as Tablet PC, we need to reduce the computation power of new talking head, Windy and make it possible to be used on Tablet PC. We have removed those special effects in Windy so that GPU is not a must for it. Table 5.5 shows the comparison between LinLin and Windy. We have applied

[3] GPU Definition, http://www.isprank.com/Glossary/GPU.html



我叫琳琳

Figure 5.9: Left hand side of the user interface. Users can input Chinese text in (1) and LinLin will speak it out. Users can have a try on different potential applications such as news reporter and booking ticket through (2).

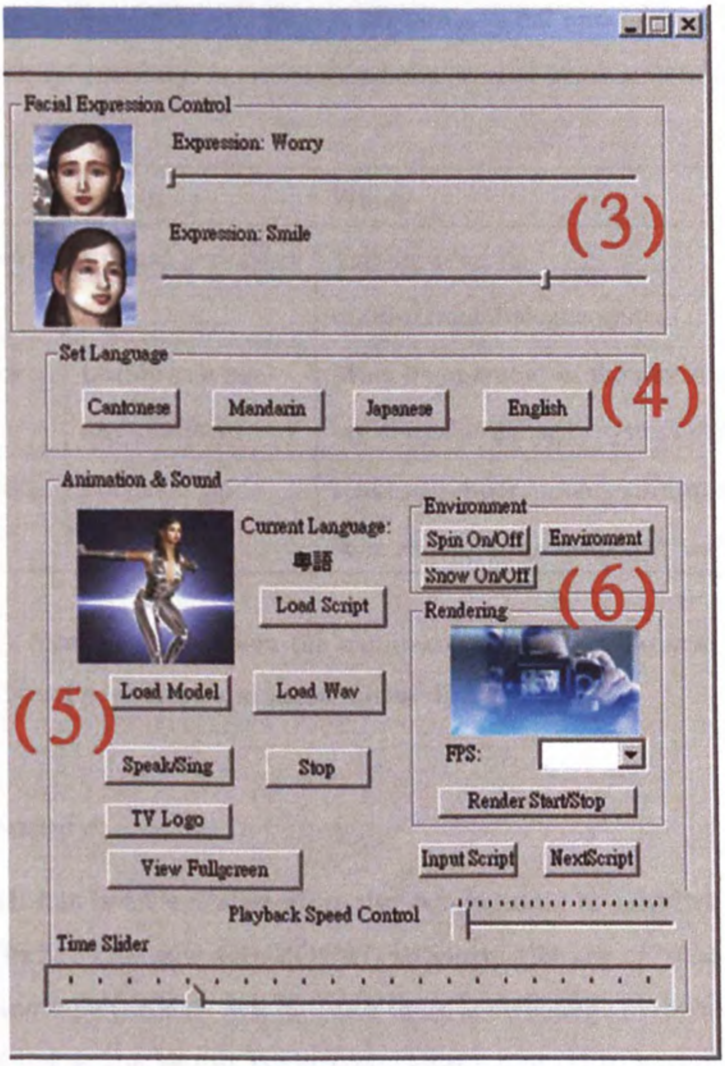


Figure 5.10: Right hand size of the user interface. Users control the weights of expression (smile and worry) through control bar (3). Users can request LinLin to sing in different languages through (4). The buttons for the system controllers such as loading model and full screen display with (5). Users can play the snowing effect through (6).

the animation and integration techniques of LinLin into Windy for providing multi-modal dialogue system. Windy supports vocal and non-verbal communication such as prosody and non-vocal-non-verbal communication like head nod.

	LinLin	Windy
Purpose	Virtual performer	Talking head in multi-modal dialogue system
Device	LinLin can use any platform	Must be operated in the platform of multi-modal dialogue system running
Function	Focus on lip-movement	Need to support more expressions such as emphasis, head nod and shake

Table 5.5: A comparison between the requirement of a TTVS system in the applications of virtual performer and multi-modal dialogue system.

5.5.1 Prosody

CU VOCAL has been updated after the development of LinLin and has been used in Windy. The new CU VOCAL supports the use of Speech Synthesis Markup Language (SSML) [63] [59] and there is no change in its output format. SSML falls below the World Wide Web Consortium (W3C) standard[4]. SSML is a XML-based markup language and aims to provide a standard way to control aspects of speech such as volume, pitch, rate, etc.

We can control the presentation of speech to emphasize something using the tag of <emphasis> in SSML. We can input the speech content with SSML tags to CU VOCAL and use its output to drive the animation. For some named entity, which may lead to ambiguation or difficult to recognize, we can

[4] World Wide Web Consortium http://www.w3c.org

explicitly slow down the speaking rate by the tag of <prosody rate='x-slow'>. Table 5.6 shows a dialogue to illustrate the usage of SSML in CU VOCAL.

User 1: 夜晚有邊個地方好玩?
System 1: 我會建議你去<emphasis level='strong'>蘇豪</emphasis>.
User 2: 蘇豪係邊到呢?
System 2: 蘇豪位於<prosody rate='x-slow'>士丹利街及伊利近街一帶 </prosody>.

Table 5.6: An example shows the use of SSML in CU VOCAL. We want to emphasize the focus, i.e. the suggested attraction name, in System 1. We also slow down the speaking rate to ensure that users can recognize the street names correctly in System 2.

5.5.2 Body Gesture

Human may show agreement and disagreement with head nod and head shake. Therefore, we add a head nod model and two head shake models (turn left and turn right) so as to simulate the actions of head nod and head shake in Windy. The animation technique of head nod and head shake is the same as the one used to express smile or worry in LinLin. We can control the blending weights of the models to achieve different levels of expression as shown in Figures 5.11 and 5.12

5.6 Chapter Summary

This chapter describes our work on TTVS systems to support multi-modal presentation. LinLin can generate highly natural synthetic speech that is precisely time-synchronized with a real-time 3D face rendering. Our Cantonese TTVS system utilizes a home-grown Cantonese syllable-based concatenative text-to-speech system named CU VOCAL. This chapter describes the extension of CU VOCAL to output syllable labels and durations that correspond to the output acoustic wave file. The syllables are decomposed and their initials/finals mapped to their nearest IPA symbols that correspond to static viseme models. We have also defined two static face models that correspond to emotions. In order to achieve 3D face rendering, we have designed and implemented a blending technique that computes the linear combinations of the static face models to effect smooth transitions in between models. We have demonstrated that this design and implementation of a TTVS system can achieve real-time performance. In tonel language (e.g. Mandarin and Cantonese), the lip position will be affected by different tone [64]. However, tone are mainly converyed by recognizing the vibrating frequencies of the larynx [65]. To simplify the TTVS system, we do not consider the tone effect on lip. Although there is no consideration on the tone effect, the TTVS system still can improve the user perception by providing the synchronized lip movements. Moreover, we adopt the animation and integration techniques from LinLin to Windy. The focus of Windy is to support the usage in multi-modal dialogue system. We have used SSML annotations on Windy so that it is able to express in different ways. Windy is able to present agreement and disagreement through head nod and head shake.

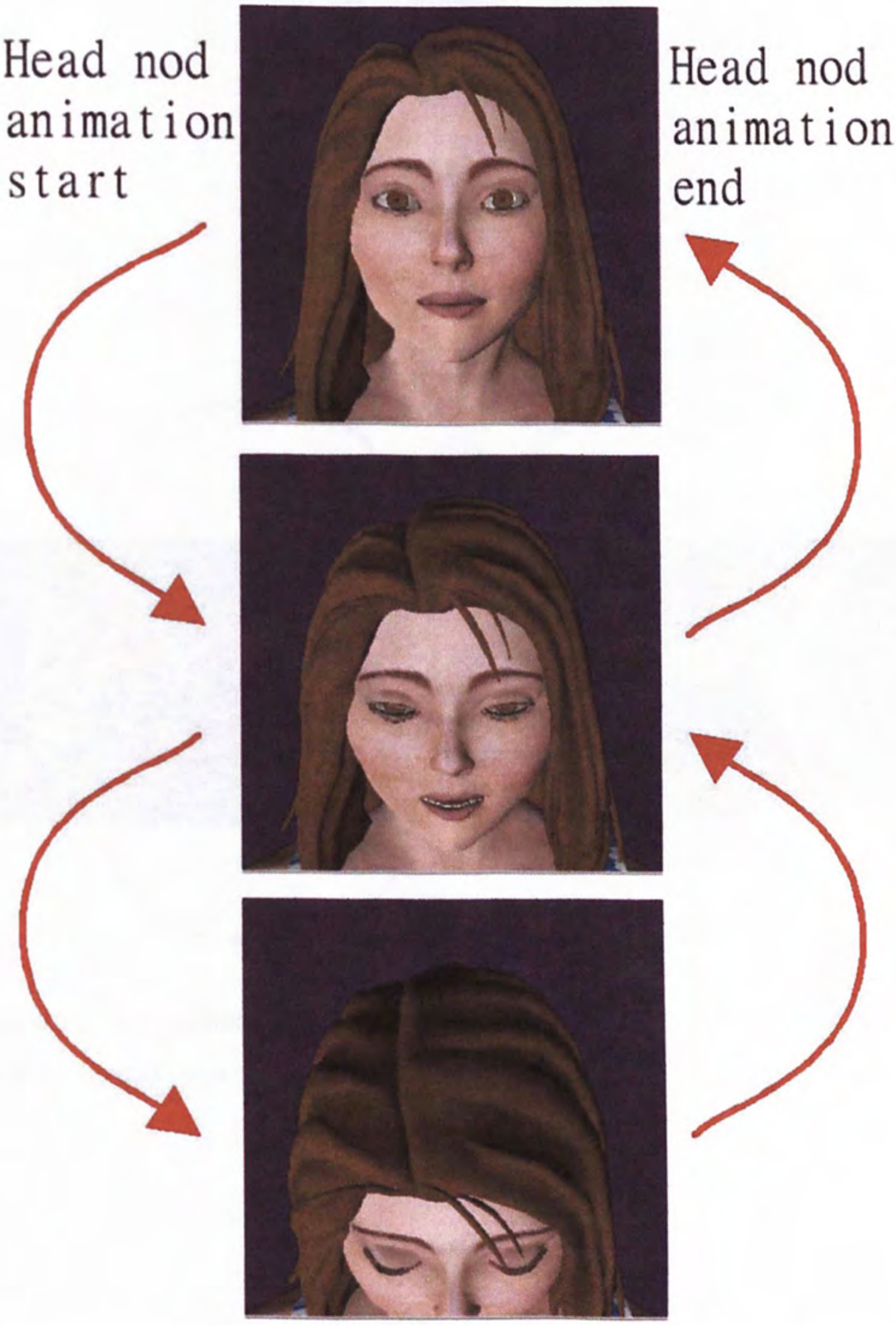


Figure 5.11: An graphical illustration of head nod of Windy. The animation is started from the top to the bottom photo and then in reverse direction to back to the normal state.

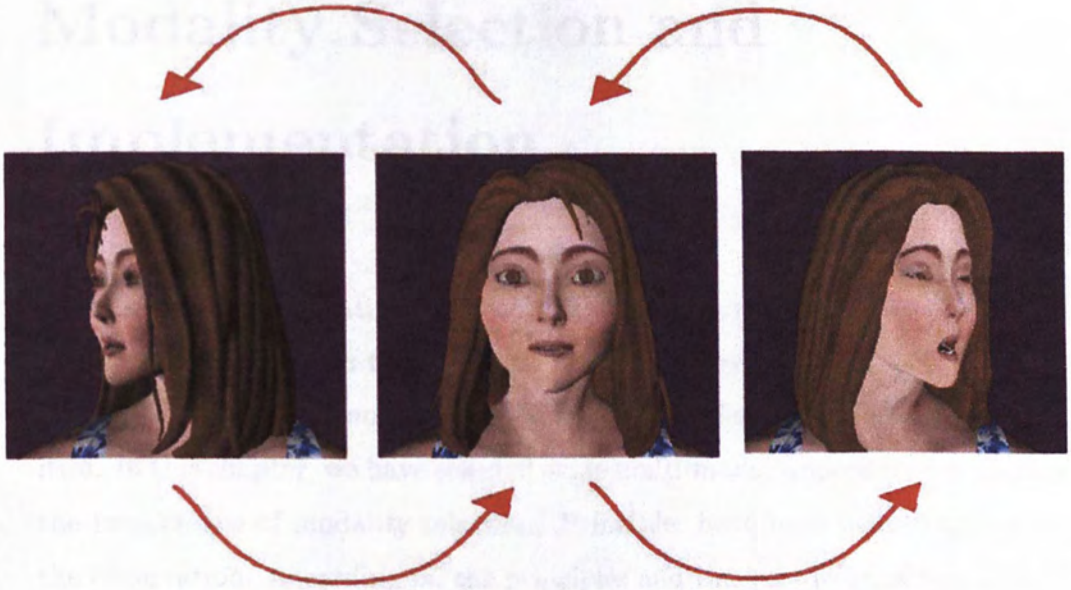


Figure 5.12: An graphical illustration of head shake of Windy. Windy shakes her head from normal state to the left and then go to the right to present a disagreement.

Chapter 6

Modality Selection and Implementation

An efficient communication relies on suitable choice of modalities. The study in Chapter 4 has shown that different intentions (task goals and dialog acts), information types and modality combinations will affect the selection of modalities. In this chapter, we have selected some multi-modal responses to illustrate the importance of modality selection. Principles have been defined based on the observation. According on the principles and the knowledge of human behavior on modality selection from Chapter 4, we have developed some heuristic rules to perform modality selection. The temporal coordination, physical layout and the use of deictic term are considered. Finally, we have implemented a preliminary system to demonstrate the multi-modal response generation.

6.1 Multi-modal Response Examples

In Chapter 4, we found that INFORM and SUGGEST in ASK_ATTRACTION use multiple modalities more frequently. In the top five dialog acts in ASK_ATTRACTION, there are 170 multi-modal responses out of 333 responses. INFORM and SUGGEST include 156 multi-modal responses (91.8%). Tables 6.1 and 6.2 summarize the information.

Uni-modal responses	163 response turns
Multi-modal responses	170 response turns
Total	333 responses turns

Table 6.1: The distribution of uni-modal and multi-modal responses in ASK_ATTRACTION task goal.

INFORM	102 response turns
SUGGEST	54 response turns
Other dialog acts	14 response turns
Total	170 response turns

Table 6.2: The distribution of dialog acts in multi-modal responses with ASK_ATTRACTION task goal.

Locative information, descriptive information and physical object information also prefer the use of multiple modalities. Therefore, we focus on the combinations of these two dialog acts and three information types. We will examine responses with single concept-value and multiple concept-values together with the same or different information types in this section. Table 6.3 shows the distribution of different number of concept-values.

6.1.1 Single Concept-value Example

The simplest multi-modal response only includes one concept-value. An example is shown in Table 6.4, in which the location of SoHo is the only concept-value.

Number of concept-value	Number of response turns
0	11 (5.6%)
1	58 (29.6%)
2	54 (27.6%)
3	37 (18.9%)
4	25 (12.8%)
5	5 (2.5%)
6	2 (1.0%)
7	4 (2.0%)
Total	196

Table 6.3: The distribution of the number of concept-values in responses with INFORM or SUGGEST dialog act and ASK_ATTRACTION task goal. Responses with zero concept-value means the response only contains some backchannel message such as 'Emm..'

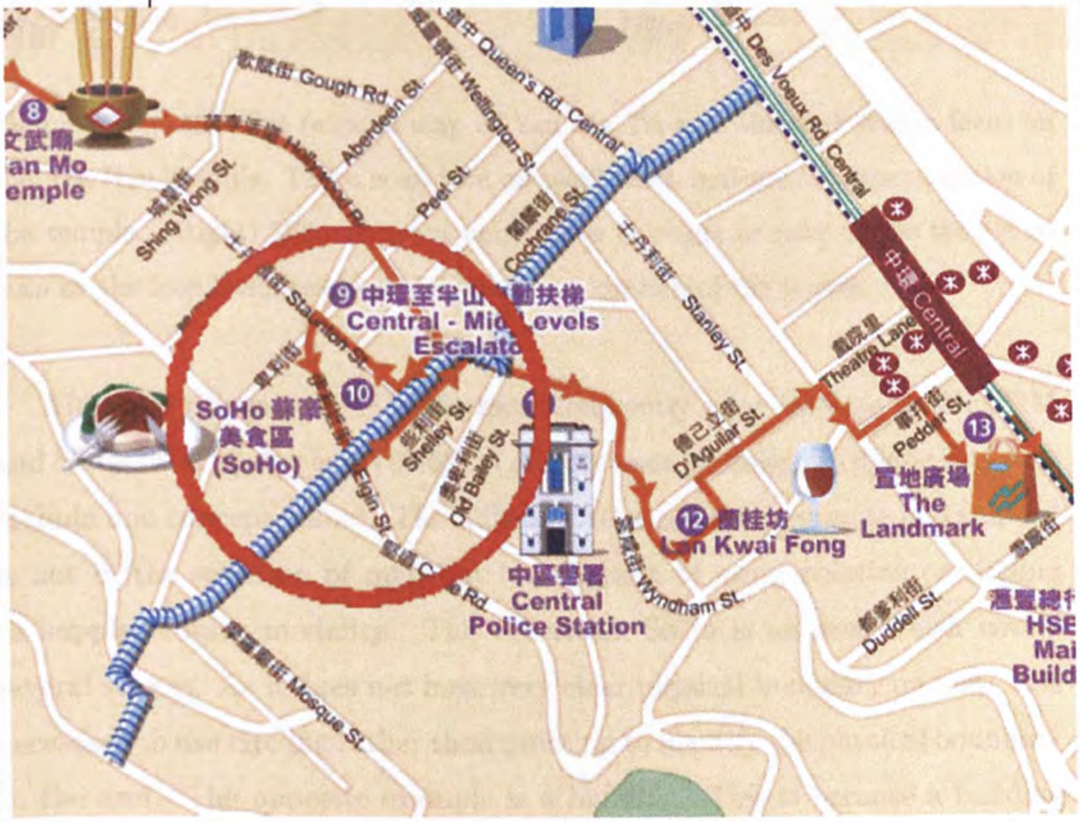
User 1:	個景點蘇豪(SoHo)係邊度架?
System 1:	就係哩度[Show the SoHo map][Circling the area]
	

Table 6.4: An example of a response with one concept-value, i.e. the location of SoHo.

(Figure) System uses a red circle to inform where the attraction is and indicates the physical boundary in System 1.

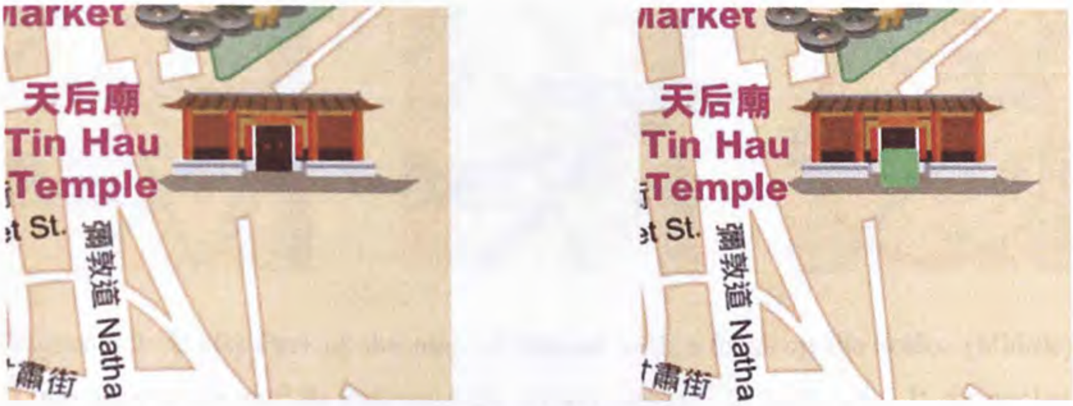


Figure 6.1: (Left) The original map of Yau Ma Tei and Mongkok with a focus on the Tin Hau Temple. There is an icon on map which indicate the exact location of the temple. (Right) We can use a point (lime in color) to refer to the temple on map as the icon has already shown the exact location of the temple.

Although it is simple, it is the most frequently occurred case in INFORM and SUGGEST dialog acts (29.6% of all responses in these two dialog acts only include one concept-value). The difficulty to handle the example in Table 6.4 is not in the selection of map but the decision of using pointing or circling as supplementary modality. The attraction SoHo is an area which covers several streets. As it does not have very clear physical boundary on map, it is necessary to use circling rather than pointing to identify the physical boundary to the user. The opposite example is a building. This is because a building on map has a clear physical boundary. Figures 6.1 and 6.2 illustrate with and without clear physical boundary.

6.1.2 Two Concept-values with Different Information Types

When there are two different information types in one response, the situation will be more complex. Since the task goal is ASK_ATTRACTION, physical object information such as building is usually used to represent an attraction.

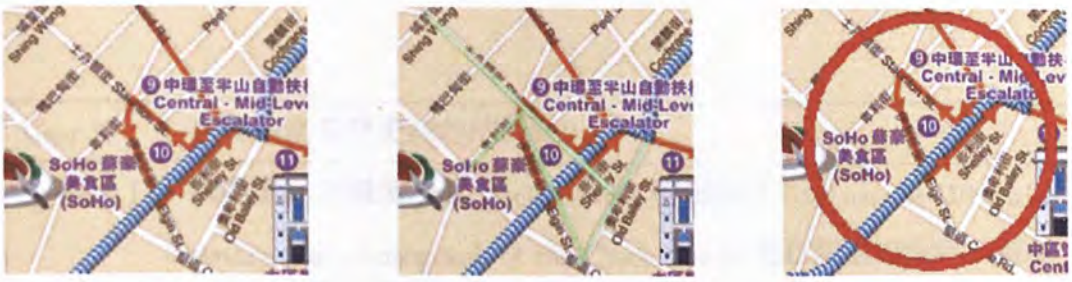


Figure 6.2: (Left) Part of the map of Central with a focus on the SoHo. (Middle) Lime lines in the middle indicated the streets included in SoHo area. It shows that the SoHo is an area and does not have a clear boundary indicated on map. (Right) We can use a circle to roughly outline the area of SoHo.

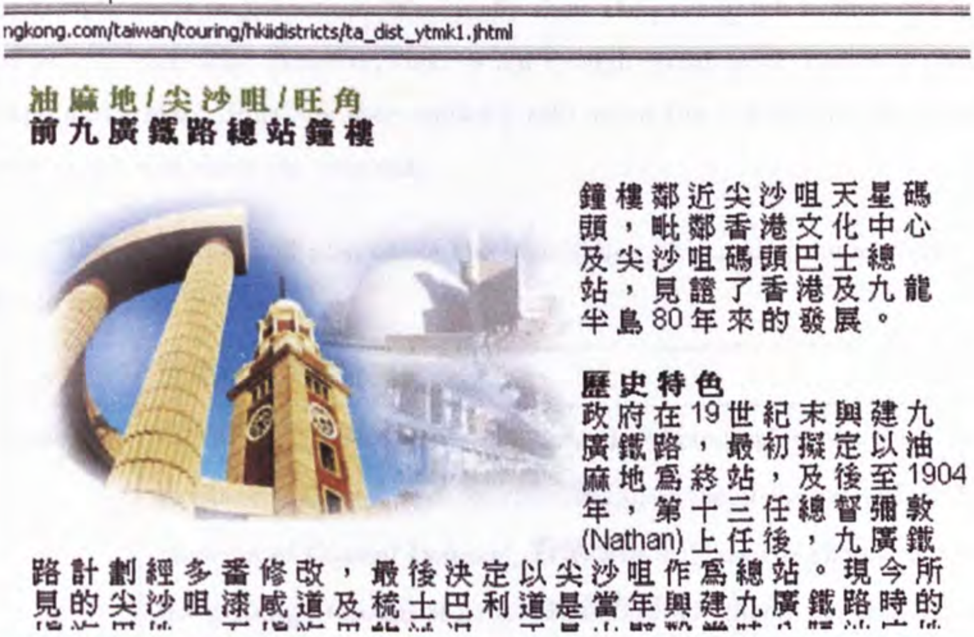
A response may provide the location or description as supplementary information. Table 6.5 shows that a response with a building which is physical object information and the history of the building which is descriptive information. It shows that the selected modality for individual concept-value may cause some unnecessary redundancy in multi-media information. It is because the URL already included the photo. The difficulty is the selection of a combination of modality which minimize the unnecessary multi-media information redundancy.

The selected modality may scramble for the visual space. The modality selection needs to decide which modality should be used to satisfy the visual space constraint. Table 6.6 shows an example which includes locative information and also physical object information.

6.1.3 Multiple Concept-values with Same Information Types Example

A response may include several concept-values that belong to the same information type. This is common for multiple attractions because the task goal is ASK_ATTRACTION. For example, system suggests several museums to a

User 1:	有邊的富有歷史特色的景點?
System 1:	我建議你去鐘樓(the Clock Tower).哩個網頁上[Show the URL with history and photograph of the Clock Tower]就有相關的資料．請睇.



ngkong.com/taiwan/touring/hkiidistricts/ta_dist_ytmk1.jhtml

油麻地/尖沙咀/旺角
前九廣鐵路總站鐘樓

鐘樓鄰近尖沙咀天星碼頭，毗鄰香港文化中心及尖沙咀碼頭巴士總站，見證了香港及九龍半島80年來的發展。

歷史特色
政府在19世紀末興建九廣鐵路，最初擬定以油麻地爲終站，及後至1904年，第十三任總督彌敦(Nathan)上任後，九廣鐵路計劃經多番修改，最後決定以尖沙咀作爲總站。現今所見的尖沙咀漆咸道及梳士巴利道是當年興建九廣鐵路時的

Table 6.5: An example of avoiding unnecessary redundancy in multi-media information. The response replies a historical building with its history. We prefer to show the photograph for the building to illustrate what it is and show an URL to provide the historical information. The URL has already included the photograph of the historical building. We will not show another photograph, as it will be an unnecessary redundancy.

(Figure) A screen capture of the URL of the Clock Tower as shown in System 1. The Clock Tower history and a photograph were shown in an URL. The wizard shows this URL to the user at the time of suggestion of the Clock Tower.

User 1:	旺角附近好似有個廟架．唔知佢係邊度呢？
System 1:	我估你所指的係天后廟(Tin Hau Temple).佢係喱度[Show the map of Tin Hau Temple]

Table 6.6: An example of scrambling of visual space. The response replied with the temple name and location. We usually show the photograph to illustrate how the temple looks like. However, there is not enough visual space to show a photograph and a map. Since the user explicitly asks about the location, we give up the photograph and show the map only.

user. The response will also cause the scrambling of visual space as shown in Table 6.7.

User 1:	香港有邊▯的博物館值得我去?
System 1:	我介紹你去香港文化博物館[Show the photograph of the Hong Kong Heritage Museum],香港海防博物館 [Show the photograph of the Hong Kong Museum of Coastal Defence], 香港藝術館 [Show the photograph of the Hong Kong Museum of Art],香港歷史博物館 [Show the photograph of the Hong Kong Museum of History]，香港科學館，香港太空館．你需唔需睇埋其他博物館的相片？

Table 6.7: An example of suggesting multiple attractions. The response suggests six major museums listed on the Discover Hong Kong website. Although we want to show the photograph of each museum, the visual space is not enough.

6.2 Heuristic Rules for Modality Selection

After observing the response patterns in multi-modal responses, we have defined a set of principles for modality selection.

6.2.1 General Principles

Chapter 4 shows that each information type has its own preferences for different modalities. Descriptive information, like history or abstraction, usually in textual form and URL is more suitable. Table 6.5 has illustrated the using of URL for history. Therefore, we define the principle of modality preference as:

Principle of Modality Preference:

Choose the most preferable modality for each information type if possible

Modality preference only consider single modality but not the combination. However, in some cases, we need to choose more modalities. For map, we use circling to define an area when there is not any explicit physical boundary (refer to Figure 6.2). We need to clearly tell where is the area. We call this principle as preciseness:

Principle of Preciseness:

Sharply define the reference's physical boundary.

Selection of more modalities does not mean that it would be better. Conciseness response is always preferred. As shown in Table 6.5, unnecessary multi-media information should be avoided. Concise also means using a few words to present the same response. One practical example is the long address, we can use a map with a deictic term rather than speak out the long address. Therefore, we come up with the third principle as:

Principle of Conciseness:

Minimize the use of words and unnecessary multi-media information.

We need to consider that a response can achieve its communication goal or not. The concept-value in focus is the most important information to achieve the communication goal. For example, when the response's communication goal is providing the location of an attraction, the concept-value in focus should be its address. When there is a modality conflict (e.g. not enough visual space), the concept-value in focus should take the most preferable modality first. Similarly, more important information should have higher priority on modality selection. The importance is necessary to be defined in the input of modality selection. Therefore, we summarize the principle of focus as follow:

Principle of Focus:

First priority of modality selection is giving to the concept-value in focus and then other important concept-value(s).

We also need to consider the visual space constraint. The visual space of an user interface is always limited. It is impossible to show a large map and also show three large photographs on screen at the same time. The visual space constraint forces us to select a modality combination to satisfy the visual space. Together with considation of the principle of focus, we can make a right modality selection to satisfy the visual space constraint.

6.2.2 Heuristic rules

Based on above principles, we have defined a set of domain specific heuristic rules. The rules consider the principles as mentioned before, the knowledge from Chapter 4 and some practical considerations. We will explain them one by one. We consider modality preference and have three rules. We present them in Table 6.8. In locative information (row two and three), according to the knowledge from Chapter 4, speech only is the most preferable one. However, we need to consider that the visitors do not know where is Mongkok (a district in Hong Kong) (Practical consideration). Therefore, we always choose the second most preferable modality which is map. Moreover, according to the principle of preciseness, we choose circling to indicate an area when no explicit physical boundary of the location can be found. The existing physcial boundary of a location is determined manually and stored into a database. Chapter 4 has already shown that when there is descriptive information, we should use URL for long description and show in row four. As mentioned before, we use a threshold of four characters to distinguish between long and short description according to the expert perference. In row six, we always show photograph to make the response becoming more interesting.

Information type	Map and Pointing	Map and Circling	Url	Photograph	Speech only
Locative information with physcial boundary	√				
Locative information without physcial boundary		√			
Long descrptive information (>four Chinese characters)			√		
Short descriptive information					√
Physical object information				√	

Table 6.8: The information and modality matrix for modality preference. Rows two and three are from rule one. Rows four and five are from rule two. The last row is from rule three.

The redundant multi-media information should be eliminated according to the principle of conciseness. Therefore, we have rule four to keep conciseness as shown in Table 6.9.

As mention before, we follow the principle of focus to select modality in rule five (Table 6.10) so as to satisfy the visual space constraint. Since we cannot display all multi-media information (e.g. the attraction photograph)

Rule 4:

If selected URL includes photograph of the same concept-value,

Choose the selected URL only (Conciseness)

Table 6.9: The heuristic rules to keep conciseness.

on the limited visual space, we will tell the user that there are other related multi-media remained and has not been shown yet. He/she can decide to look at them or not as another request.

Rule 5:

If there is no visual space, (Visual space constraint)

Keep the selected modality for information in focus but (Focus)

Choose speech only for the less important information one by one

Until the visual space constraint is satisfied.

Ask user want to have more information or not.

Table 6.10: The heuristic rules for visual space constraint and focus.

The heuristic rules need to be adapted to different task goals and dialog acts. This is because different possible concepts (e.g. in ASK_TOUR task goal, tour name will be a possible concept that will not occur in ASK_ATTRACTION task goal). However, we still need to consider the four principles that we have mentioned before.

6.2.3 Temporal Coordination for Synchronization

For temporal coordination, map (with pointing or circling) and photograph should be shown at the same time of the spoken reference. Due to the practical necessary, URL needs some loading time (around one second even used buffer). As the aim is to complete loading at the same time of the spoken reference is

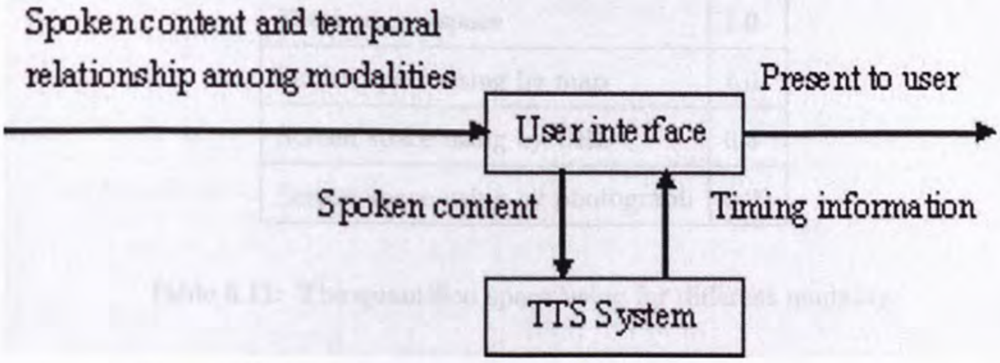


Figure 6.3: The flow to get the timing information of each modalities.

spoken out. We start to load the URL a little bit before the time of the spoken reference. Figure 6.3 illustrates the flow about getting the timing information of each modalities. We send the spoken content, the relationship between reference and modality to the user interface. The user interface sends the spoken content to TTS system and get the timings of different modalities. Then, the user interface presents all information.

6.2.4 Physical Layout

For the physical layout, one of the problems is how to know that visual space is enough or not. The visual space constraint is closely related to the implementation of the user interface. To simplify the checking of visual space constraint, we quantify the space needed for different modalities in Table 6.11. The total available visual space is set to one. We find that a map is often large so the space need is one also. For URL, we set to 0.5 to reflect the ability of scroll down the web browser to read the content by users. For photograph, we set to 0.25 for each and we scale down the large photograph if necessary.

Although this quantification is an ad hoc setting and implementation-orientation, the setting can still reflect the usual usage of space for differ-

Total screen space	1.0
Screen space using by map	1.0
Screen space using by URL	0.5
Screen space using by photograph	0.25

Table 6.11: The quantified space using for different modality.

ent modalities. We have built a task-dependent multi-media database. The database labeled with the information that whether a locative information has an explicit physical boundary or not on the map. The relative coordinates of pointing or circling of map will also be stored in the multi-media database. Moreover, the database stores the association URL of description and photograph included in an URL.

6.2.5 Deictic Term

Deictic term has been commonly used in the spoken content. When we refer to the study in Chapter 4, deictic terms are preferred by URL and speech combination. We will only use deictic term if there is one URL in the response but not multiple urls. This is because we try to avoid any mis-understand when user resolves the deictic term. For the modalities combination with map, we will use deictic term to indicate the pointing or circling on map.

6.2.6 Example

We apply the heuristic rules to determine suitable modalities for given response semantic frame (Table 6.12). A semantic frame includes the task goal, dialog act and concepts with textual values. The first column in Table 6.12 is intention and concept. The second column is their corresponding value(s). The numeric value at the beginning of a concept-value is the importance of the value where one is the most important (focusing) value. If we select multiple modalities, the related multi-media will be retrieved from the database. After that, we determine the use of deictic term giving selected modalities.

Response Semantic Frame	
Task_goal:	ASK_ATTRACTION
Dialog_act:	SUGGEST
Attraction_description:	2_見證了香港及九龍半島...
Building:	1_鐘樓
Expected response	
我介紹你去鐘樓．相關的資料就係哩個網頁上 [show the Clock Tower URL].	

Table 6.12: Examples of response semantic frame.

The building concept is a kind of physical object information. The attraction description is a kind of descriptive information and the number of Chinese characters is more than four. According the Table 6.8, we choose photograph for building and URL for attraction description.

We retrieve the building's photograph and the URL included the attraction description fro multi-media database. We find that the URL already included the building's photograph. According to the rule four, we modify the modality choice of building from photograph to speech only. Next, we check the visual space using. We only need 0.5 (one URL) which is less than 1.0. Rule five will not be fired.

We need to determine using deictic term or not for the URL about attraction description. Since only one URL will be shown in this response, deictic term will be used for the attraction description concept.

6.3 Spoken Content Generation

We have developed a preliminary system to generate the response so as to demonstrate how to realize a multi-modal response. As mentioned before, our focus is the ASK_ATTRACTION task goal with two dialog acts that are INFORM and SUGGEST, and three information types including locative, descriptive and physical object information. We get the selected modalities, multi-media information and using deictic term or not from before process. Then, we create a sequence of words to present the response using a set of spoken content generation templates and deictic term generation templates.

We have made reference to Yip [54] to design the spoken content generation templates as shown in Table 6.13. Each spoken content generation template may include one or more verbalization options(separate by a vertical bar). We select one of them randomly to make the response to be more interesting.

Each verbalization option includes specified concept (denoted by '#'). In a multi-modal response, we may use deictic term which is not included in Yip [54]. Therefore, we extended the work and designed deictic term generation templates as shown in Table 6.14. If the concept does not use any deictic term, its value should be obtained from the response semantic frame. Otherwise, the value should refer to its corresponding deictic term generation templates. The letters 'w', 'p' and 'm' stand for URL, photograph and map respectively. For example, when the ATTRACTION_DESCRIPTION is presented by URL and decided to use deictic term, the deictic term generation template ATTRACTION_DESCRIPTION will be selected. The words that associate with 'w' will be selected ('哩個網頁上').

A set of template selection rules are developed to select suitable templates as shown in Table 6.15. As shown in Table 6.7, we need to handle multiple concept-values. Therefore, we fire some selection rules continuously for handling the various number of concept-values. More than one template selection

Spoken Content Generation Templates:

Template label: SUGGEST_BUILDING
Template contents: 我介紹你去 <#BUILDING>|不如你去<#BUILDING>.

Template label: MORE_BUILDING
Template contents: ,<#BUILDING>.

Template label: PROVIDE_DESCRIPTION_DEICTIC
Template contents: 相關的資料就係<#ATTRACTION_DESCRIPTION>.
|<#ATTRACTION_DESCRIPTION>就有相關的資料.

Template label: PROVIDE_DESCRIPTION
Template contents: 這個景點<#ATTRACTION_DESCRIPTION>.

Table 6.13: Examples of spoken content generation templates.

Deictic Term Generation Templates:

Template label: BUILDING
Template content: (w,p):哩個建築物, m:哩到

Template label: ATTRACTION_DESCRIPTION
Template content: w:哩個網頁上, p:哩副相, m:哩張地圖上

Table 6.14: Examples of deictic term generation templates.

rules will be fired if they are satisfied.

Response Dialog State: ASK_ATTRACTION, SUGGEST	
Templates selection rules	**Selected templates**
Concept BUILDING present	SUGGEST_BUILDING
Concept BUILDING with different values present (repeat this rule until no more BUILDING values)	MORE_BUILDING
Concept ATTRACTION_DESCRIPTION present and using deictic term	PROVIDE_DESCRIPTION_DEICTIC
Concept ATTRACTION_DESCRIPTION present and not using deictic term	PROVIDE_DESCRIPTION

Table 6.15: Examples to illustrate template selection.

We illustrate the spoken content generation in Table 6.16 for the given response semantic frame in Table 6.12.

6.4 Chapter Summary

In this chapter, we have presented several examples which required the use of multiple modalities. We generalize some principles including *Modality Preference, Preciseness, Conciseness* and *Focus.* Heuristic rules are developed for multi-modal response generation implementation purpose. A preliminary system is developed to show the generation of multi-modal responses. We have borrowed the idea from Yip [54] and extended it to capable of generating suitable deictic term in the system developed.

After applying the heuristic rules and determine the use of deictic term(s) Concept: BUILDING, Selected modality: Speech only, Using deictic: No Concept: ATTRACTION_DESCRIPTION, Selected modality: Url, Using deictic: Yes
Response Dialog State: ASK_ATTRACTION, SUGGEST Associated Spoken Content Generate Templates (Excerpt): Option 1: SUGGEST_BUILDING Option 2: MORE_BUILDING Option 3: PROVIDE_DESCRIPTION_DEICTIC Option 4: PROVIDE_DESCRIPTION Template Selection Condition 1: Concept Building present Selected Template: SUGGEST_BUILDING Template Selection Condition 3: Concept ATTRACTION_DESCRIPTION present and using deictic term Selected Template: PROVIDE_DESCRIPTION_DEICTIC Selected Deictic Term Generation Template's content: 哩個網頁上 Generated Response: 我介紹你去鐘樓．相關的資料就係哩個網頁上 [show the URL of the Clock Tower]

Table 6.16: Illustration of spoken content generation templates and deictic term generation templates selection process. The URL will be shown at the same time of corresponding deictic term is spoken out.

Chapter 7

Conclusions and Future Work

7.1 Summary

In this thesis, we have described the preliminary development of a multi-modal response generation in the Hong Kong tourism domain. The multi-media information required appropriate use of multi-modal to convey a message effectively. We have collected a corpus of multi-modal dialogues Wizard-of-Oz setup so as to understand human behavior on multi-modal response. Our study is based re-designed responses by an expert. The finding included identification of some preferable multi-modal combinations. URL with speech and any modalities combination including map prefer the use of deictic term. INFORM and SUGGEST dialog acts in ASK_ATTRACTION task goal prefer the use of multi-modal. Locative, descriptive and physical object information prefer the use of multiple modalities.

We have increased the expressive power of computer by developing a text-to-audiovisual speech (TTVS) system, LinLin. LinLin is a real-time Cantonese TTVS system. We investigated the real-time facial animation techniques using blending process among several static viseme models. LinLin included the audio-visual synchronization techniques using linear interpolation to control the facial animation by acoustic signal. We found that the TTVS system can improve the accuracy in user perception experiment from 80.6% to 89.0%. We want to further increase the computer expressive power. Therefore, an extension system of LinLin has been developed to support acoustic expression, head nod and shake.

We selected some typical multi-modal response examples, that cover most of the cases of multi-modal responses, and see how to select suitable modalities to convey the message. We define four principles that are modality preference, preciseness, conciseness and focus. Based on the principles and study from collected multi-modal responses, we have designed five heuristic rules to select modalities. A preliminary system, including the talking head Windy, has been implemented to demonstrate the generation of multi-modal responses.

7.2 Contributions

In this thesis, the following contributions are made to the research area of multi-modal response generation:

- In the multi-modal dialogue collection, we have found:
  - Several some preferable multi-modal combinations. For example, map usually occurs with mouse gesture, URL and photo are fewer,
  - The use of deictic terms is depending on the modalities combinations. The modalities combinations including URL and speech, or map and other modalities, prefer to use deictic terms,
  - The use of multiple modalities will frequently occur in the task goal of ASK_ATTRACTION together with the dialog acts SUGGEST and INFORM,
  - The use of multi-modal responses will frequently occur in locative, descriptive and physical object information.

- In the TTVS area, we have designed and implemented a real-time Cantonese TTVS system capable to:
  - Provide a method to perform a real-time facial animation with blending technique,
  - Provide a method to generate synthetic audio-visual speech,
  - Extend to support multi-modal response generation.
- In the modality selection, we have designed some heuristic rules to handle the most common multi-modal responses with the considerations on:
  - Modality Preference: Choose the most preferable modality for each information type if possible,
  - Preciseness: Sharply define the reference boundary,
  - Conciseness: Minimize the use of words and unnecessary multimedia information,
  - Focus: First priority of modality selection is giving to the concept-value in focus and then other important concept-value,
  - Visual space constraint of the user interface.
- A preliminary system has implemented based on the TTVS techniques and heuristic rules mentioned above.

We have developed the TTVS techniques to provide a tool to maximize the expressive power which is the original thesis goal. We cover the re-designed responses to design the guidelines of modality selection in the Hong Kong tourism domain. Both of them contribute to the thesis goal.

7.3 Future work

Deeper understanding of human behavior on multi-modal responses is necessary for simulating human-like multi-modal response. The work listed below need to be done to support the understanding of human behavior:

- Developing an automatic annotation tool

  The current dialogues collected are annotated manually so it is very time consuming to perform annotation. A tool should be developed to reduce the human effort on annotation.

- Exploring the annotated data

  Many knowledge still hide in the collected multi-model response. For example, Oviatt [55] defined nine temporal relations between modalities. Do all modality combinations have the same temporal relations? Can we observe the same modality usage pattern in other task goals? Will there any difference between the redesigned responses by two different experts?

- Studying the expression in multi-modal data

  The expert in the multi-modal response collection used many facial expressions and emphasis. It is an interesting task to know when the facial expression or emphasis should be used. Since our TTVS system can support some facial expressions and emphasis, we can simple incorporate the founding into the current TTVS system.

We can further improve the multi-modal response generation system by improving the expressive power of computer. Several improvements are:

- Incorporating the tone information in TTVS system.

  According to Hoole [66], tone three in Mandarin shows a different tongue, jaw and head position with other tone. In our TTVS system, we do not consider any tonal effect on the talking head. More precise facial expression can be generated with the consideration on the tonal effect.

- Importing skeleton to Windy

  In computer animation, if a model includes a skeleton model, all vertices will associate to the skeleton. In order to move a finger to point something, We can move the related bones only. Then, the vertices associated with it will be changed accordingly. The skeleton allows us to have more body gestures in multi-modal dialogue system.

- Using human face photo as a texture
  We can incorporate a human face photo on the talking head model to improve the realization of TTVS system. However, facial adaption is necessary before mapping of a human face photo. Facial adaption means that modify the genesis face to fit on a specific human shape. Moreover, the voice need to fit the adapted face to increase the realistic [67]. We do not expect a young lady face will have a low masculine voice.

Our ultimate goal is to build a multi-modal dialogue system. A multi-modal dialogue system includes different task goals. The next step is generalizing the modality selection to various task goals. One possible way as follows:

- Improving the modality selection by lattice searching
  Modality choices and concepts can form a matrix. We have already discovered some probabilities among information type, intention and modality. It is possible to use them as the cost for lattice searching. However, we need to understand how the cost related to the probabilities.

Spoken content generation is very important before generation of a multi-modal response because most responses include more or less spoken content. Therefore, reduction of the manual work on generating spoken content and improvement of the quality of spoken content are necessary. Therefore, possible work in this direction includes:

- Generating the spoken content generation templates automatically
  Our preliminary system only includes a few numbers of spoken content

generation templates. Hand design all necessary templates is nearly impossible and time consuming. Some statistical approaches in text generation can help to write the necessary templates.

- Investigating the use of anaphoric reference
  Anaphora is the use of a pronoun instead of repeating a word. It can increase the conciseness of response by reducing the repeated content. Currently, we have not fully utilized the use of anaphora. It is possible to improve the template-based approach for spoken content to support the use of anaphora.

Our preliminary system is only a part of multi-modal dialogue system. In order to handle multi-modal input, we should integrate a multi-modal input understanding system and dialogue manager to understand user's request and manage the flow of conversation.

Appendix A

XML Schema for M3 Markup Language

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" attributeFormDefault="unqualified" >
  <xs:element name="M3ML" >
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="group" maxOccurs="unbounded" />
        <xs:element ref="sequence" maxOccurs="unbounded" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
```

```
<xs:element name=“group” >
 <xs:complexType>
  <xs:sequence>
   <xs:element ref=“sequence”  maxOccurs=“unbounded” />
   <xs:element ref=“interpretation”  maxOccurs=“unbounded” />
  </xs:sequence>
 </xs:complexType>
</xs:element>
<xs:element name=“sequence” >
 <xs:complexType>
  <xs:sequence>
   <xs:element ref=“group”  maxOccurs=“unbounded” />
   <xs:element ref=“interpretation”  maxOccurs=“unbounded” />
  </xs:sequence>
 </xs:complexType>
</xs:element>
<xs:element name=“interpretation” >
 <xs:complexType>
  <xs:sequence>
   <xs:element ref=“content”  maxOccurs=“unbounded” />
  </xs:sequence>
  <xs:attribute name=“id”  type=“xs:string”  use=“required” />
```

```
<xs:attribute name=“medium” >
 <xs:simpleType>
  <xs:restriction base=“xs:string” >
   <xs:enumeration value=“acoustic” />
   <xs:enumeration value=“tactile” />
   <xs:enumeration value=“visual” />
  </xs:restriction>
 </xs:simpleType>
</xs:attribute>
<xs:attribute name=“mode” use=“required” >
 <xs:simpleType>
  <xs:restriction base=“xs:string” >
   <xs:enumeration value=“speech” />
   <xs:enumeration value=“pointing” />
   <xs:enumeration value=“circling” />
   <xs:enumeration value=“highlight” />
   <xs:enumeration value=“blinking” />
   <xs:enumeration value=“webpage” />
   <xs:enumeration value=“photo” />
   <xs:enumeration value=“map” />
  </xs:restriction>
 </xs:simpleType>
</xs:attribute>
<xs:attribute name=“start” type=“xs:integer” />
<xs:attribute name=“duration” type=“xs:integer” />
<xs:attribute name=“time-ref-uri” type=“xs:anyURI” />
<xs:attribute name=“offset-to-start” type=“xs:integer” />
```

```
<xs:attribute name=“time-ref-anchor”  default=“start” >
  <xs:simpleType>
    <xs:restriction base=“xs:string” >
      <xs:enumeration value=“start” />
      <xs:enumeration value=“end” />
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
<xs:attribute name=“semantic-to-start”  >
  <xs:simpleType>
    <xs:restriction base=“xs:string” >
      <xs:enumeration value=“just_before” />
      <xs:enumeration value=“just_after” />
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
<xs:attribute name=“semantic-to-duration” >
  <xs:simpleType>
    <xs:restriction base=“xs:string” >
      <xs:enumeration value=“same_duration” />
      <xs:enumeration value=“longer” />
      <xs:enumeration value=“shorter” />
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
<xs:attribute name=“self-x-coordinate”  type=“xs:integer”  />
<xs:attribute name=“self-y-coordinate”  type=“xs:integer”  />
<xs:attribute name=“coordinate-ref-uri”  type=“xs:anyURI”  />
<xs:attribute name=“x-offset-locate”  type=“xs:integer”  />
<xs:attribute name=“y-offset-locate”  type=“xs:integer”  />
```

```
    <xs:attribute name=“coordinate-ref-anchor” >
      <xs:simpleType>
        <xs:restriction base=“xs:string” >
          <xs:enumeration value=“top_left” />
          <xs:enumeration value=“top_right” />
          <xs:enumeration value=“bottom_left” />
          <xs:enumeration value=“bottom_right” />
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    <xs:attribute name=“semantic-coordinate” >
      <xs:simpleType>
        <xs:restriction base=“xs:string” >
          <xs:enumeration value=“above_top” />
          <xs:enumeration value=“below_top” />
          <xs:enumeration value=“above_bottom” />
          <xs:enumeration value=“below_bottomt” />
          <xs:enumeration value=“left_leftside” />
          <xs:enumeration value=“right_leftside” />
          <xs:enumeration value=“left_rightside” />
          <xs:enumeration value=“right_rightside” />
          <xs:enumeration value=“inside” />
          <xs:enumeration value=“outside” />
        </xs:restriction>
      </xs:simpleType>
    </xs:attribute>
    </xs:complexType>
  </xs:element>
    <xs:element name=“content” type=“xs:string” />
</xs:schema>
```

Figure A.1: The full XML schema definition of M3ML.

Appendix B

M3ML Examples

Example 1

```
<?xml version=“1.0” standalone=“yes ”?>
<M3ML xmlns:xs=“http://www.w3.org/2001/XMLSchema”
xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
xsi:noNamespaceSchemaLocation= =“c:\xtemp\m3ml.xsd”>
  <M3ML:group>
    <M3ML:interpretation M3ML:medium=“map” M3ML:mode=“visual”
     id=“m1”>
      <content>tsimshatsui.jpg</content>
    </M3ML:interpretation>
    <M3ML:interpretation M3ML:medium=“acoustic”
     M3ML:mode=“speech” id=“s1” M3ML:time-ref-uri=“m1”
     M3ML:time-ref-anchor=“end” M3ML:semantic-to-start=“just_after”>
      <content>你的酒店就係哩度 </content>
    </M3ML:interpretation>
  <M3ML:interpretation M3ML:medium="acoustic"
   M3ML:mode="speech"  id="s2"  M3ML:time-ref-uri="s1"
   M3ML:time-ref-anchor="end"  M3ML:offset-to-start=0>
    <content>而碼頭呢就係哩度 </content>
  </M3ML:interpretation>
  <M3ML:interpretation M3ML:medium="pointing"
   M3ML:mode="tactile"  M3ML:time-ref-uri="s1"
   M3ML:time-ref-anchor="end"  M3ML:semantic-to-start="just_after"
   M3ML:coordinate-ref-uri="m1"
   M3ML:semantic-coordinate="inside"/>
  </M3ML:group>
</M3ML>
```

Figure B.1: Example one of M3ML annotation.

Example 2

```
<?xml version=“1.0” standalone=“yes ”?>
<M3ML xmlns:xs=“http://www.w3.org/2001/XMLSchema”
xmlns:xsi=“http://www.w3.org/2001/XMLSchema-instance”
xsi:noNamespaceSchemaLocation= =“c:\xtemp\m3ml.xsd”>
   <M3ML:group>
     <M3ML:interpretation M3ML:medium=“acoustic”
      M3ML:mode=“speech” id=“s1”>
       <content>你可以睇下哩度 </content>
     </M3ML:interpretation>
     <M3ML:interpretation M3ML:medium=“map” M3ML:mode=“visual”
      M3ML:time-ref-uri=“s1” M3ML:time-ref-anchor=“end”
      M3ML:semantic-to-start=“just_after”>
       <content>stanley.jpg</content>
     </M3ML:interpretation>
   </M3ML:group>
</M3ML>
```

Figure B.2: Example two of M3ML annotation.

Appendix C

Domain-Specific Task Goals in the Hong Kong Tourism Domain

ASK_ATTRACTION	
Definition:	The dialog turn contains content related to ATTRACTION information seeking
Example:	‘有無關於香港歷史的博物館?’ ‘有，你可以去香港歷史博物館.’
ASK_FEE	
Definition:	The dialog turn contains content related to FEE information seeking
Example:	‘請問博物館入場費幾多?’ ‘你需要成人票還是小童票?’
ASK_INFO	
Definition:	The dialog turn contains content related to general information seeking
Example:	‘香港有幾多區？’ ‘香港有十八區.’

ROUTE_SEEKING	
Definition:	The dialog turn contains content related to ROUTE or transportation information seeking
Example:	‘點樣去哩度?’ ‘你想塔咩野類型的交通工具?’
ASK_SUGGEST	
Definition:	The dialog turn is related to a suggestion
Example:	‘我無咩行山經驗，但又想試下．有咩好建議?’ ‘我建議你行山頂先．這是最短路線．.’
ASK_TOUR	
Definition:	The dialog turn is related to TOUR information seeking
Example:	‘有無大嶼山的旅行團?’ ‘這是大嶼山旅行團的資料．’
RESERVATION	
Definition:	The dialog turn is related to RESERVATION
Example:	‘要三張船飛.’ ‘我幫你訂左 三張船飛啦.’

Table C.1: Definition and Examples of seven task goals corresponding to the Hong Kong tourism domain

Appendix D

Dialog Acts for User Request in the Hong Kong Tourism Domain

APOLOGY

Definition: With an APOLOGY the tourist solely signals regret to wizard.

Example: ‘對不起．’

BACKCHANNEL

Definition: With a BACKCHANNEL the tourist solely signals that he is still following the conversation, without really take the turn himself.

Example: ‘嗯嗯．’

CLOSE

Definition: With a CLOSE the tourist says good bye or certain phrases to the wizard, thereby closing the dialog.

Example: ‘再見.’

COMMIT

Definition: With a COMMIT the tourist explicitly commits him/herself to do one or more specific actions to the wizard.

Example: ‘我想遲些再去.’

CONFIRM

Definition: With an utterance expressing a CONFIRM the tourist wraps up the result of the negotiation (or a part thereof). This is done by repeating parts of the completed task.

Example: ‘我睇到個網頁啦去.’

DEFER

Definition: The tourist explicitly suggests or announces the interruption of the topic currently dealt with in the dialog.

Example: ‘等我想想先.’

FEEDBACK_NEGATIVE

Definition: With an utterance expressing FEEDBACK_NEGATIVE the tourist reacts to a contribution of the dialog partner in a negative way. A FEEDBACK_NEGATIVE can signal rejection or dislikes of the contents of a previous contribution or it can express a negative answer to a yes/no question.

Example: ‘我想佢太貴啦.’

FEEDBACK_POSITIVE

Definition: With an utterance expressing FEEDBACK_POSITIVE the tourist reacts to a contribution of the dialog in a positive way. A FEEDBACK_POSITIVE can signal acceptance of the content of a previous contribution or it can express a positive answer to a yes/no question.

Example: ‘佢睇來幾好啊.’

GIVE_PREFERENCE

Definition: With a GIVE_PREFERENCE the tourist signals his/her preference on the content of previous conversation. It includes the preference on activities, venue, interests and reservation. immediately preceding and/or following the context.

Example: ‘我鐘意室內活動多點.’

GIVE_REASON

Definition: A dialog segment is labeled with GIVE_REASON if it contains the reason/justification/motivation for a statement, made in the immediately preceding and/or following the context.

Example: ‘我驚星期日人多濟迫.’

GREET

Definition: GREET is used for all kinds of initial greetings.

Example: ‘哈囉.’

INFORM

Definition: The label INFORM is reserved for cases where none of the categories apply. If not enough information is available in the content to label the given dialog segment as any of those it can be labeled as INFORM.

Example: ‘我會早點去架啦.’

INTRODUCE

Definition: The utterance contains information about the speaker, e.g. his/her name, country, time and trip planning.

Example: ‘我係由美國黎.’

REQUEST_ACTION	
Definition:	The tourist explicitly requests to perform on or more specified actions. For example, request for reservation service.
Example:	‘我想睇下哩兩個地方的相片 ?’
REQUEST_COMMENT	
Definition:	With an utterance expressing a REQUEST_COMMENT the tourist explicitly asks the wizard to make a confirmation or clarification. It is often used to yield the turn; in that case it prompts the dialog partner to respond.
Example:	‘你明 唔明我講咩 ?’
REQUEST_INFO	
Definition:	With an utterance expressing a REQUEST_INFO the tourist asks the dialog partner to present information or more information about something that has already been either explicitly or implicitly introduced into the discourse.
Example:	‘塔山頂覽車時我應該注意什麼 ?’
REQUEST_SUGGEST	
Definition:	With an utterance expressing a REQUEST_SUGGEST the tourist asks the dialog partner to make a suggestion or proposal.
Example:	‘有邊度値得我去參觀.’
THANK	
Definition:	With an utterance expressing a THANK the tourist expresses his gratitude to the dialog partner.
Example:	‘唔該晒 .’

Table D.1: Definition and Examples of eighteen user request dialog acts corresponding to the Hong Kong tourism domain

Appendix E

Dialog Acts for System Response in the Hong Kong Tourism Domain

APOLOGY

Definition: With an APOLOGY the wizard solely signals regret to tourist.

Example: ‘對唔住，我顯示錯左網頁.’

BACKCHANNEL

Definition: With a BACKCHANNEL the wizard solely signals that he is still following the conversation, without really take the turn himself.

Example: ‘嗯嗯.’

CLOSE

Definition: With a CLOSE the wizard says good bye or certain phrases to the tourist, thereby closing the dialog.

Example: ‘再見.’

COMMIT

Definition: With a COMMIT the wizard explicitly commits him/herself to do one or more specific actions to the tourist.

Example: ‘我會顯示地圖比你睇.’

CONFIRM

Definition: With an utterance expressing a CONFIRM the wizard wraps up the result of the negotiation (or a part thereof). This is done by repeating parts of the completed task.

Example: ‘你想買手信，明白啦.’

DEFER

Definition: The wizard explicitly suggests or announces the interruption of the topic currently dealt with in the dialog.

Example: ‘請等一陣.’

FEEDBACK_NEGATIVE

Definition: With an utterance expressing FEEDBACK_NEGATIVE the wizard reacts to a contribution of the dialog partner in a negative way. A FEEDBACK_NEGATIVE can signal rejection or dislikes of the contents of a previous contribution or it can express a negative answer to a yes/no question.

Example: ‘我幫你唔到.’

FEEDBACK_POSITIVE

Definition: With an utterance expressing FEEDBACK_POSITIVE the wizard reacts to a contribution of the dialog in a positive way. A FEEDBACK_POSITIVE can signal acceptance of the content of a previous contribution or it can express a positive answer to a yes/no question.

Example: ‘好呀.’

GIVE_REASON

Definition: A dialog segment is labeled with GIVE_REASON if it contains the reason/justification/motivation for a statement, made in the immediately preceding and/or following the context.

Example: ‘哩度好適合遊水.’

GREET

Definition: GREET is used for all kinds of initial greetings.

Example: ‘早晨.’

INFORM

Definition: The label INFORM is reserved for cases where none of the categories apply. If not enough information is available in the content to label the given dialog segment as any of those it can be labeled as INFORM.

Example: ‘哩個博物館係關於香港文化同歷史.’

INIT

Definition: The dialog act INIT is used to describe utterance where the topic of the interaction to follow is introduced.

Example: ‘等我向你介紹香港的購物中心先.’

INTRODUCE

Definition: The utterance contains information about the speaker, e.g. his/her name and nature.

Example: ‘歡迎使用本系統.’

OFFER

Definition: The speaker explicitly offers to perform one or more specified actions.

Example: ‘有咩野可以幫到你？’

REQUEST_ACTION

Definition: The wizard explicitly requests to perform on or more specified actions.

Example: ‘你可唔可以畫比我睇?’

REQUEST_COMMENT

Definition: With an utterance expressing a REQUEST_COMMENT the wizard explicitly asks the tourist to comment on a proposal. It is often used to yield the turn; in that case it prompts the dialog partner to respond.

Example: ‘咁海洋公園好不好?’

REQUEST_INFO

Definition: With an utterance expressing a REQUEST_INFO the wizard asks the dialog partner to present information or more information about something that has already been either explicitly or implicitly introduced into the discourse. The information is neither comment nor preference.

Example: ‘請問點稱呼?’

REQUEST_PREFERENCE

Definition: With an utterance expressing a REQUEST_PREFERENCE the wizard explicitly asks the tourist to give a preference about something.

Example: ‘你鍾意咩野類型活動?’

SUGGEST

Definition: With an utterance expressing a SUGGEST the speaker proposes an explicit instance or aspect of the negotiated topic.

Example: ‘我建議你去鐘樓.’

THANK

Definition: With an utterance expressing a THANK the tourist expresses his gratitude to the dialog partner.

Example: ‘唔該.’

Table E.1: Definition and Examples of twenty dialog acts corresponding to the Hong Kong tourism domain

Appendix F

Information Type and Concepts

Information type	Concepts	Examples
Locative	LOCATE_DISTRICT, LOCATE_AREA, LOCATE_STREET	(位於)中西區, (位於)尖沙咀附近
Descriptive	DESCRIPTION	地道的，最大的
Physical Object	BUILDING, WARE	三棟屋博館, 鐘樓 美利樓
Nominal	ATTRACTION_NAME, STREET, DISTRICT, AREA, CITY, REGION	赤柱市場, 花墟, 山頂, 時代廣場
Temporal	VISIT_DAY, VISIT_REFERENCE_DAY, VISIT_ABSOLUTE_DAY, VISIT_DATE, VISIT_PART_OF_DAY, VISIT_ABSOLUTE_TIME	早上, 星期日
Abstract	BUILDING_CLASS, PLACE_CLASS, GARDEN_CLASS, ANIMAL_CLASS SHOP_CLASS	商場, 海灘, 海岸公園

Table F.1: Six information types and associated concepts. The word inside the brackets is not part of the value, which are shown to illustrate the meaning of that value.

Appendix G

Concepts

BUILIDNG_CLASS ⟶ 商場 | 購物商場 | 大型商場 | 電腦商場 | 博物館 | 酒店 | 主要博物館 | 大樓 | 樓宇 | 古跡

PLACE_CLASS ⟶ 郊遊地點 | 沙灘 | 海灘 | 大平原 | 郊區 | 購物區 | 地方 | 島 | 漁村 | 市場 | 景點

PARK_CLASS ⟶ 海岸公園 | 郊野公園

ANIMAL_CLASS ⟶ 海豚 | 雀仔 | 動物

SHOP_CLASS ⟶ 賣金魚舖頭 | 海味的舖頭 | 相機專門店 | 舖頭 | 商戶 | 食肆

ANIMAL ⟶ 中華白海豚 | 鸚鵡

PARK ⟶ 公園 | 海洋公園 | 香港迪士尼 | 城寨公園

ATTRACTION_NAME ⟶ 花墟 | 太平山頂 | 蘭桂坊 | 淺水灣 | 星光大道 | 機場 | 金紫荊廣場 | 黃金海岸

PIER ⟶ 中環碼頭 | 荃灣碼頭 | 中環港外線碼頭 | 碼頭

CITY ⟶ 香港 | 澳門 | 深圳

REGION ⟶ 九龍 | 新界 | 香港島

DISTRICT	⟶	油尖旺區 | 中西區 | 西貢 | 元朗 | 沙田 | 屯門 東區 | 離島 | 九龍城 | 葵青 | 觀塘 | 北區 | 深水 | 深水土步 | 深水步 | 南區 | 大埔 | 荃灣 | 灣仔 | 黃大仙
STREET	⟶	鴨寮街 | 太子街 | 雀仔街 | 金魚街 | 女人街 | 玉器街 | 花園街 | 海鮮街 | 廟街 | 通菜街 | 赤柱大街 | 鳥街 | 西洋菜街 | 廣東道 | 上海街 | 怡和街
AREA	⟶	油麻地 | 尖沙咀 | 旺角 | 赤柱 | 中環 | 香港仔 | 大嶼山 | 東坪洲 | 索罟灣 | 南丫島
CULTUREAL_CENTER	⟶	香港文化中心 | 鐘樓 | 藝術博物館 | 美利樓 |
SHOPPING_CENTER	⟶	文化中心 | 科學館 | 香港藝術博物館 | 海港城 | 文化中心 | 科學館 | 香港藝術博物館 | 海港城 |
MUSEUM	⟶	大佛 | 凌霄閣 | 海洋中心 | 觀音廟 | 寶蓮寺 | 天后廟 | 亭 | 海運中心 | 香港歷史博物館 | 車公廟 |
BUILDING	⟶	歷史博物館 | 禮賓府 | 山頂廣場 | 新城市廣場
TEMPLE	⟶	觀音廟 | 寶蓮寺 | 天后廟 | 車公廟
SHOP	⟶	榕記 | 珍寶王國 | 小熊國
NUMBER_BUILDING	⟶	NUMBERVALUE 至 NUMBERVALUE 間 |NUMBERVALUE 間
PRICE_RANGE	⟶	比較平的 | 比較貴的 | 比較便宜的 | 平 | 貴
WARE	⟶	衣服 | 電腦 | 電子產品 | 電話 | 電器 | 相機 | 衫 | 手信 | 金飾玉器 | 玩意
POSITION	⟶	隔離 | 上面 | 近住 | 附近 | 傍邊 | 一帶 | 近

LOCATE_STREET ⟶ (位於) STREET

LOCATE_TRANSPORTATION_STATION ⟶ (位於) TRANSPORTATION_STATION

LOCATE_REGION ⟶ (位於) REGION

LOCATE_AREA ⟶ (位於) AREA

LOCATE_DISTRICT ⟶ (位於) DISTRICT

LOCATE_UNIVERSITY ⟶ (位於) 香港中文大學 | 香港大學 | 城市大學 | 香港科技大學 | 香港理工大學 | 嶺南大學

LOCATE_BUILDING ⟶ (位於) BUILDING

LOCATE_ATTRACTION_NAME ⟶ (位於) ATTRACTION_NAME

DESCRIPTION ⟶ 地道的 | 最大的 | 特別的 | 好多賣花的 | 好熱鬧

ACTIVITY ⟶ 買野 | 週圍行 | 欣賞海景 | 飲野 | 睇魚 | 行山 | 參觀 | 影相 | 行街 | 睇下

EAT_FOOD_STYLE ⟶ 地道小食 | 海鮮 | 中菜 | 西餐 | 中國菜 | 日本菜

OPENING_PART_OF_DAY ⟶ (開放時間係) 早上 | 上晝 | 朝早 | 晚上 | 夜晚 | 下午 | 下晝

OPENING_HOUR ⟶ (開放時間係)NUMBERVALUE 點 |NUBMERVALUE 時 | 通宵

OPENING_DAY ⟶ (開放日子爲) 星期 NUMBERVALUE| 星期日

OPENING_DURATION_HOUR ⟶ (開放時間係)NUMBERVALUE 小時

CLOSING_PART_OF_DAY ⟶ (關門時間係) 晚上 | 夜晚 | 下晝 | 下午

CLOSING_HOUR ⟶ (關門時間係)NUMBERVALUE 點 |NUBMERVALUE 鐘

EVENT_PERIOD	⟶	十一黃金週
VISIT_DAY	⟶	(建議) 星期 NUMBERVALUE| 星期日 (去)
VISIT_REFERENCE_DAY	⟶	(建議) 後日 | 明天 | 明日 | 大後天 | 大後日 (去)
VISIT_ABSOLUTE_DAY	⟶	(建議) 第 NUBMERVALUE 日 (去)
VISIT_DATE	⟶	(建議)NUBMERVALUE 月 NUBMERVALUE 日 (去)
VISIT_PART_OF_DAY	⟶	(建議) 上半日 | 夜晚 | 早上 | 下晝 | 晚上 (去)
VISIT_ABSOLUTE_TIME	⟶	(建議)NUMBERVALUE 點鐘 |NUMBERVALUE 點 (去)
VISIT_REFERENCE _TIME	⟶	(建議) 中午之後 | 晚飯前 | 晚飯後 (去)
PATH	⟶	鴨寮街路徑 | 行山路徑 | 行山路線 | 路線圖 | 路徑
WHERE	⟶	地方 | 其他地方
WHEN	⟶	地點
JOURNEY	⟶	旅行團
NUMBER_JOURNEY	⟶	NUMBERVALUE 個 |NUMBERVALUE 團
TRANSPORTATION_ TOOL	⟶	地鐵 | 纜車 | 電車 | 小巴 | 巴士 | 機場巴士 | 船 | 車 | 山頂纜車 | 通宵小巴 | 火車 | 的士 | 地下鐵路 | 專線小巴 | 紅色小巴 | 輕鐵 | 的士
TRANSPORTATION _COMPANY	⟶	新巴 | 城巴 | 地鐵公司 | 九鐵公司
ROUTE_NUMBER	⟶	NUMBERVALUE 號
DESTINATION_ ATTRACTION_NAME	⟶	(目的地係) ATTRACTION_NAME
DESTINATION_ BUILDING	⟶	(目的地係) BUILDING
DESTINATION_TRANSPORTATION_STATION	⟶	(目的地係) TRANSPORTATION_STATION

<table>
<tr><td>DESTINATION_DISTRICT</td><td>⟶ (目的地係) DISTRICT</td></tr>
<tr><td>DESTINATION_AREA</td><td>⟶ (目的地係) AREA</td></tr>
<tr><td>DESTINATION_STREET</td><td>⟶ (目的地係) STREET</td></tr>
<tr><td>DESTINATION_POINT</td><td>⟶ 街頭 | 街尾</td></tr>
<tr><td>DESTINATION_REGION</td><td>⟶ (目的地係) REGION</td></tr>
<tr><td>DEPARTURE_STREET</td><td>⟶ (從) STREET (出發)</td></tr>
<tr><td>DEPARTURE_AREA</td><td>⟶ (從) AREA (出發)</td></tr>
<tr><td>DEPARTURE_DISTRICT</td><td>⟶ (從) DISTRICT (出發)</td></tr>
<tr><td>DEPARTURE_BUILDING</td><td>⟶ (從) BUILDING (出發)</td></tr>
<tr><td>DEPARTURE_PIER</td><td>⟶ (從) PIER (出發)</td></tr>
<tr><td>DEPARTURE_TRANSPORTATION_STATION</td><td>⟶ (從) TRANSPORTATION_STATION (出發)</td></tr>
<tr><td>DEPARTURE_ATTRACTION_NAME</td><td>⟶ (從) ATTRACTION_NAME (出發)</td></tr>
<tr><td>DEPARTURE_HOUR</td><td>⟶ (出發時間係)NUMBERVALUE 點</td></tr>
<tr><td>DEPARTURE_POINT</td><td>⟶ (係) 街頭 | 街尾 (出發)</td></tr>
<tr><td>TRANSPORTATION_SCHEDULE</td><td>⟶ 船期表 | 班次 | 最後一班船係 NUMBERVALUE 點 | ⟶ . . .</td></tr>
<tr><td>TRANSPORTATION_DURATION_MINUTES</td><td>⟶ NUMBERVALUE 分鐘 |NUMBERVALUE 至 NUMBERVALUE 分鐘</td></tr>
<tr><td>TRANSPORTATION_DURATION_HOUR</td><td>⟶ NUMBERVALUE 個鐘 |NUMBERVALUE 至 NUMBERVALUE 個鐘</td></tr>
<tr><td>MTR_EXIT</td><td>⟶ 地鐵站出口 | 銀行中心出口 | 始創中心出口</td></tr>
<tr><td>TRANSPORTATION_FEE</td><td>⟶ 港幣 NUMBERVALUE 元 | 免費 |NUBMERVALUE 蚊</td></tr>
<tr><td>TICKET_CLASS</td><td>⟶ 成人單程票 | 來回票 | 小童單程票 | 月票</td></tr>
<tr><td>NUMBER_TRANPORSTATION_STATION</td><td>⟶ NUMBERVALUE 個站 |NUMBERVALUE 幾個站</td></tr>
<tr><td>NUMBER_STREET</td><td>⟶ NUMBERVALUE 條街 |NUMBERVALUE 個街口</td></tr>
<tr><td>NUMBERVALE</td><td>⟶ 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 | 十 | 十一 | 十二 |NUMBERVALUE NUMBERVALUE</td></tr>
</table>

TRANSPORTATION_STATION	⟶	火車站 | 巴士總站 | 地鐵站 | 電車站
INDICATION	⟶	指示 | 沿路的指示 | 路牌指示 | 文字顯示架
DIRECTION	⟶	東面 | 南面 | 西面 | 北面
REASON	⟶	信得過 | 最方便 | 唔會呃你 | 快點 | 坐得舒服點
TRANFFIC_JAM_TIME_PERIOD	⟶	上下班 (時間比較塞車)

Table G.1: The concepts and their corresponding value. The characters inside bracket are not part of value, which are shown to illustrate the meaning of concept.

Bibliography

[1] Oviatt, S. L. Ten myths of multimodal interaction. *Communications of the ACM 42(11)*, 1999.

[2] Dix, A., J. Finaly, G. Abowd and R. Beale. Human-Computer Interaction, Second Edition, Section 15.2, pp. 555, 1998.

[3] Bernsen, N. O. Defining a taxonomy of output modalities from an HCI perspective. In: *Rist et al.*, 1997.

[4] Bordegoni, M. G. Faconti, S. Feiner, M. T. Maybury, T. Rist, S. Ruggieri, P. Trahanias and M. Wilson. A Standard Reference Model for Intelligent Multimedia Presentation Systems. In: *Rist et al.*, 1997.

[5] Seneff, S. Response planning and generation in the MERCURY flight reservation system. In the *Proceedings of Computer Speech and Language*, 2002.

[6] Oviatt, S. L. Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems. Advances in Computers, Academic Press, vol. 56, pages 305-341, 2002.

[7] Oviatt, S.L, P.R. Cohen, et al. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art Systems and future research directions. *Human Computer Interaction*, 2000.

[8] Bangalore, S. and M. Johnston. Robust Multimodal Understanding. In the *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.

[9] MIT Project Oxygen http://oxygen.lcs.mit.edu/

[10] Wahlster, W. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In the *Proceedings of the Human Computer Interaction Status Conference*, 2003.

[11] Gustafson, J., N. Lindberg and M., Lundeberg. The August spoken dialogue system. In the *Proceedings of Eurospeech*, 1999.

[12] W3C Multimodal Interaction Activity. http://www.w3.org/2002/mmi/.

[13] VoiceXML Forum's X+V language. http://www.voicexml.org.

[14] SALT. http://www.saltforum.org.

[15] EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft 14 December 2004. http://www.w3.org/TR/emma.

[16] Oviatt, S. L. Multimodal Interfaces. A chapter in *Handbook of Human-Computer Interaction* (ed. by J. Jacko and A. Sears), Lawrence Erlbaum: New Jersey, 2002.

[17] Ezzat, W. and P. Tomaso. MikeTalk: a talking facial display based on morphing visemes. In the *Proceedings of the Computer Animation Conference*, 1998.

[18] Mary, E. F. Deliverable 6.1 State of the art review: Multimodal fission. *COMIC Document*, 2002.

[19] Meng, H., W. L. Yip, O. Y. Mok and S. F. Chan, Natural Language Response Generation in Mixed-Initiative Dialogs using Task Goals and

Dialog Acts. In the *Proceedings of the 8th European Conference on Speech Communication and Technology*, September 2003.

[20] Walker, M., R. Prasad and A. Stent. A Trainable Generator for Recommendations in Multimodal Dialog In the *Proceedings of Eurospeech*, 2003.

[21] Michelle, X. Z. and V. Aggarwal. An Optimization-based Approach to Dynamic Data Content Selection in Intelligent Multimedia Interfaces. In the *Proceedings of the Seventeenth annual ACM symposium on User Interface software and technology*, pages 227-236, 2004.

[22] André, E. The generation of multimedia documents, In: A Handbook of Natural Language Processing: techniques and Applications for the Processing of Language as Text. edited by R Dale, H Moisl, and H Somers, pages 305-327. Marcel Dekker Inc., 2000, URL:http://www.dfki.de/imedia/papers/handbook.ps.

[23] Feiner, S. K. and K. R. Mckeown. Automating the generation of coordinated multimedia explanations. *IEEE Computer 24(10):33-41*. Reprinted in Maybury and Wahlster, 1998.

[24] Coutaz, J., L. Nigay, D. Salber, A. Blandford, J. May and R. Young. Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. In the *Proceedings of INTERACT' 95*, 1995.

[25] Arens, Y., E. Hovy and M. Vossers. On the knowledge underlying multimedia presentations. In: Maybury (1993a), pages 280 - 306. http://www.isi.edu/natural-language/multimedia/knowledge-models.ps, Reprinted in Maybury and Wahlster, 1998.

[26] Malandro, L. A., L. L. Barker, and D. A. Baker. Nonverbal Communication. Random House, New York, Chapter one, Second Edition, 1989. (ISBN: 0-394-36526-7)

[27] Foster, M. E. Corpus-based planning of deictic gestures in COMIC. In the *Proceedings of INLG-04 (Student Session)*, 2004.

[28] Schiel, F., S. Steininger, N, Beringer, U. Turk and S. Rabold Integration of multi-modal data and annotations into a simple extendable form: the extension of the BAS Partitur Format In the *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002*, pp. 39-44., 2002.

[29] Rapp, S. and M. Strube. An Iterative Data Collection Approach for Multimodal Dialogue Systems. In the *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 29-31 May, 2002.

[30] World Wide Web Consortium. http://www.w3.org.

[31] Christoph Müller and Michael Strube MMAX: A Tool for the Annotation of Multi-modal Corpora. In the *2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2001.

[32] Garg, S., B. Martinovski, S. Robinson, J. Stephan, J. Tetreault and D. R. Traum. Evaluation of Transcription and Annotation tools for a Multi-modal, Multi-party dialogue corpus In the *Proceeedings of the Forth International Conference on Language Resources and Evaluation*, pp. 2163 - 2166, 2004.

[33] Oviatt, S., R. Coulston, R. Lunsford. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In the *The Sixth International Conference on Multimodal Interfaces*, October 14-15, 2004.

[34] Kruijff-Korbayová, I., N. Blaylock, C. Gerstenberger, V. Rieser, T. Becker, M. Kaiser, P. Poller, J. Schehl. Presentation Strategies for Flexible Multimodal Interaction with a Music Player. In the *Ninth Workshop On The Semantics And Pragmatics Of Dialogue (Semdial)*, France, June 9-11 2005.

[35] Tao, J. H. and T. I. Tan. Emotional Chinese Talking Head System In the *Proceeding of ICMI*, 2004.

[36] Massaro, D. W. A Computer-Animated Tutor for Spoken and Written Language Learning. In the *Processing of ICMI*, 2003.

[37] Nakamura, S. and E. Yamanoto. Speech-to-lip movement synthesis by maximizing audio-visual joint probability based on the EM algorithm. In the *Journal of VLSI Signal Processing 27*, 2001.

[38] Bailly, G. Audiovisual Speech Synthesis from Ground Truth to Models. In the *Processing of the International Conference on Spoken Language Processing (ICSLP)*, 2002.

[39] Bregler, C., C. Michele and S. Malcolm. VideoRewrite: driving visual speech with audio. In the *Proceedings of SIGGRAPH 97*, 1997.

[40] Karlsson, I., A. Faulkner and G. Salvi. SYNFACE —a talking face telephone. In the *Proceedings of the Eurospeech*, pages 1297-1300, Geneva, Sweden, September 2003.

[41] Terzopoulos, D. and W. Keith. Analysis of Facial Image Using Physical and Anatomical Models In the *Proceedings of Computer Vision*, pages 727-732, 4-7 December 1990.

[42] Roberto, P. Roberto Pockaj's Home Page, Research Activities, http://www-dsp.com.dist.unige.it/ pok/RESEARCH/MPEG/fdpspec.htm.

[43] Cosi, P., A. Fusaor and G. Tisato LUCIA a New Italian Talking-Head Based on a Modified CohenMassaro's Labial Coarticulation Model. In the *Proceedings of EUROPSEECH*, pages 2269-2272, Geneva, Sweden, September 2003.

[44] Bailly, G. Audiovisual Speech Synthesis. In the *Proceedings of ETRW on Speech Synthesis*, Perthshire, Scotland, 2001.

[45] Noh, J., D. Fidaleo and U. Neumann. Animated deformations with radial basis functions. In the *Proceeding of ACM Virtual Reality and Software Technology*, 2000.

[46] Roberto, P. Roberto Pockaj's Home Page, Research Activities, http://www-dsp.com.dist.unige.it/ pok/RESEARCH/MPEG/fapspec.htm.

[47] Zhong, J. L. Flexible face animation using MPEG-4/SNHC parameter streams. In the *Proceedings of Image Processing 1998*, 1998.

[48] Hong, P., Z. Wen and T. Huang. An integrated framework for face modeling, facial motion analysis and synthesis. In the *Proceeding of ACM Conference on Multimedia*, 2001.

[49] Nakamura, S., E. Yamamoto and K. Shikano. Speech-to-lip movement synthesis maximizing audio-visual joint probability based on EM algorithm. In the *Proceeding of IEEE Second Workshop on Multimedia Signal Processing*, 1998.

[50] André, E., W. Finkler, W. Graf, T. Rist, A. Schauder and W. Wahlster. WIP: The automatic synthesis of multimodal presentations. In: Maybury, pages 75-93, 1993.

[51] André, E., T. Rist and J. Müller. Employing AI methods to control the behaviour of animated interface agents. In: *Applied Artificial Intelligence Journal*, 1999.

[52] Stent, A. Content planning and generation in continuous-speech spoken dialog systems. In the *Proceedings of the KI'99 workshop, "May I Speak Freely?"*, September 1999.

[53] Alexandersson, J. B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz and M. Siegel. Dialog Acts in VERBMOBIL-2 Second Edition. In *Verbmobil Report 226, Universitat Hamburg*, DFKI Saarbrucken, Universitat Erlangen, TU Berline, 1998.

[54] Yip, W. L. Natural Language Response Generation in Mixed-Initiative Dialogs. *Master Thesis*, the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong.

[55] Oviatt, S. L., A. DeAngeli and K. Kuhn. Integration and Synchronization of Input Modes during Multi-modal Human-Computer Interact. In the *Proceedings of the Conference on Human Factors in Computing Systems*, pages 415-422, 1997.

[56] Sutcliffe, A. and P. Faraday. Designing presentation in multimedia interfaces. In the *Proceeding of ACM SIGCHI*, 1994.

[57] Gupta, A. and Anastasakos, T. Integration Patterns during Multimodal Interaction. In the *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[58] Picard, R. *Affective Computing*, MIT Press, 1997.

[59] CU Vocal, http://www.se.cuhk.edu.hk/cuvocal/.

[60] Lee, T., H. M. Heng, W. Lau, W. K. Lo and P. C. Ching. Microprosodic Control in Cantonese Text-to-Speech Synthesis. In the *Proceedings of Euorspeech*, volume 4, pages 1855-1858, September 1999.

[61] Virtual Tutorials in Phonology –Hong Kong Word. http://www.cbs.polyu.edu.hk/VTP/hkword/s/s1.htm.

[62] Parke, F. I. and K. Waters. Computer Facial Animation. *A. K. Peters Ltd.*, 1996.

[63] W3C Speech Synthesis Markup Language (SSML) Version 1.0. http://www.w3.org/TR/speech-synthesis/. 2004.

[64] Hoole, P. and F. Hu. Tone-Vowel Interaction in Standard Chinese. In the *Proceedings of the International Symposium on Tonal Aspects of Languages With Emphasis on Tone Languages*, 2004.

[65] HU, X., A. Fourcin, A. Faulkner and J. I. Wei. Speechreading of Words and Sentences by Normally Hearing and Hearing Impaired Chinese Subjects: the Enhancement Effects of Compound Speech Patterns. In the *Speech, hearing and language: work in progress (1996) Volume 9, page 119 - 131*, 1996.

[66] Hoole, P. and F. Hu. Tone-Vowel Interaction in Standard Chinese. In the *Proceeding of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, China, 2004.

[67] Mersiol, M., N. Chateau and V. Maffiolo. Talking Heads: Which Matching between Faces and Synthetic Voices? In the *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, 2002.