Multilingual Multimodal Integration of Sketch and Speech: A Generic Speech Representation Model for Spatial Description

Lee-Na Teh and Alvin W. Yeo
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak
leena.teh@gmail.com, alvin@fit.unimas.my

Abstract

This paper details how multiple languages are accommodated in the multimodal integration of sketch and speech, specifically, in spatial applications. The study encompasses English, Malay, Mandarin, and two under-resourced languages in Malaysia, i.e. Melanau and Iban. The preliminary study revealed that not all spatial terms (prepositions) appear in all languages. Based on these findings, we propose a method to assist in the design and development of multilingual multimodal applications. This method employs a generic representation model for spatial description.

1. Introduction

Human-to-human communication usually involves more than one modality. For instance, these modalities could be pen gesture, speech, eye gaze, hand gestures, lip movements, face expressions, body gestures and so forth. As a computer is "taught" to mimic the human, human-machine communication would preferably involve more than one modality. With today's technologies, multimodal interaction is applied to spatial queries, spatial descriptions, brain storming, robot navigation and augmented reality environments [1] [2] [3].

To perform effective multimodal interaction in the afore mentioned situations, it very much depends on multimodal integration. Multimodal integration, which aligns multiple modalities into a correct order, is very important in delivering more precise information, as well as disambiguates information. For example, a command is given to a person verbally: "Please hand me that book on the table". Apparently, it is a very unequivocal command, if and only if, there is only one book on the table. But, what if there is more than one book on the table or maybe there is more than one table with books. It would be vague if any of these situations mentioned earlier occurs, as illustrated in Figure 1. However, it would be different if a hand gesture is involved in this scenario, as illustrated in Figure 2, whereby two modalities conveyed a clearer message to the audience.

Figure 1. Books on a table.

Figure 2. Involvement of a hand gesture

According to the Schlaisich and Egenhofer (2001), interactions between humans on spatial information often occur with talking and drawing at the same time [4]. Therefore, we believe that the utilisation of spatial information in multimodal integration would increase the accuracy of aligning the correct pairs of information from the different modalities. However, current multimodal integration techniques are designed to accommodate only interactions in English. Multimodal integration techniques in other languages are theoretically distinct from each other. To make multimodal applications easier for people to use, it is important to adapt it to a language that they are comfortable with. As each language is unique, a

978-0-7695-3904-1/09 $26.00 © 2009 IEEE
DOI 10.1109/IALP.2009.13