Multiple-Level Models for Multi-Modal Interaction

Martin Russell[1], Stephen Cox[2], Antje Meyer[3], Alan Wing[3]

[1]School of Engineering, The University of Birmingham
[2]School of Computing Sciences, The University of East Anglia
[3]School of Psychology, The University of Birmingham

m.j.russell@bham.ac.uk

Abstract

Though people usually speak in order to communicate, linguistic utterances, taken in isolation, are rarely unambiguous. Nevertheless interlocutors usually feel that they achieve a reasonable degree of mutual understanding. A number of factors, such as shared world knowledge, a common discourse model and awareness of the same visual world, contribute to communicative success. In addition, speakers can augment their utterances by non-verbal means, most notably by looking at the intended discourse referents and by pointing at them. Both eye gaze and pointing have been shown to have important functions for listeners as well as speakers. Thus, communicative success arises because interlocutors can use other types of information than just the speech signal. In theories of speech production and comprehension the core processes of retrieving and combing words are much better described than the processes of using world and discourse knowledge, eye gaze or gestures. There are, to our knowledge, no computationally implemented models capturing these processes.

Modern approaches to automatic interpretation of spoken language lack the knowledge and the mathematical formalisms to accommodate the processes which give rise to ambiguity, and instead focus on the relationship between word sequences and the acoustic signals to which they correspond. Even then, variations in the acoustic realisation of a sequence of words are treated as noise, and little or no attempt is made to model the underlying factors which might enable us to understand this variation. Arguments for the incorporation of additional levels of description in these models are compelling. For example, they have the potential to characterise the processes which give rise to the differences between read and conversational speech, or between the speech of an adult and that of a child, or to constrain a recogniser to give better performance in noise. These arguments also apply to modalities other than speech. Moreover, in addition to being critical to the understanding of variations in the realisation of a particular communicative goal within each modality, multiple-level representations may be needed for a proper understanding of the relationships between modalities

We propose a project investigating experimentally how speakers and listeners use eye gaze and pointing along with the speech signal. In parallel, we will develop computationally useful models that characterise the variations in the 'surface' realisations of these modalities by incorporating representations of the causes of variability.

In the experimental research we would use various referential communication and matching tasks and would record the participants' speech,. eye movements, and gestures. Of particular interest are the temporal relationships between eye gaze, gesture and speech. We know, for instance, that a speaker's eye gaze is typically ahead of their speech. By contrast, the listener's eye gaze usually follows the speech signal. Thus, though speakers and listeners look at the same objects, but not in perfect synchrony. The consequences of this asynchrony for the listener's speech processing are unknown. In addition, both interlocutors spend some time looking at each other, in order to obtain some nonverbal assurance of understanding as well as to determine where the other person is fixating, and, finally, speakers and listeners quite often look at their own and the interlocutor's gestures. Whether this actually facilitates comprehension does not seem to be known. Important goals of the empirical research would be to determine, first, how the speakers' and listeners' eye gaze are co-ordinated in discourse tasks when, for instance, the difficulty of the tasks or the time pressure to fulfil the task are varied, (b) to determine the functional significance of eye gaze and gesture information by

either allowing or preventing mutual eye contact between the interlocutors, (c) to determine the importance of precise temporal co-ordination of the speaker's and listener's gaze. (This could, for instance be investigated by asking listener to follow video-typed instruction in which the audio- and video-channel are, in some places, desynchronised.

The objective of the modeling research is to improve automatic identification of human communicative goals through the development of multiple-level models of speech, eye-movement and gesture which capture the levels of representation which are key to understanding variations in the surface representations of these modalities. This will be informed by, and use data from, the experimental work describe above. We aim to understand the statics and dynamics of these representations and the relationships between them *within* a particular modality, and to understand and model the relationships at different levels of representation *between* different modalities. To achieve this we will need to develop new, formal, trainable models, which incorporate several levels of parameterization, characterize the relationships between these layers, and incorporate continuous-state models of dynamics within the layers. Some progress has already been made.

The project will combine expertise in psycholinguistics, speech and language processing, and mathematical modeling.