Technology for Processing Non-verbal Information in Speech

Current speech technology is founded upon text. People don't speak text, so there is often a mismatch between the expectations of the system and the performance of its users. Talk in social interaction *of course* involves the exchange of propositional content (which *can* be expressed through text) but it also involves social networking and the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. A computer-based system that processes human speech, whether an information-providing service, a translation device, part of a robot, or entertainment system, must not only be able to process the text of that speech, but must also be able to interpret the underlying intentions, or *acts*, of the speaker who produced it. It is not enough for a machine just to know *what* a person is saying; it must also know *what that person is doing* with each utterance as part of an interactive discourse.

Tone of voice

Previous work carried out in Japan has shown that more than half of interactive speech in everyday conversations takes the form of nonverbal utterances which cannot adequately be transcribed into text. These stylised utterances as well as non-lexical *affective* speech sounds, such as laughs, feedback noises, and grunts, also carry important interpersonal information related to the states, intentions, and beliefs of the discourse participants, and to the progress of the social interaction as a whole. They constitute a small finite set of highly variable sounds in which most of the information is carried by prosody and tone-of-voice. It is this component of speech especially that makes it such a rich and expressive medium for human interaction, but this is an element of the signal that is not yet well modelled, if at all, by machine processing.

A human interlocutor intuitively interprets the nonverbal information in speech and tone-of-voice to aid in the interpretation of each utterance in context. It has been shown, for Japanese, that a machine can be programmed to perform similar interpretation of speech utterances, and currently research is being carried out to generalise and further develop these findings using speech data from other languages. While the academic goal of such research is to show that the use of nonverbal utterances in conversation is a characteristic of human speech *in general* and not limited to only one particular culture or language, the technical goal of the work is to produce devices that are *specifically adapted to interactive or conversational speech* that will enable a friendlier and more efficient speech interface for public services and entertainment.

Recognising that *social actions* are the essential component of intercourse, and that actions, rather than words are the prime units to be processed in a discourse, future speech research must specifically address the question of how new technologies can be produced which are capable of processing not only the lexical content of an utterance, but also its underlying intentions. This might be done by processing prosody & tone-of-voice.

To further the development of such speech technology, it is therefore essential to collect a representative corpus of spoken interactions wherein participants display the *full range of their daily speech strategies* and to use that material to train new modules for interactive speech processing (whether for synthesis or recognition) that can make use of such higher-level information. However, such a corpus requires the prior development of recording techniques that are unobtrusive, and environments which are felicitous.

Discourse dynamics

There is growing international interest in multimodal interaction processing (see e.g., UC *(Universal Communication)* in Japan, AMI *(Augmented Multimodal Interaction)* in Europe, and CHIL *(Computers in the Human Interaction Loop)* in the US) and in the collection of multimodal conversational speech data, which was identified as a principal future task at the LREC (Language Resources and Evaluation Conference) last year.

Whereas traditional approaches to spoken interaction and dialogue systems have tended to assume a “ping-pong” or “push-to-talk” model, wherein either the system or the interlocuting human is active at any given time, it is becoming increasingly apparent that the dynamics of spoken interaction is an important element in itself for speech information processing, and that the typical flow of speech is fragmented and multi-faceted, rather than forming a single uninterrupted stream. This is supported by many recent findings in conversation and discourse analysis, where the definition of a “speech-turn”, or even an “utterance” is proving to be very complex.

People apparently don’t “take turns” to talk in a typical conversational interaction; rather they each contribute actively *and interactively* to the joint emergence of a “common understanding”. The apparent "no gap no overlap" alternation of spoken utterances is actually emergent from a background of continuous behavioural coordination at different levels of behavioural organization. This *interaction synchrony* is a feature yet to be incorporated in modular speech processing technology and might prove to be an important element for dialogue interface design. It should therefore be taken into consideration as a key component of corpus design.

Corpus control

Speech data will continue to be collected from a variety of sources using a variety of capture devices. Techniques will be developed to deal robustly with impoverished or "less-than-perfect" materials, and a corresponding robustness will be reflected in the technology produced as a result. Conversely, in order to derive useful and reliable components for speech information processing, we should ensure that the corpora we collect are representative of the styles and mannerisms of interactive conversational speech, so that future users of this technology will be presented with interface designs that match their (unconscious) expectations and that are able to process the full range of information that is carried by inflections of the voice and from the characteristics of timing and turn-taking.

Conclusion

As we envisage the incorporation of speech processing modules in more and more sophisticated commercial applications, including machine interpretation, robotics, games, and customer-services, a key element of the research will be to develop methods that enable the efficient collection of conversational and interactive speech data without the need for extensive or invasive recordings. Privacy considerations may prevent the use of naturally-occuring samples, so this work may require the development of both capture devices (cameras and recorders) and capture environments (equivalent to a recording studio) that encourage participants ro relax informally and maximise their range of speaking styles and formats.

Nick Campbell Trinity College Dublin, February 2009