Foresight Speech, Language and HCI Abstract

Leslie Smith, Department of Computing Science, University of Stirling.

Language (and to a lesser extent speech) is a quintessentially human activity. Much research has been carried out on language acquisition, and on the use of language in babies and children. From the viewpoint of computational utility, understanding language and using it effectively in HCI could allow computers to take the step from being useful machines for the trained individual to being ubiquitous assistants usable by anyone. But this project has many levels: these range from detecting the basic elements of language from their carrier (sound for speech, moving images for sign), to piecing together words and phrases (and importantly too, capturing mood and intonation), to detecting the semantics of the assembled annotated phrases. Language generation is sometimes seen as the simpler task. However this is not simply producing a spoken (or signed) stream from a text: any children's teacher will tell you that spoken language requires the speaker to have an understanding of the meaning for effective communication.

One view of the task is that it can be described in terms of non-intersecting layers: sound to phonemes (or syllables), then syllables to words, then words to phrases and sentences. There are many problems with this view: as is well known, continuous speech has few word boundaries, and the pronunciation of syllables is very dependent on the surrounding ones. The standard techniques of speech recognition (HMMs, etc) are designed to take account of this. In addition, however, factors such as mood and intonation are detectable in the original sound, and these need to be taken account of in interpretation. Thus, a simple layered model, while attractive from an engineering viewpoint, is not necessarily appropriate.

At a lower level still, we need to accept that the sound signal is rarely clean. If we are to use speech interactively with ubiquitous machines, we need their capabilities to approach human capabilities. Normal hearing individuals can easily understand speech in poor signal-noise environments, even when the speech is further degraded by reverberation. Speech is produced within two constraints: what is biologically possible to produce, and what can be understood under realistic conditions. The articulatory side of speech production has been thoroughly researched: there is more to be gained now by understanding how sound (and speech) is processed in the auditory brainstem and cortex. Neurophysiological studies show the importance of short-term co-variations in different parts of the spectrum in the fusing of energy allowing foreground sounds to be segregated from the background. Biological techniques are highly parallel: many concurrent transformations are carried out on the sound, though only a selected set of results are attentively processed. Recent work at Stirling, Sheffield, and Ohio (amongst other places) suggests that these biological techniques can be engineered, providing a radically different "front end" for speech interpretation. Further, these types of technique can simultaneously provide information on intonation.

We are proposing (as part of a larger scale project)

- A neurophysiologically inspired front end. This is highly parallel, and continuously computes and groups features. To be useful, this will need to run in real-time, implying hardware implementation.

- A feature based recogniser. This recognises temporal sequences of grouped features. This is likely to produce a probabilistic output, and to be trained on the vocabulary to be recognised.
- An intonation and mood recogniser, based primarily on the outputs feature based front end.

For speech recognition, the larger scale project would include top-down prior setting. The overall project should include semantics: this is particularly critical for speech (as opposed to text read aloud), as the semantics of speech is often tightly coupled to the situation in which the speech occurs.