[*Translating and the computer 11*. Proceedings of a conference … 16-17 November 1989, ed. Catriona Picken (London: Aslib, 1990)]

Speech recognition, artificial intelligence and translation: how rosy a future?

Henry Thompson

Department of Artificial Intelligence, Centre for Cognitive Science, Human Communication Research Centre, University of Edinburgh

Note to the reader: This is a lightly edited version of the slides which accompanied the talk given at *Translating and the Computer 11.* Those interested in a more extended discussion of some of the points mentioned here are referred to (Thompson, 1984), (Thompson, 1986) and (Thompson, 1988).

SPEECH RECOGNITION – WHAT MAKES IT HARD?

Real problems

— The signal is lousy – it is noisy, and talkers are sloppy and lazy (Figure 1). The schwa (annotated as @) in the first syllable *of potato* is barely there at all – only about 20 milliseconds long, with much reduced excursion of the diagnostic properties. In general, talkers put no more information into the signals they produce than absolutely necessary. Hearers must use all sorts of 'filters' to reduce the resulting uncertainty of analysis.

— There are no reliable cues to the division of the signal into words - silence is if anything negatively correlated with word boundaries.

There are three words and a bit in the display in Figure 2 – can you tell where they are divided?*

There is good evidence that people use at least three different sorts of 'filters' – lexical, syntactic and semantic/pragmatic – to cope with these sources of uncertainty. The lexical filter effectively says 'Not just any sequence of English phonemes (the sound alphabet) will do, they must make up English words.' The syntactic filter says 'Not just any sequence of English words will do, they must make up (more or less) grammatical English utterances.' And finally the semantic/pragmatic filter says 'Not just any English utterance will do, it must make sense given what I know about the talker, the conversation so far and the world'.

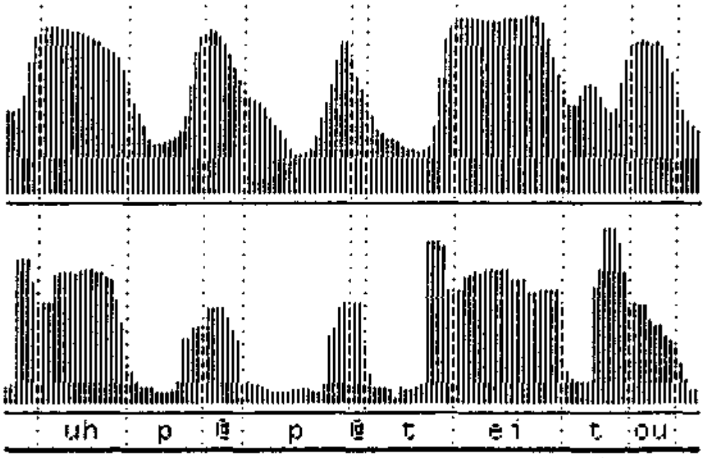

Figure 1. Some of the signal properties for an utterance of 'up a potato'

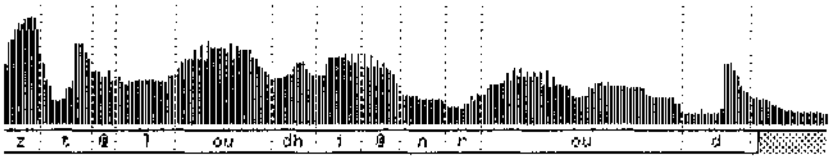

Figure 2. Continuous speech

Technology problems

— The sound-to-symbol mapping is imperfectly realised – context and individual differences have profound and complex effects, which we cannot yet capture in our computer systems.

We don't have complete theories about what aspects of context and what properties of the speaker determine the acoustic properties of (the sound which corresponds to) an individual phoneme. Our systems only imperfectly embody such incomplete theories as we do have.

— 'Does this make sense' filter cannot be done yet, and available approximations all are flawed.

Consider the following, which shows a small subset of the words which might be found to account for a part of a simple utterance, given a quite good effort from the first stages of processing:

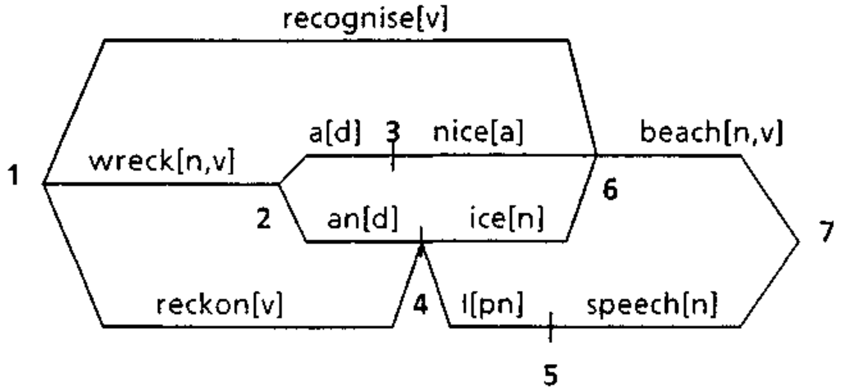

Figure 3. Alternative analyses of /r e k @ n ai s b ii ch/

Of the 15 paths through this lattice, two are syntactically consistent with the prologue ‘People can easily . . .’ But only one makes sense in context. Can artificial intelligence implement the ‘Does this make sense’ filter we need?

ARTIFICIAL INTELLIGENCE DEFINED

Artificial intelligence is a methodology, not a discipline. The methodology is the effective computational deployment of knowledge to solve problems.
Two related tasks arise in any application:

1. *Represent* the knowledge in a computationally tractable form;
2. *Design and implement algorithms* which effectively employ that knowledge so represented to achieve the desired processing.

But performing those two tasks in support of providing a ‘Does this make sense?’ filter cannot yet be accomplished in the general case. Our ability to represent and employ general knowledge of the required sort is just not up to it. In other words

> Fully Automatic High Quality Unrestricted Continuous Speech Recognition/ Machine Translation/Scene Analysis/. . . crucially involves *understanding* and is therefore out of reach for the time being.

There are two possible responses: 1) give up *unrestricted;* 2) give up *fully automatic.*

THE STATE OF THE ART

All existing commercial systems, and most research prototypes today, have taken the first of two routes, or a combination of both:

— large vocabulary isolated word systems, speaker trained, post-editing required

— small vocabulary, mandated grammar systems, speaker independent, 98 per cent, words correct.

NEAR TO MEDIUM TERM PROSPECTS

More of the same. Breakthroughs in the knowledge representation problem are not obviously imminent. There will be quantitative improvements (what counts as small, etc.) but no qualitative leaps.

WHAT ABOUT TRANSLATION?

Live within the constraints described above? Niche markets at best, hard to imagine but might be possible. For instance, one might dignify with the descriptor *translation* the invocation of spoken phrase book entries via single word speech recognition, in which case the technology exists for spoken machine translation.

Wait for the revolution, and work like hell while you wait? A sensible course of action.

Full speed ahead and damn the torpedoes? The Japanese ATR Interpreting Telephony Project has full end-to-end spoken language interpretation as its goal. The Japanese approach to technological progress is often based on setting unachievably high goals, with the confidence that much beneficial spin-off will be generated in pursuing them. The current state of that effort can be gleaned from a number of papers by, *inter alia,* Kurematsu, Iida, Komori, Inoue and Takeda in (Tubach & Mariani, 1989).

*Figure 2 displays the end of the utterance 'Which bus goes to Lothian Road?'

REFERENCES

Thompson, H.S. 1984. ‘Artificial Intelligence and Speech Processing: The Good News and the Bad News’. In: J.N. Holmes, ed., *Proceedings of the 1st International Conference on Speech Technology.* Institute of Acoustics (Speech Group), UK.

Thompson, H.S. 1986. ‘Knowledge Based Systems for Speech Recognition: What Kind of Knowledge?’. In: *Proceedings of the First International Congress on Knowledge and its Engineering,* Polytechnic of Madrid.

Thompson, H.S. 1988. 'Speech Technology and the Language Industry: Opportunities and Pitfalls'. In: *Proceedings of Online Information 88,* Learned Information Ltd., Oxford.

Tubach, J.P. and J.J. Mariani, eds. 1989. *Proceedings of the European Conference on Speech Communication and Technology,* European Speech Communications Association, published by CEP Consultants, Edinburgh.

AUTHOR

Henry Thompson, Department of Artificial Intelligence, Centre for Cognitive Science, Human Communication Research Centre, University of Edinburgh, 80 South Bridge Street, Edinburgh EH1 1HN, UK.