Speech, Hearing and Language: work in progress

Volume 11

Opportunities for re-convergence of engineering and cognitive science accounts of spoken word recognition.

Mark HUCKVALE

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

Opportunities for re-convergence of engineering and cognitive science accounts of spoken word recognition[1]

Mark HUCKVALE

Abstract

This article traces the roots of the divergence between the engineering community and the cognitive science community accounts of word recognition. It argues that although there are cultural differences, when looked at objectively, there is considerable overlap in the desires and motivations of the two communities. It suggests that the criticisms of engineering systems that caused the original divergence in the late 1970s are much less valid today, that re-convergence is timely and will help create a theory of speech processing which will explain both primary and emergent phenomena. It proposes that the study of LVCSR systems as if they were human, and the study of humans as if they were LVCSR systems, could lead to a research agenda which would benefit both communities. It introduces elements of a programme to encourage joint research and co-operation.

1. Introduction

As users are only too aware, contemporary large vocabulary speech recognition systems do not respond to speech in the same way as humans. The dictation systems that are in use today are very sensitive to dysfluencies, restarts, background noise and change of speaker or voice quality. Furthermore the recognition errors they make seem to be very different to the ones that humans make even when listening in poor environments. There is no doubt that recognition systems will only become more comfortable to use when they act more like a human listener. This should mean that scientific knowledge about how humans process speech is relevant and important in the design of these systems. However it is now the case that scientific research into the human processing of language has diverged from research into artificial language processing systems. We have separate and independent fields of 'psycholinguistics' and 'spoken language engineering'.

This article explores the relationship between the engineering and cognitive science communities within the relatively well-defined sub-field of spoken word recognition. That is we shall be mainly concerned with the processes by which word sequences are recovered from acoustic input. The starting point for the article is that these two communities had similar aims and methods for word recognition in the early 1970s but that these had diverged markedly by the early 1980s. The main argument of the article is that opportunities now arise to build bridges between the communities: that engineered systems are sophisticated enough that parallels can be drawn with human recognition at a suitable level of abstraction. Such parallels allow the cross-fertilisation of ideas between cognitive science and systems engineering.

The article is in three parts: the roots of the divergence between engineering and cognitive science accounts of word recognition are explored in the first part. Differences in motivation, methodology and culture are all seen to play a part and are explored in a historical context. The second part of the article discusses the potential benefits of a re-convergence of the two scientific fields and argues that the time is ripe for progress now. Engineering systems are

[1] This is a revised version of an article published in the Proceedings of the Institute of Acoustics, Vol. 20 (1998) pp9-20.

stable and successful enough to be worth interpreting in cognitive terms, while they are sophisticated enough to allow useful comparisons with humans to be undertaken. The final part of the article proposes some elements of a joint research programme which could act as a stimulus for the two communities to work together. Highlighted are the cognitive accounts of priming phenomena which relate to recent engineering work in adaptation, and cognitive accounts of morphological processing which relate to engineering problems of vocabulary selection and use. Other possibilities relate to phonetic reduction phenomena at the low end, and semantic grouping or phrasing at the high end of both human and machine recognition.

2. Background

2.1 Historical Background

The systems at the peak of the artificial intelligence approach to speech understanding in the 1970s: Hearsay (Erman & Lesser, 1979) and HWIM (Wolf & Woods, 1979) operated using symbolic processing paradigms which remain familiar and comfortable within cognitive science today: independent knowledge sources containing production rules, distinctions between long-term and working memory, and management systems for setting rule-firing priority with little attempt at knowledge integration or optimisation. Thus Hearsay and HWIM, unlike later systems, could be both engineering implementations and acceptable cognitive accounts of human word recognition.

The development of the Harpy system (Lowerre & Reddy, 1979) is usually taken as a watershed in recognition systems development. Harpy was the first large vocabulary continuous speech recognition (LVCSR) system of reasonable performance, and it achieved its success by making a significant break with the architectures of Hearsay and HWIM. Harpy replaced multiple knowledge sources with an integrated network of spectral templates, and rule firing by graph search. Harpy was influential because it showed that good recognition performance could be achieved through good engineering without good quality linguistic knowledge. However the use of pattern recognition algorithms formally outside the domain of artificial intelligence had other consequences: it threatened to split the field into those that would accept any computational framework for recognition providing it would do the job from those that sought an explanation of human processing using familiar symbolic manipulation.

The potentially explosive consequences of such a split were not lost on two prominent scientists of that period. Both Alan Newell and Dennis Klatt studied Harpy to try to deduce lessons for a theory of human processing. Klatt's analysis lead to a cognitive model called LAFS (Lexical Access from Spectra; Klatt, 1980), while Newell's led to an attempt to link Harpy's integrated search into the AI paradigm of production systems, and thereby to absorb Harpy's success back into the conventional AI paradigm (Newell, 1980).

It is probably fair to say that neither of these attempts at bridge-building across the engineering/cognitive-psychology divide was acceptable to either side. LAFS was never a successful implementation, nor taken seriously as a cognitive model. Newell's attempt to re-establish the dominance of production systems did not lead to an AI implementation of Harpy, nor to cognitive studies based on production systems. Donald Norman (1980) is particularly scathing about both attempts. His main criticisms are important because we will claim later in this article that they have been largely addressed in the intervening period.

Norman's criticisms of Harpy as a cognitive model can be summarised as:

1. Harpy's performance, though better than knowledge-based systems, was still considerably worse than a human.

2. Harpy's architecture was only one of many potential architectures for speech recognition. Reddy (1980) had estimated that there were over 1,000,000 possible architectures, so that the importance of its specific structure could not be stated.

3. Harpy did not show how higher level linguistic constraints relating to syntax, meaning or discourse could be incorporated in the search.

Norman could not see the value in studying humans as if they were implementations of such an arbitrary system design.

In the late 1970s, just as engineers were being given new direction by Harpy's success, cognitive psychology benefited from a new type of theoretical model of word recognition. Marslen-Wilson's Cohort theory (1978), spurred an explosion of interest in the time course of human lexical access. Cohort theory did not pretend to be a recognition architecture that could be implemented to take in signals and recover word identities. Rather it aimed to account for the phenomenological properties of human word recognition: that listeners were able to identify a word as soon as sufficient of it had been heard to reduce the number of lexical candidates to one. It made testable predictions about the results of experiments that could be undertaken in any psychology laboratory, and it did so using a symbolic processing paradigm involving phonetic segments. This reliance on accurate bottom-up segmentation and labelling was another factor in this story. The one lesson that all engineers took from Harpy was that high performance came from postponing decisions: not to identify segments until such time as top-down information about potential word sequences were available. By integrating word sequence and pronunciation constraints, Harpy was able to perform the trick of recognising words without making decisions about the identity or location of segments in the signal.

The basis of modern engineered speech recognition systems arose in the work of the IBM research team in the early 1980s (Bahl, Jelinek & Mercer, 1983). Here the emphasis on good engineering over good linguistic knowledge rose to a peak. Current LVCSR systems continue to exploit the architecture pioneered by the IBM team: the use of a separate acoustic model and language model, the use of Bayes' theorem to underlie search, time-synchronous decoding with beam pruning, and partial traceback to generate output during search (Young, 1996). Here too system word accuracy was placed above all other criteria for success - regardless of contemporary linguistic or psycho-linguistic wisdom - confirming the divergence started by Harpy.

But while the engineers sought better word accuracy, the psychologists explored other phenomena related to human recognition: a preference for real words over nonsense words, or a preference for high-frequency words over low-frequency words. A major competitor to the Cohort theory came along with the connectionist revolution in cognitive science. The TRACE model (McClelland & Elman, 1986) was a connectionist model of human word recognition based on an interactive activation architecture. TRACE went beyond the predictions of the Cohort model to these other effects. Although implemented with both a speech signal input and a phonetic feature input, it was very limited in the vocabulary size it could deal with. TRACE suffered from the same unrealistic assumptions about bottom-up phonetic

transcription as did the Cohort theory: its ‘explanations’ of human processing relied on simulated input. Nevertheless TRACE is important to our argument because as a computer program it showed that cognitive theories of speech recognition could be implementable (and conversely, that an implementation could be a cognitive theory).

The most recent cognitive model we shall introduce is the Shortlist model (Norris, 1994). Here, perhaps in a small acceptance of the emerging LVCSR systems, problems of dealing with large vocabularies become relevant. In Shortlist, Norris sees a primarily bottom-up word hypothesis component, fed by symbolic phonetic input, which feeds an interactive activation architecture of word competition. Prior to word competition, the input stage generates a short list of candidates on phonetic grounds. Although this process has some similarity with the ‘Fast Match’ procedures used in LVCSR to subset vocabulary prior to search (Gopalakrishnan & Bahl, 1996), the motivations are quite different. In Shortlist a candidate list is generated to get round the need for top-down feedback on the phonetic analysis, while in LVCSR a reduction in word candidates is only needed to reduce the amount of processing and memory required. In the latter case, the quality of the match between the input and the hypothesised words is still expressed in terms of phone probabilities.

In summary, the cognitive models have been created to account for the results of experiments in the time course of human recognition, or the human reaction to ambiguity, but not to explain human word accuracy. Conversely, the engineering models have only been created to approach human performance in word accuracy, and on the whole have not been used to explain other aspects of human word recognition. Cognitive models have not been designed as working recognition systems, and working recognition systems have not been designed as cognitive models.

2.2 Motivations

We return to the issue of the motivation behind the scientific research undertaken in the two fields. We need to find an expression of the key issues that separate the two communities. Underlying the slogans that cognitive scientists want to "explain human behaviour", while spoken language engineers only want to "build a working system", are two different issues.

We can use quotes by Alan Garnham to introduce these:

> "A working program is of psychological interest only if it is based on general explanatory principles about the way the mind works." (Garnham, 1989)

In other words, not all computational architectures are acceptable as models of human cognition. The general defence of this position is usually presented by cognitive psychologists using the analogy of chess-playing. The search strategy used by machine chess programs is generally accepted to be different to how humans play. From this example of how machines play a mathematical game, we are meant to infer that how machines process human language is equally invalid. As to which aspects of LVCSR architectures are most objectionable, the only commentators I have found refer to the use of Hidden Markov Models (HMMs) (Massaro, 1996). HMMs are seen as too general a mechanism, with too many free parameters to be the basis for a parsimonious account of cognition. We shall simply note for now that HMMs are just a mechanism to deliver a table of phone probabilities against time: a mechanism that can be replaced quite satisfactorily with a connectionist model (Hochberg *et al*, 1995) that is fundamentally no different from TRACE.

The second issue:

> "Realistic outputs do not indicate that any theoretically useful analysis of language understanding has been made." (Garnham, 1989)

In other words, even if our engineering model had the same behaviour as a human, it might not be working in the same way as a human. This statement, while true, does however deny the possibility of scientific research in the field. Any scientific theory may be false, and that is why we do experiments to try to differentiate alternative hypotheses. On the other hand, if we compare an engineering system that recognises speech with an accuracy that makes it a viable commercial product with a cognitive system that gives essentially random behaviour for the same task, there is no question which is the better theory. As far as modelling human word recognition accuracy is concerned, the problem is not that we have competing theories, but that we only have one working theory with which we can do experiments.

To ever consider the engineering community and the cognitive science community working together in word recognition, we must address these two issues, which we shall call (i) the cognitive architecture issue, and (ii) the multiple methods issue.

2.3 Methodological Divergence

Another aspect to the divergence emerged in our historical account and is worth investigating further. The engineers, quite openly, chose to pursue human word accuracy as the sole goal in their research, while the cognitive scientists pursued fidelity to human behaviour apart from accuracy. We should emphasise the difference here. The engineers were interested in *primary* behaviour: the ability to actually recognise the identity of a word accurately from the sound stream. On the other hand the cognitive scientists were interested in *emergent* behaviour: the side-effects of recognition. Thus the measure of the engineer is percent correct, while the measures of the psychologists are typically response times as a function of word frequency, ambiguity or linguistic context.

This divergence in methodology does not mean that engineering theories cannot be used to predict emergent behaviour nor that cognitive models cannot be used to predict primary behaviour. Cognitive models can be equipped with an acoustic-phonetic front-end and used to explain how words can be recognised from signals (which they do rather poorly), and engineering models can be used to explain, say, why word frequency has an effect on the resolution of ambiguity (which in fact they do quite well).

There is no doubt that a future ‘theory of speech communication’ would have to explain both the primary and the emergent behaviour. This is another indication why it is necessary to bring the communities together.

2.4 Cultural Divergence

Briefly we can mention some other aspects of the divergence associated with the mathematical tools and scientific culture in the two communities.

Mathematically, the engineers use data modelling techniques applied to very large speech and text corpora, hoping for structure to be learned rather than specified. The results of this data modelling are statistical likelihoods rather than deterministic rules. The philosophical problem with this is that data-driven models could seem arbitrary rather than based on testable principles. On the other hand, a great deal of morphological and phonological coding

is arbitrary. There are also differences in decoding procedures which are sequential in LVCSR while they are parallel in TRACE and Shortlist. In this case it is easier to see compromises: it is likely that parallel processing equivalents of Viterbi search can be found.

Culturally, the two communities tend to inhabit different university departments, publish in different journals, and obtain research funds from different funding sources. We should not underestimate the power of the 'not invented here' syndrome which blinds workers to the achievements of others. If we are to establish a future joint research programme, it will not be easy for any one group to give up the determination of the research agenda.

2.5 Why Reconverge Now?

Before we discuss how a re-convergence might be obtained, it is worth discussing why the time is appropriate now rather than in the past or in the distant future.

Let us return to the criticisms made by Donald Norman (1980) of Harpy as a cognitive model:

(i) Harpy's performance was not very good compared to humans. The performance of modern LVCSR is radically better than Harpy, the best recent figures for research systems on read speech are around 95% word accuracy on vocabularies of 65,000 words (Young, 1996). Human performance, particularly on spontaneous speech and on speech in noise is still significantly better (Lippman, 1997). However not even the psycholinguists are suggesting that ideas from TRACE or Shortlist will make much of an impact on this discrepancy.

(ii) Harpy's architecture was only one of many. Curiously, the fact that LVCSR architectures have remained stable since the 1980s shows that they are not arbitrary or readily open to alternatives. While details of implementations change, such as the use of triphones or recursive neural networks for the acoustic model, the overall construction has stood the test of time. If it wasn't capturing some useful properties it would have been replaced completely in the past 15 years.

(iii) Harpy's architecture did not allow for the incorporation of higher-level linguistic knowledge. Modern LVCSR systems certainly have more sophisticated language models than Harpy, and work continues to incorporate prosody, syntax and task constraints. Modern systems still make a separation between the word recognition component and an interpretive component which decodes shallow word-lattices with respect to the task. However the cognitive evidence for more top-down influence than this *in the recognition of the words themselves* is rather weak (Tanenhaus *et al*, 1979). Work in topic adaptation and trigger pairs (e.g. Iyer & Ostendorf, 1996; Lau *et al*, 1993) can be seen to overlap considerably with cognitive accounts of semantic priming (e.g. Swinney, 1979; Zwitserlood, 1989).

Another significant factor which makes re-convergence timely, is the ready availability of LVCSR systems for experiment. There are a number of toolkits available for researchers to build their own systems, and some complete systems that can be downloaded over the Internet. (for example the Abbot system of Tony Robinson *et al*, `ftp://ftp-svr.eng.cam.ac.uk/comp.speech/recognition/AbbotDemo`).

Reconvergence is also timely for cognitive science accounts of word recognition. We have seen how the design of Shortlist has been influenced by a need to demonstrate how a cognitive model could function with an everyday sized vocabulary. Revisions to the Cohort

model (Marslen-Wilson, 1987) have been necessary to accommodate less than perfect phonetic analysis. TRACE itself has been criticised for being a less than realistic computational architecture (Norris, 1994).

2.6 Can LVCSR address Cognitive Science issues?

To make headway with re-convergence proper, we return to the two significant motivational issues. Firstly the cognitive architecture problem: can LVCSR architectures be considered analogous to cognitive architectures? Well clearly not at the level of the hardware or the lowest level of the software. No one could claim that double-precision floating point numbers or multiple-mixture gaussian distributions are used in the brain. This is however, to miss the point. We can establish a level of abstraction of an LVCSR system where analogies can be made. These could include:

a. the use of continuous values to represent phone likelihoods,

b. the time synchronous construction of a word lattice,

c. competition between sentence-fragment hypotheses,

d. the integration of concordance derived likelihoods into sentence fragment scores,

e. the lack of on-line phonological processing,

f. the lack of on-line morphological processing,

g. the lack of influence of interpretation on sentence fragment scores in the current sentence.

Possibly items a-c. are uncontroversial, d. needs to be argued, while e-g. seem to be clearly mistaken as far as cognition is concerned. However the experiments to determine the relative importance of e-g. to humans have yet to be done. When and if it can be shown that human primary recognition behaviour benefits from using, say, on-line phonological recoding of the lexicon in context, then such benefits might also accrue to the engineering system.

Looking at the issue from the other direction, Altmann (1990) has provided a list of human word recognition issues that cognitive models need to address. We should see what an abstracted LVCSR system has to say about these:

- "How does acoustic input contact the lexicon?" Input is explained using knowledge of the likelihoods of the acoustic realisations of segments constrained by the lexicon and language model.
- "What is the nature of the intermediate representations?" Sentence-fragment hypotheses are built in a time-synchronous manner.
- "What strategies are used to facilitate recognition?" Continuously scored hypotheses are influenced by higher level sequential likelihoods.
- "What is the locus of word frequency effects?" Bayes' theorem is used to combine conditional with unconditional word probabilities.
- "What factors influence word competition?" Acoustic similarity, prior frequency, and sequential probability interact through likelihoods.
- "How do contextual effects arise?" Likelihoods are affected by the context through the use of the language model and though adaptation.

In summary, it is possible to finesse the cognitive architecture issue by choosing a level of abstraction of LVCSR systems which is detailed enough to make testable predictions, but fuzzy enough to hide their current implementation on sequential digital computers.

The second significant issue was "multiple methods" - even if an LVCSR system does the job, it might not do it in a way that parallels how humans do it. We have already attacked this position as anti-science - it would deny progress to any theory simply on the basis that it might not be ‘true’. Another argument against this was indicated in the first paragraph of this article. Engineers are not trying to solve an artificial game but to process human speech with the flexibility with which humans process speech. The signals are not artificial but generated by humans; on the whole they are not strongly adapted to a human’s perception of machine abilities, but remain similar to the signals that humans generate when talking to each other. Thus engineering systems are processing human specified and human produced language designed for human processing and human interpretation. They might do it in a radically different way to humans, but only by comparing both their primary and emergent behaviours can we identify where there are significant discrepancies.

3. Benefits

Why should a cognitive scientist want to investigate LVCSR systems? What benefits are there to an engineer to make her system more human-like? We now turn to the benefits to both communities that could be obtained by reconvergence.

I would like to argue that existing computer implementations of cognitive models (such as TRACE) are less than convincing explanations of human recognition in terms of either their primary or their emergent behaviour. Their explanations of how a cognitive model determines a ‘cohort’ of word candidates, or how a word sequence is extracted from continuous input are seriously weakened by the use of predigested phonetic units rather than signals as input. This assumption sweeps under the carpet the enormous problems of noise, speaker variability and coarticulation on transcription. It also denies any research that relates these issues to lexical access. Contemporary issues in psycholinguistics relating to the processing of errorful input (e.g. Marslen-Wilson, 1987) are also hindered by such early decision making.

Worse still are the supposed ‘explanations’ for emergent effects. Thus prior word activations in TRACE (and other cognitive models) are meant to account for human preference to resolve ambiguity in the direction of frequent words over infrequent words. But to argue that frequent words are more readily recognised because they have greater prior activation is simply to build the phenomenon into the model. Presumably if it had been found that words beginning with /b/ were more readily recognised, then these would have been given extra activation. A true explanation of an emergent effect is one that is an inevitable consequence of the primary processing. In machine recognition, more frequent words are chosen because, on average, that maximises the likelihood of correct identity. In contrast, this is a simple, direct and falsifiable claim.

The most significant benefit to be derived from the substitution of phenomenological models of word recognition with an LVCSR system would be that alternative hypotheses about cognitive processing could be tested against one another using real speech data. LVCSR systems can be ‘opened up’ to give access to the table of phone probabilities, or to the lattice of word hypotheses, see Figure 1. Data from the beam search can be extracted on a frame-by-frame basis. Relative activations (probabilities) can be measured and manipulated, prosodic cues incorporated, effects of semantic priming modelled. All this can be done within a

computational framework directed towards the primary goal of maximising recognition performance. Such a goal is also reasonable to assume for the human listener.

The benefits of reconvergence to the engineering of LVCSR systems is mainly to provide new foci and new directions of research. A criticism that has been directed at engineers since Pierce (1969) has been that experiments have been conducted for no theoretical reasons, but merely because they were possible. A critical view would maintain that progress in LVCSR has been due to a 'ratchet' effect - keeping the most productive of thousands of random changes to existing systems - rather than because of well-motivated research and development.

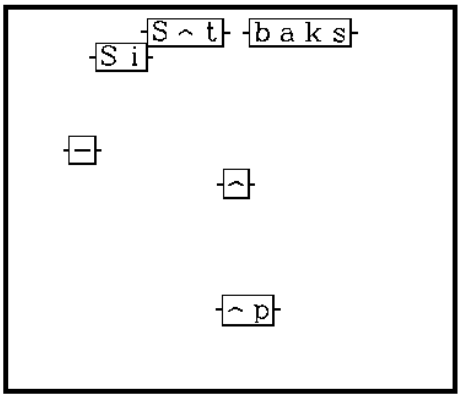
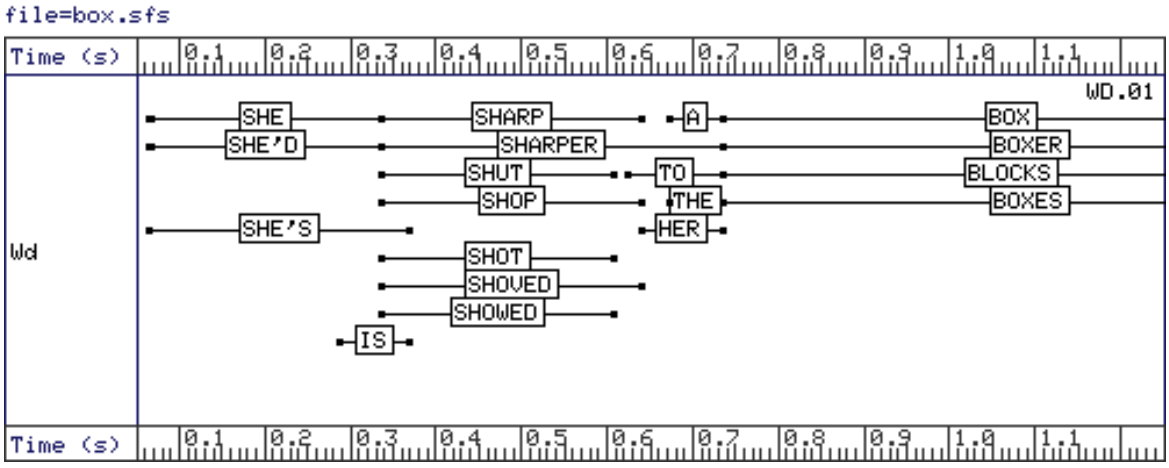


Figure 1. *Left: output from TRACE (McClelland & Elman, 1986) for the simulated sentence "She shut a box". Below: output from the Abbot LVCSR system (Hochberg et al, 1995) for a real version of the same sentence. The vertical dimension in both graphs is activation.*



The problems facing LVCSR system designers today: noise, speaker variability, contextual variability, dysfluency, or the differences between read and spontaneous speech, are only considered 'problems' because human listeners are only weakly affected by such things. There may be engineers who want to create recognisers better than humans, but I suspect most would be happy with a performance that equals human. Most would agree that ultimate performance would mean a much deeper 'understanding' of the communication than we could expect of a machine. Thus understanding how humans have been able to 'solve' these problems is directly relevant to how machines could be improved.

Reconvergence would thus provide a clear research agenda for the developers of LVCSR systems - to investigate the discrepancies between engineered and human systems. By this I don't mean merely to discover that humans are better (as Lippman (1997) has done), nor to

borrow psycho-acoustic results blindly (Bourlard *et al*, 1996), but to perform experiments to expose differences in processing and representation. In this way, a direction can be given to engineering efforts and the convergence of the communities cemented.

4. Re-convergence

4.1 Getting Communities together

To make a start on reconvergence, we could consider a small number of initiating activities:

- **Joint research programme**. A small number of topics could be chosen where there is clear interest and overlap of expertise across communities. Some first suggestions are given below.
- **Shared tools and data**. The communities could open up access to the programs and signals each use for experimentation.
- **Exchange of research workers**. The communities could build channels of communication and trust through visits, exchanges and joint meetings.

Sponsorship through research funding councils would help, as would support from journal publishers.

4.2 Outline programme

The first part of a joint programme should be to explore the differences between LVCSR and human recognition, not in terms of absolute performance, but in terms of emergent behaviour. The utility of this has been denied by the cognitive scientists using the same arguments we reviewed in part 1, but what experiments have been done have lead to interesting results (Cutler & Robinson, 1992). It has been easy to suggest that because LVCSR systems don't have some feature of cognitive model X then one can not make any useful conclusions. An alternative view is that the very discrepancies would indicate exactly which phenomena are just side effects of recognition and which indicate linguistic specialisation. In the former I would put word frequency effects, and in the latter I would put prosodic phrasing.

In conjunction with this analysis of LVCSR systems as if they were human is the cross comparison of humans as if they were LVCSR systems. To what extent do humans actually use on-line syntactic constraints in word recognition as opposed to simple concordance likelihoods? Could the results of gating experiments be predicted by a simple template recogniser?

In the absence of these cross comparisons, we can only speculate about the most productive areas of a joint research programme. There are two relatively clear groups of relevant activities: the probabilistic modelling of human linguistic processing, and the study of adaptation at a number of levels.

It is easy to identify aspects of human linguistic processing not yet incorporated into LVCSR systems. This may be because recognition can be performed without them, or that they cannot be described well enough or in the right way. A first topic is the way human speakers use prosodic cues to group words into chunks. There seem to be complex links between grouping and meaning which could be modelled by statistical means and hence be amenable for combining with language model likelihoods. A second is the way humans process morphologically complex words in different ways. Inflexional morphology appears to be treated differently to synchronically productive morphology and differently also to historically fossilised morphology. Language and pronunciation modelling could be adapted to process

morphs or lemmas using the same statistical methods currently used for words. A third aspect of human linguistic processing is phonetic reduction during production: dropping/merging of syllables, smoothing of diphthongs, lenition of stops to fricatives, etc. The probabilistic distribution of such phenomena might not only improve phonetic recognition but might also serve as a testing ground for phonological theory.

A second separate part of a potential joint programme is to look at adaptation at various levels in the linguistic hierarchy and at various time scales. Considerable amounts of current LVCSR research is related to tuning systems for particular speakers, noise conditions or topics. The premise is a large system adapted on the basis of a small amount of known evidence. It is possible to forsee adaptation being accepted as a general mechanism by which recent experience is used to maximise recognition accuracy. This might then apply not only to the signal processing or the language model, but even within a sentence to explain the semantic coherence of sentence hypotheses. This last issue relates directly to much recent work in psycholinguistics on semantic priming (Swinney, 1979; Zwitserlood, 1989), for which there is as yet no computational implementation.

5. Summary

In this article I have traced the roots of the divergence between engineering and cognitive accounts of word recognition. I have tried to show that although there are cultural differences, there is considerable overlap in the desires and motivations of the two communities and opportunities for mutual support. I have suggested that the criticisms of engineering systems that caused the original divergence are much less valid today. I have suggested that convergence will help create a theory of speech processing which will explain both primary and emergent phenomena. I believe that the study of LVCSR systems as if they were human, and the study of humans as if they were LVCSR systems, could lead to a research agenda which would benefit both communities. I have proposed the beginnings of a programme to encourage joint research and co-operation and indicated areas which I predict will be of interest and utility.

Acknowledgements

Thanks to Bill Ainsworth, Anne Cutler, Georg Meyer, Roger Moore and Andy Faulkner. Responsibility for the views represented here remains mine.

References

Altmann, G., ‘Cognitive Models of Speech Processing: An Introduction’, in Cognitive Models of Speech Processing, ed G. Altmann, MIT Press, 1990.

Bahl, L., Jelinek, F., Mercer, R., ‘A Maximum Likelihood Approach to Continuous Speech Recognition’, IEEE Trans. Pattern Analysis and Machine Intelligence, 5 (1983) pp179-190.

Bourlard, H., Hermansky, H. & Morgan, N., ‘Towards Increasing Speech Recognition Error Rates’, Speech Communication 18 (1996) pp205-231.

Cutler, A., Robinson, A., ‘Response Time as a Metric for Comparison of Speech Recognition by Humans and Machines’, Proc. ICSLP, Banff Canada, (1992) pp189-192.

Erman, L. & Lesser, V., ‘The Hearsay-II Speech Understanding System’, in Trends in Speech Recognition, ed W. Lea, Prentice Hall, 1979.

Garnham, A., Psycholinguistics, Routledge (1989).

Gopalakrishnan, P. & Bahl, L., 'Fast Match Techniques', in Automatic Speech and Speaker Recognition, ed C. Lee, F. Soong, K. Paliwal, Kluwer (1996).

Hochberg, M., Renals, S., Robinson, A., 'ABBOT: The CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System', in Proc. Language Technology Workshop, Austin Texas, Jan 1995, Morgan Kaufmann.

Iyer, R., Ostendorf, M., 'Modelling Long Distance Dependence in Language: Topic Mixtures vs. Dynamic cache Models', Proc. ICSLP, Philadelphia, (1996) pp236-239.

Klatt, D., 'Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access', in Production and Perception of Fluent Speech, ed R. Cole, Lawrence Erlbaum, 1980.

Lau, R., Rosenfeld, R., Roukos, S., 'Trigger-based Language Models, a Maximum Entropy Approach', in Proc. ICASSP (1993), pp45-48.

Lippmann, R., 'Speech Recognition by Machines and Humans', Speech Communication, 22 (1997) pp1-15.

Lowerre, B. & Reddy, R., 'The Harpy Speech Understanding System', in Trends in Speech Recognition, ed W. Lea, Prentice Hall, 1979.

Marslen-Wilson, W. & Welsh, A., 'Processing interactions and lexical access during word-recognition in continuous speech', Cognitive Psychology 10 (1978) pp29-63.

Marslen-Wilson, W., 'Functional Parallelism in Spoken Word Recognition', Cognition 25 (1987), pp71-102.

Massaro, D., 'Modelling Multiple Influences in Speech Perception', in Computational Psycholinguistics ed. T Dijkstra & K de Smedt, Taylor & Francis, 1996.

McClelland, J. & Elman, J., 'The TRACE Model of Speech Perception', Cognitive Psychology 18 (1986) pp1-86.

Newell, A., 'Harpy, Production Systems, and Human Cognition', in Production and Perception of Fluent Speech, ed R. Cole, Lawrence Erlbaum, 1980.

Norman, D., 'Copycat Science or Does the Mind Really Work by Table Look-Up?', in Production and Perception of Fluent Speech, ed R. Cole, Lawrence Erlbaum, 1980.

Norris, D., 'Shortlist: A Connectionist Model of Continuous Speech Recognition', Cognition 52 (1994) pp189-234.

Pierce, J., 'Whither Speech Recognition?', J.Acoust.Soc.Amer. 46 (1969) pp1049-51.

Reddy, R., 'Machine Models of Speech Perception', in Production and Perception of Fluent Speech, ed R. Cole, Lawrence Erlbaum, 1980.

Swinney, D., 'Lexical Access During Sentence Comprehension: (Re)considerations of Context Effects', Journal of Verbal Learning and Verbal Behaviour, 18 (1979) pp645-659.

Tanenhaus, M., Leiman, J., Seidenberg, M., 'Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Context', Journal of Verbal Learning and Verbal Behaviour, 18 (1979) pp427-441.

Wolf, J. & Woods, W., 'The HWIM Speech Understanding System', in Trends in Speech Recognition, ed W. Lea, Prentice Hall, 1979.

Young, S., 'A Review of Large Vocabulary Continuous Speech Recognition', IEEE Signal Processing Magazine (September 1996) pp45-57

Zwitserlood, P., 'The Locus of the Effects of Sentential-Semantic Context in Spoken-Word Processing', Cognition 32 (1989) pp25-64.