Presented at NATO ASI Workshop Speech Pattern Processing, Jersey 1997

10 THINGS ENGINEERS HAVE DISCOVERED ABOUT SPEECH RECOGNITION

Mark Huckvale

Dept. Phonetics & Linguistics
University College London
Gower Street, London WC1E 6BT
M.Huckvale@ucl.ac.uk

ABSTRACT

While there is still much room for improvement, current speech recognition systems have remarkable performance. Rather than asking what is still deficient, we ask instead what engineers must be doing right. The answers conflict in many ways with the conventional paradigm of linguistic science.

1. INTRODUCTION

In the study of spoken communication, the gulf between Linguists and Speech Engineers may be clearly demonstrated through their attitudes to the word 'performance'. For linguists, performance is about the act of producing language, of speaking, of performing - it is incidental to and independent of 'competence', knowledge about language structure. To a linguist, performance is a kind of veil, something that hides the true nature of the simple, elegant, parsimonious, universal structures that underly language. For speech engineers on the other hand, performance means degree of success at communication, how readily and accurately a linguistic message can be communicated. To an engineer good performance is primary, and the underlying structures are only interesting to the extent that they make communication more accurate and more reliable. So, if the regularities and constraints in language can be captured and expressed in a mathematical formalism that can exploited to make man-machine communication practical, then it doesn't matter to the engineer that such a formalism allows all kinds of unknown phenomena to occur, or is beyond human cognition. On the other hand, the linguist has a stronger sense of understanding than the engineer, who on the whole is content to communicate word strings (but see [3]).

The differences between linguists and engineers is important because of this rather astonishing observation: *after 40 years of modern linguistics the most successful computational linguistic applications use very little modern linguistic knowledge*. Grammar checkers are based on templates, information retrieval systems treat documents as bags of words, speech recognition systems don't know about phrase structure. The reason? Systems in the knowledge-based tradition have lamentable (engineering) performance: speech recognition systems with 15% word accuracy, parsing systems that only provide acceptable syntactic structures for 60% of sentences found in newspapers, translation systems with outputs that require trained humans to interpret.

This paper is really about what happens when you put performance first. It discusses automatic speech recognition systems and contrasts the conventional symbolic and logico-deductive framework for linguistic decoding with the architectures that engineers (working by trial and error) have found to actually work. In this way I hope that lessons can be drawn from *what the engineers had to do* to solve particular problems.

In the rest of this paper I shall give 10 lessons that can be learned from the experience of building speech recognition systems. The reader is directed to Huckvale [6], for a more detailed discussion.

2. LESSONS

1. Real speech is more variable than you think

The conventional view of phonetic segments (or indeed, handwritten letters) is that they have realised identifiable and distinguishing properties or *features* by which they may be recognised. The engineering view is that segments are associated with regions in a multi-dimensional space formed from a fixed set of continuously-valued observational parameters. This shift from specific features to generic parameters has been necessary owing to two facts: (i) the enormous variability in the realisations of segments according to producer, context, environment and repetition, and (ii) the imperfections of pattern recognition systems for feature detection. Together, these mean that the presence or absence of a detected feature cannot reliably indicate the presence or absence of an underlying segment.

However by modelling the observational path, the engineer can study how realisations of segments cluster in parameter space. Parameters can be chosen to make these clusters small and separated. Variety in realisation can be related to distance in this space, so that any observation can be assessed for similarity to every segment. This brings an important shift in emphasis whereby a segment becomes defined in terms of the other segments, in terms of the space that it doesn't occupy. Segments are no longer identified by their individual character, but by their relative positions within parameter space.

2. Fine detail is irrelevant

Conventionally, human speech perception is considered to be exquisitely sensitive to fine acoustic detail in the signal. Many believe that the special characteristics of auditory processing are an essential part of recognition. However the shift from fine-grained features of a segment to the positioning of a segment in a multi-dimensional space has an important engineering consequence: the space must have relatively few dimensions. Systems tend to use relatively coarse temporal and spectral characteristics of the signal, typically only 10-16 parameters per 10ms.

The reason is that pattern recognition is hard in high dimensional spaces: where each observation is made up from a large number of features. The problem is to do with combining evidence across features: which has the most weight, how should conflicting information be reconciled? In high dimensional spaces the co-variation of features is difficult to establish, and so distances can not be measured reliably.

Not co-incidentally, a rough spectro-temporal representation is likely to be robust to channel effects and speaker differences. Support for such simplicity of acoustic coding is also found from human performance with cochlear implants and from speech coding systems.

3. Low-level performance is worse than you think

Conventionally, it is assumed that human segmental recognition performance is very high. There are a number of reasons to suspect that where higher level linguistic constraints are not applicable, human performance is actually not that good: consider your difficulty in recognising people's names, or the problems you have even segmenting fluently-spoken foreign language. An ability to transcribe nonsense words is not counter evidence since it could still rely on a great deal of phonological knowledge. The engineers have found this too - the best phone recognition rate without phone-sequence constraints seems to be about 70%.

In conventional terms, a 70% correct phone transcription would be a disaster because it would lead to an explosion of word and sentence hypotheses. However engineers have learned to expect poor low-level performance, and to accommodate it by postponing decisions on phone identity. By delaying phonetic transcription until after the word sequence is identified, poor segment-level performance can be converted to good word-level performance using higher constraints. And good word-level performance can lead to artificially high phonetic transcription performance. This is the engineers' explanation for the apparent human transcription skill.

4. Speakers are more similar to others than themselves

The conventional view is that speaker differences are so large that they need to be 'normalised' to a standard speaker prior to phonetic processing [2]. However in speech recognition systems, it is remarkable how little attention is paid to speaker differences: not only are the models of segments built without regard to speaker, the recognition system does not even impose constraints of speaker continuity through an utterance. However, these systems do have and seem to need different models for segments depending on the phonetic context in which they occur. The corollary of this must be that contextual variability is bigger than speaker variability: the realisations of a segment across speakers in one context are more similar than realisations across contexts in one speaker.

This is not to say that adaptation to speakers at multiple linguistic levels is unnecessary, just that it is much less important than other sources of variability.

5. The consequence of obscurity is ambiguity

Conventionally the view is that some sections of utterances are 'clearer' than others, that some sections are 'more important', that some sections seem to be 'emphasised' by the speaker and 'attended to' better by the listener. There are many observers of these phenomena who therefore conclude that recognisers need to target such sections specifically. They suggest that lexical access is driven by information from stressed syllables, and hence these are 'landmarks' or 'islands of reliability' to which more attention should be paid.

Speech engineers have not targeted stressed syllables as deserving of more processing effort. Although some engineers have tried multiple models for stressed and unstressed vowels, the results are not conclusive [4]. What engineers have clearly avoided is putting more weight on the results of recognising a stressed syllable than an unstressed one. This may seem odd if stressed syllables are clearer and more important. An alternative view is that the recognition system already does *exactly the right thing* under these circumstances: in clear regions of speech, the system generates a few good hypotheses, while in unclear regions, the system generates many weak hypotheses [1]. Since the circumstance of a few good hypotheses will constrain lexical access much more than the circumstance of many weak hypotheses, a clear region will carry more weight.

Notice that the recognition system does not need to know in advance that a given piece of speech is clear or not. It does not have 'stressed syllable detectors', nor indeed does it make any early and firm decisions about clarity or importance; instead the system adapts automatically to changes in clarity - *by having the appropriate decoding process*.

6. Segment your knowledge, not your representation

Conventionally speech is considered to consist of a sequence of discrete segments. "How is the word 'cat' made up?" "'k' and 'a' and 't'." Phonological systems are based on phonetic transcription: a system of describing speech sounds as a sequence of symbols from a finite inventory (even metrical phonologies have 'slots' for melodic information).

That words have a regular structure made up from a small number of possible units is probably a necessary consequence of having to remember them. We couldn't possibly remember words as just sounds - who could identify 50,000 different sounds? - so our memory imposes a logical organisation based on a sequential ordering of a very small number of basic distinctions.

But just because our mental organisation has this segmented structure, this does not mean that we need to recognise an utterance as segments. The speech engineers have implemented it this way: the recognition task is to find an underlying segmented representation that is the most likely source of the observed signal. That is, given a model of speech generation that happens to have a segmented input, what do we have to put in to get this utterance out?

The lesson is twofold: firstly that segments are simple-minded but quite workable as underlying entities for modelling pronunciation variation, and secondly that segmentation is only a character of the underlying level and can only be inferred on the signal after recognition. The hadwriting

analogy make this clear: we know that writing is made from 26 discrete symbols, yet connected handwriting doesn't show boundaries and we cannot segment the handwriting without knowing what segments have taken part (i.e. after recognition) [5].

7. Mediocrity at everything

It would be wrong to take as a lesson from the preceding discussion that speech recognition systems have any single linguistic component that describes effectively all necessary phenomena at any one level. It is more the case that the linguistic knowledge in such systems is barely sufficient. Contemporary systems only use bigram models of word order constraints, single pronunciations for words, no modelling of speaker variability at any level, and so on.

Yet if we are to explain the fair performance of such systems, it must be because although they may be mediocre, they are *comprehensively mediocre*. In other words they cover all aspects of the recognition problem even if rather poorly. This comprehensive quality is essential for a system that is robust, that doesn't fail for 'non-grammatical' sentences, for intrusive noises, for odd word pronunciations, for disfluencies.

Consider a comprehensive model of syntax. It has already been indicated that current parsers only assign reasonable structures to 60% of newspaper sentences. So such a parser is useless to filter out possible from impossible sentence hypotheses - even if the word accuracy was extremely high, the system would have very poor overall performance if it only relied on the syntactic judgements of modern parsers. This is why speech recognition systems, even today, use n-gram probabilities instead. For a sentence hypothesis, the likelihood of the sentence can be calculated from the word frequencies and the word transition probabilities. All sentences can be processed in this way, and all hypotheses can be ranked as (relatively) good or bad. While a probabilistic scheme may only give good average estimates of worth, a true parser only needs to make one poor judgement to ruin a whole recognition.

8. Unity is strength!

Conventional levels of linguistic description and processing have very different purposes. For example, the role of phonology is to parse transcription, the role of the lexicon is to find words which match the phonological sequence, the role of syntax to erect phrase structure, the role of semantics to form a logical expression.

Engineering systems have linguistic levels too, but they have a common purpose: estimating probabilities. The role of phonology is to predict the likelihood that a phonetic representation is possible for a word, the role of syntax is to predict the likelihood that a syntactic structure is possible for a word sequence, the role of semantics is to predict the likelihood that a sentence has a certain interpretation.

This design arises from the need to balance constraints at the different levels - under errorful input which constraint is best to break: a phonotactic one or a syntactic one? To reconcile constraints across the disparate theories and models of linguistics requires an underlying common purpose to those theories. Speech recognition systems make this common purpose explicit: the role of the theories, models, constraints and knowledge is to estimate the probability that a certain structural relationship could occur. In this way the probability that a formant frequency is 100Hz higher than normal can be balanced against the probability that 'red' is more likely to be an adjective than a noun.

Significantly, there has been a recent shift towards the influence of pragmatics - the study of language use - in linguistics and speech perception. It may be possible that 'ensuring the correct interpretation' will be the common purpose for a future linguistic science.

9. Evaluate don't analyse

Perhaps the reason it took so long for integrated, probabilistic speech recognition systems to replace knowledge-based ones, is that the mechanisms by which knowledge is represented and used in probabilistic systems seems backwards. Modern systems contain knowledge about speech production, not speech perception. They contain the probability that a word might follow another word, not the probability that a given utterance might contain that sequence. They use the probability that a segment might give rise to a spectral vector, not the likelihood

that a spectral vector is evidence for a particular segment.

In a speech recognition system, hypotheses are not generated on the basis of analytical knowledge (words from segments, phrases from words). Instead the knowledge is only there to evaluate hypotheses generated by the search engine.

While most of modern linguistics has been in the generative tradition, in fact it is the engineers that have explained why this is necessary. Recognition takes place at the lowest levels: in terms of spectra predicted and observed. Systems need to be able to predict the acoustic form of any utterance so that all linguistic information can be exploited in recognition.

10. Recognition is just optimisation

The ability to use a speech production model for recognition is built on the ability to find, for any input utterance, the most likely input to the model that would have produced such an utterance. The most likely input is also, of course, the best interpretation given all the knowledge available.

In ASR the fundamental support for such a scheme comes from a graph search algorithm which effectively evaluates all possible utterances. Every syntactic path, every word, every pronunciation, every acoustic realisation of every possible utterance is weighed up to find the one that best fits the observation. For the kinds of recognition systems we are discussing the search space is enormous, and the number of possible paths is vast. Yet the simple principle of optimality that underlies the graph search makes the search for the single best path highly efficient.

Thus the final lesson we should take from speech recognition systems is that we should not miss the opportunity to apply *all* the knowledge we have at any one time to the decoding of a single spectrum. The identification of the lowest allophonic variation can be influenced by semantic context. The best interpretation of an utterance is the one that fits best with all of what we know and what we expect.

This is important because although graph-search seems an unlikely cognitive mechanism, when viewed as constraint satisfaction recognition is really just optimisation. Such optimisation problems are solved by physical systems, plants and animals all the time. There is no need for computation at all - just thermodynamics.

Speech recognition engineers have perhaps stumbled on a major cognitive discovery: not only that everything can make a difference, but that everything can be taken into account in a recognition framework based on optimisation. The size-complexity of such a task needn't prevent the right algorithm on the right kind of processing device searching the entire space of possibilities. Such an engineering solution could equally apply to the cognitive processing of language by people.

REFERENCES

[1] Bridle, J.S., Personal communication.

[2] Disner, S. (1980), "Evaluation of vowel normalisation procedures", JASA 67 pp253-61.

[3] Gorin, A (1995), "On automated language acquisition", JASA 97 pp3441-3461.

[4] Hieronymus, J.L., McKelvie, D., McInnes, F.R. (1992), "Use of acoustic sentence level and lexical stress in HSMM speech recognition", in IEEE ICASSP-92, pp225-228.

[5] Huckvale, M.A. (1992), "Illustrating speech: analogies between speaking and writing", *Speech Hearing and Language - Work in Progress 6*, Phonetics and Linguistics, UCL.

[6] Huckvale, M.A. (1996), "Learning from the experience of building automatic speech recognition systems", *Speech Hearing and Language - Work in Progress 9*, Phonetics and Linguistics, UCL.