

Modelling Human Speech Recognition using Automatic Speech Recognition Paradigms in SpeM

Odette Scharenborg¹, James M. McQueen², Louis ten Bosch¹, and Dennis Norris³

¹A²RT, Department of Language and Speech, University of Nijmegen, The Netherlands

²Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

³Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK

O.Scharenborg@let.kun.nl

Abstract

We have recently developed a new model of human speech recognition, based on automatic speech recognition techniques [1]. The present paper has two goals. First, we show that the new model performs well in the recognition of lexically ambiguous input. These demonstrations suggest that the model is able to operate in the same optimal way as human listeners. Second, we discuss how to relate the behaviour of a recogniser, designed to discover the optimum path through a word lattice, to data from human listening experiments. We argue that this requires a metric that combines both path-based and word-based measures of recognition performance. The combined metric varies continuously as the input speech signal unfolds over time.

1. Introduction

The SPEech-based Model of human speech recognition (SpeM [1]) is based on procedures and techniques used in automatic speech recognition (ASR), but attempts to account for the performance of human listeners. SpeM therefore implements the same core theoretical assumptions about human speech recognition (HSR) as are implemented in the HSR model Shortlist [2,3]. SpeM is an advance on Shortlist in at least two ways (see [1] for further details). First, SpeM can take real speech as input, while the input of Shortlist consists of an error-free string of discrete phonemes. Second, SpeM can deal with the pronunciation variants in real speech caused by processes such as insertion and deletion. The lexical search process in Shortlist is unable to deal with a mismatch between the number of phones in the input and the number of phones stored in the canonical pronunciations stored in the Shortlist lexicon.

In the present paper, we show that SpeM is able to account for key aspects of human listening ability. We compare its performance to that of the Shortlist model, and show that SpeM, like Shortlist before it, can recognise the words in stretches of speech that are lexically ambiguous. Most data on human spoken word recognition involves measures of how quickly or accurately words can be identified. A central requirement of any model of human speech recognition is therefore that it should be able to provide a continuous measure (usually referred to as 'activation' in the psychological literature) of how easy each word will be for listeners to identify. We address the problem of relating the performance of a path-based model of continuous speech recognition to word-based data from psycholinguistic experiments.

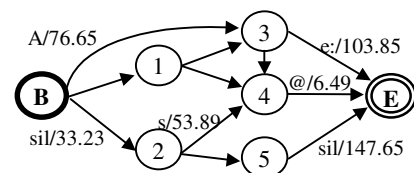
2. SpeM

SpeM segments continuous speech fragments by finding the optimal path through a lexicon and the input using a dynamic programming algorithm. For the word search process, SpeM uses a lexicon containing the words it should be able to recognise. Internally, the lexicon is transformed into a lexical tree in which the entries share common beginnings of phones and each path through the tree represents a word. The tree has one root node and as many end nodes as there are words in the lexicon.

The input for SpeM consists of a phone graph. Figure 1 shows a graphical representation of an input phone graph. A graph has one begin node ('B') and one end node ('E'). Each arc (connection between two nodes) carries a phone and its bottom-up evidence in the acoustic signal (acoustic cost). For the sake of clarity, not all phones and their acoustic costs are shown. This input can be provided by an automatic phone recogniser (APR), which takes real speech as input and generates probabilistic phone graphs such as that shown in Figure 1. In the simulations in the present paper, the inputs were the 1-best outputs of an ideal (flawless) APR.

The search for the best-matching sequence of words for a given input in SpeM is the search for the cheapest path through the graph that is the product of the input phone graph and the lexical tree. As described in [1], the total cost of each path is the sum of a number of costs, including, critically, the acoustic cost of that path (i.e., the negative log likelihood determined by the APR), costs associated with mismatches between the input and the lexical tree due to phone insertions, deletions, and substitutions, the Possible Word Constraint cost [4], and a word entrance penalty (the cost associated with starting a new word). As also described in [1], the model generates a set of N-best paths, after pruning mechanisms have removed duplicate and/or improbable paths.

Figure 1. Graphical representation of an input phone graph in SpeM.



3. Lexically ambiguous utterances

3.1. Materials

We examined the behaviour of SpeM when it was confronted with the following lexically ambiguous utterances (in DISC notation):

1. [k{t@IQg] (orthography: catalogue)
2. [SIpINkw2@ri] (orthography: ship inquiry)
3. [SIpI?kw2@ri] (orthography: ship i?quiry)

The first input is either the transcription of the word *catalogue*, or of the sequence of words *cat a log*. Human listeners, and the original Shortlist model [2], will recognise this sequence of phones as the longer word. Will SpeM be able to correctly parse such sequences?

In the second input, the continuous speech fragment consists of the phonemic transcriptions of the words *ship* and *inquiry*. In the third input, the [N] of *inquiry* is replaced by an ambiguous phone: [?]. (The phone [?] matches all other phones with the same small cost.) This sequence is again lexically ambiguous, at least until the penultimate phoneme: up to the [r], the sequence is consistent with the interpretation *shipping choir*. Furthermore, there is more support in this sequence for the incorrect lexical hypothesis *shipping* than for the correct hypothesis *ship*. Will SpeM be able to parse this input correctly, like human listeners? That is, will the model be able to use the information at the end of the sequence to correctly interpret the information at the beginning of the sequence, even though the lexical hypotheses (*ship* and *inquiry*) do not overlap in time? The ability of listeners to recognise such sequences has been taken as evidence that HSR entails the relative evaluation of multiple lexical hypotheses beginning at different points in the signal [2], and experimental evidence supports this observation [4].

The lexicon used in these simulations is identical to that used in the Shortlist simulations in [3]. Each word has one canonical phonemic representation, totalling 26,449 entries.

3.2. Calculating word activation

Shortlist provides a straightforward measure of human performance. It is a localist connectionist model, with separate nodes for each candidate word involved in the current lexical search [2]. Each of these nodes has an activation value, which changes over time as more of the input is processed. In Shortlist, therefore, word activation provides a time-varying measure of the strength of different lexical hypotheses. Word activation can be directly compared to performance by listeners in psycholinguistic experiments, where they are often required to make word-based decisions. How then can the path-based analysis in SpeM be related to human performance, and to the performance of Shortlist?

Although a word choice is implicit in the choice of the best path in SpeM, the total score of a path (the path score hereafter) does not give us direct estimates of the activation of individual words. Furthermore, the path score is computed incrementally, as the input unfolds over time. Therefore, as the *ship/shipping* example makes clear, words on the best path at one point during the input may not be on the best path at a later time. But the best path does indicate *which* words are most likely to be in the input.

The main problem with using the path score as the measure of word activation, however, is that the path score can be dramatically reduced by the presence of a single difficult to recognise word early on in an utterance. It is unlikely that such words make clearly spoken words later in the utterance harder to identify. For example, when the input contains an error or an ambiguous phone (e.g., Input 3), and if the word activation is based on the path score, the probability of both *inquiry* and *ship* will be lower than in the case where there is perfect input (i.e., Input 2). While this is plausible for the degraded word *inquiry* itself, it is not a satisfactory account for the word *ship*, since there is no error in the input with respect to *ship*. A second problem with the path score is that although it gives us an incremental measure of path likelihood, it does not give an incremental measure of the activation of individual words.

In addition to the total score of a path, SpeM also provides the bottom-up evidence for each word in the stretch of the input it corresponds to. The bottom-up evidence increases over time while the candidate word matches the input and it is directly related to whether or not there is a match between the input and the candidate word. For example, when encountering [@] in Input 1, there is no longer a match between the candidate word *cat* and the input. It is only during the input [k{t] that the word *cat* has bottom-up support. Since it should also be possible for a candidate word to be activated *after* the word's offset in the input, the word's bottom-up evidence should also not be used as the measure of the activation of the word.

In order to obtain a measure of word activation that is both incremental at the level of the word and will provide an activation measure after the word's offset in the input, both the path score and the bottom-up evidence of the word should be taken into account. However, there is another problem with the bottom-up evidence of a word and the path score. Both measures are denoted in posteriori log probabilities. These probabilities 1) decrease over time even when there is a perfect match between the input and the candidate word, whereas activation always increases in this case; 2) the lower the log probability, the more likely the candidate is, whereas in terms of word activation, the higher the activation, the more likely the word is.

A measure based on log probabilities can be converted into a measure that increases over time (when there is a perfect match between the input and the candidate word), and where the most likely word (or path) has the highest activation. This new measure is called the Bayesian activation (Act_B). The Bayesian activation is based on the costs mentioned in Section 2 above (without the cost to start a new word) and is calculated for both paths and words.

Our measure of word activation ($Act(Word)$) then, is based on the product of the Bayesian activation of the word ($Act_B(Word)$) and the Bayesian activation of each path ($Act_B(Path)$) the word lies on:

$$Act(Word) = Act_B(Word) \bullet Act_B(Path) \quad (1)$$

This measure takes both path and individual word scores into account, and therefore does not suffer from the same problems as the purely path-based and purely word-based measures of word activation.

4. Results

The results of the ‘catalogue’ and ‘ship inquiry’ simulations are shown in Figure 2-4. In all figures, the upper panel shows the raw bottom-up scores of the candidate words (acoustic score plus costs associated with insertions, deletions, and substitutions) and the middle panel shows the path score. The y-axis of both these panels is in log probabilities. Note that the higher the log probability, the more likely is the candidate word, and that log probabilities decrease over time. Finally, the bottom panel in all figures displays the word activation as calculated using Equation 1. Only the most highly activated words are plotted. The path associated with the activation of those candidate words is shown inside brackets.

Figure 2 shows that, as in the original Shortlist model (see Figure 3, p. 213, in [2]), SpeM prefers the single lexical hypothesis *catalogue* to the alternative parse *cat a log*. The path-based analysis (middle panel) shows that the *cat a log* parse is slightly worse than the *catalogue* parse. They both have the same amount of bottom-up support; however, in the case of *cat a log*, three word entrance penalties are added to the path cost, whereas in the case of *catalogue* only one entrance penalty is added. This implies that sequences of words are less likely than single words, thus that the parse prefers longer segments.

These results show that the model offers an optimal lexical interpretation of the input at any moment in time, just like it appears that human listeners do [5,6].

Figure 2. The raw acoustic scores of the candidate words (upper panel), the total cost of the path (middle panel), and the word activation (bottom panel) when the input was [k{t@lQg].

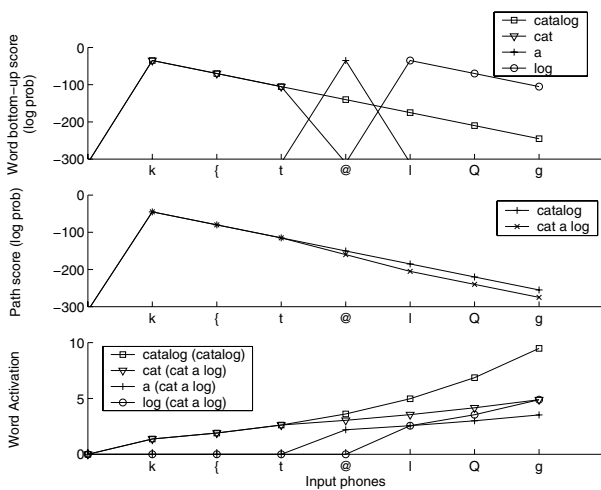


Figure 3 shows the results obtained on the ‘ship inquiry’ simulation given perfect input (Input 2). The upper panel clearly shows that there is more bottom-up evidence for *shipping* (dotted line and triangles) than for *ship* (dotted line and squares): the word bottom-up score of *ship* decreases earlier than the word bottom-up score of *shipping*. Furthermore, the word with the greatest degree of bottom-up support at the end of the input is *inquiry*.

The middle panel of Figure 3 shows that the parses *ship inquiry*, *shipping choir*, and *shipping query* have the same bottom-up evidence until the phoneme [w]. After the [w], the

parse *shipping query* becomes unlikely. (Note that the difference in path costs around the [I] and [N] are related to the word entrance penalty that is added to the path cost of *ship inquiry* at an earlier point in time than for the other two parses.) The penultimate phoneme [r] disambiguates between the *ship inquiry* and *shipping choir* parses. The parse *shipping choir* is penalised for the mismatching phonemes causing its path cost to decrease more than the parse *ship inquiry*.

The word activation of *shipping* (see bottom panel) is indeed higher than the word activation of *ship* until the penultimate phoneme, as is to be expected on the basis of the bottom-up evidence. However, since the parse *ship inquiry* is more likely than *shipping choir* or *shipping query*, the word activation of *ship* is higher than the word activation of *shipping* at the end of the input.

Figure 3. The raw acoustic scores of the candidate words (upper panel), the total cost of the path (middle panel), and the word activation (bottom panel) when the input was [SIpINkw2@ri].

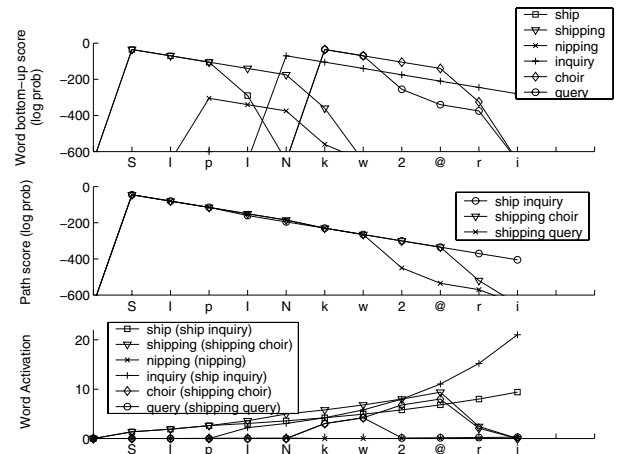


Figure 4. The raw acoustic scores of the candidate words (upper panel), the total cost of the path (middle panel), and the word activation (bottom panel) when the input was [SIpI?kw2@ri].

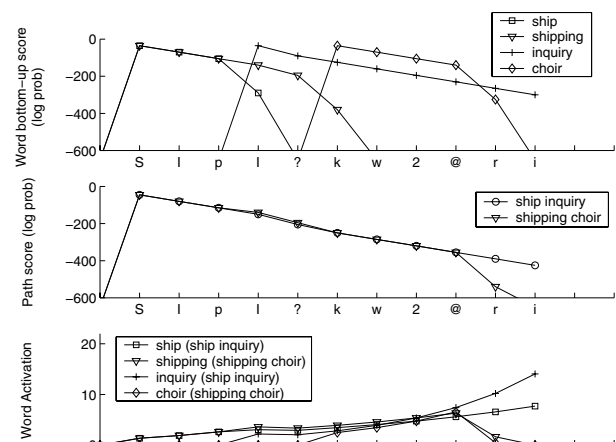


Figure 4 shows a similar picture for the degraded input ‘ship i?quiry’ (Input 3). In spite of the ambiguous second phoneme in the second word, the words *ship* and *inquiry* are the most highly activated lexical candidates at the end of the input. Note, however, that the activation of *ship* is lower at the offset of the ambiguous sequence (bottom panel, Figure 4) than at the offset of the unambiguous sequence (bottom panel, Figure 3), even though the degree of bottom-up support for *ship* was the same in both cases. As we discuss in more detail below, this finding questions the validity of the measure *Act(Word)*. These simulations nevertheless show that SpeM is able to parse correctly lexically ambiguous continuous speech fragments such as ‘ship inquiry’, like human listeners and like the original Shortlist model (see Figure 8, p. 220, in [2]).

5. Discussion and Conclusion

In this paper, we have investigated the modelling of human spoken-word recognition using ASR-based alignment scores in a word search algorithm based on dynamic programming techniques. The resulting model (called SpeM) is able to simulate word activations in accordance with findings based on human speech recognition experiments.

SpeM is still able to correctly parse the input when the input contains an error or an ambiguous phone. From HSR, we know that word activations should increase during the processing of the associated stretch of speech in case of correct input, and that the activations should not vanish after the word’s offset. In this paper, we stipulate that word activation is a function of two factors, namely, the Bayesian word activation (based on the bottom-up evidence in the signal of the word itself), and the Bayesian activation of the search path on which the word occurs. This computational implementation shows adequate but not entirely satisfactory simulation results. The model is able to capture correctly the observations that human listeners perform an optimal analysis of the speech signal and that this analysis changes continuously over time, as more speech is heard [5,6]. But the model makes the counter-intuitive prediction that a word will be less strongly activated when another word in the same path is degraded.

The precise role of path scores in word activation thus needs more investigation. It is clear from our results so far that a psychologically plausible measure of word activation needs to avoid various problems associated with path scores: the fact that path scores decrease in longer utterances, and our finding that poor path scores can reflect unduly on the activation of words which fully match the input. Nevertheless, path scores in a path-based model should play some kind of role in word activation: it is after all the overall better fit of the path *ship inquiry*, relative to the parses beginning with the word *shipping*, that allows *ship* to emerge as the winning candidate, in spite of the overall better bottom-up support for *shipping*.

Our current research is therefore focussing on a theoretical basis for word activation in which path scores are incorporated in a more indirect way in the final measure. One possibility is to relate the HSR notion of ‘word activation’ to the ASR notion of ‘confidence score’. Another possibility that we are considering is one in which word activation is a measure derived by summing over all paths which contain that word, rather than, as in the current implementation, a measure computed separately for each path. We are also

examining whether longer words should be favoured over shorter words in the final measure of lexical activation (cf. the difference between *ship* and *inquiry* in the bottom panel of Figure 3) and how to include measures of word frequency (i.e., prior probabilities of individual words).

The longer-term aim of this research project is to simulate the performance of human listeners in specific psycholinguistic experiments (e.g., [3]). That is, we hope to compare the activation values generated by SpeM to reaction time and error rate measures in word recognition tasks. Such a comparison clearly requires a satisfactory word-based measure of recognition performance. Our current *Act(Word)* measure is a promising step towards this goal, in that it captures the way in which human listeners continuously update their lexical interpretations of continuous speech as the speech signal unfolds over time.

We have shown that SpeM is able to correctly parse lexically ambiguous continuous speech fragments, like human listeners and like the original Shortlist model. Furthermore, since its word search implementation is based on a transparent and computationally elegant dynamic programming technique, it is able to handle insertions and deletions in the input adequately (see [1] for details).

6. Acknowledgements

Part of this work was carried out while the first author was visiting the Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK.

The authors would like to thank Lou Boves for fruitful discussions about this research and Gies Bouwman for his help in implementing SpeM.

7. References

- [1] Scharenborg, O., ten Bosch, L, Boves, L., “Recognising ‘Real-life’ Speech with SpeM: A Speech-based Computational Model of Human Speech Recognition,” To appear in *Proceedings of Eurospeech*, 2003.
- [2] Norris, D., “Shortlist: a Connectionist Model of Continuous Speech Recognition,” *Cognition* 52, 189-234, 1994.
- [3] Norris, D., McQueen, J.M., Cutler, A., Butterfield, S., “The Possible-Word Constraint in the Segmentation of Continuous Speech,” *Cognitive Psychology* 34, 191-243, 1997.
- [4] McQueen, J.M., Norris, D., Cutler, A., “Competition in Spoken Word Recognition: Spotting Words in Other Words,” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 621-638, 1994.
- [5] McQueen, J.M., Dahan, D., Cutler, A., “Continuity and Gradedness in Speech Processing,” in A.S. Meyer & N.O. Schiller (eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, Berlin, Mouton de Gruyter, in press.
- [6] Norris, D., McQueen, J.M., Cutler, A., “Merging Information in Speech Recognition: Feedback is Never Necessary,” *Behavioral and Brain Sciences* 23, 299-325. 2000.