‘EARLY RECOGNITION’ OF WORDS IN CONTINUOUS SPEECH

Odette Scharenborg, Louis ten Bosch, and Lou Boves

A^{2}RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{O.Scharenborg,L.tenBosch,L.Boves}@let.kun.nl

ABSTRACT

In this paper, we present an automatic speech recognition (ASR) system based on the combination of an automatic phone recogniser and a computational model of human speech recognition – SpeM – that is capable of computing ‘word activations’ *during* the recognition process, in addition to doing normal speech recognition, a task in which conventional ASR architectures only provide output after the end of an utterance. We explain the notion of *word activation* and show that it can be used for ‘early recognition’, i.e. recognising a word before the end of the word is available.

Our ASR system was tested on 992 continuous speech utterances, each containing at least one *target word*: a city name of at least two syllables. The results show that early recognition was obtained for 72.8% of the target words that were recognised correctly. Also, it is shown that word activation can be used as an effective confidence measure.

1. INTRODUCTION

Most mainstream automatic speech recognition (ASR) systems use some kind of integrated search: they compute the best path through the complete utterance, and then trace back to identify the words that make up that path. This is contrary to the human capability of recognising (polysyllabic) words well before their acoustic realisation is complete [1]. This human capability is compatible with the view in psycholinguistics that words (or phrases) in the mental lexicon are activated by the acoustic speech input (possibly ‘primed’ by expectations based on linguistic or non-linguistic context, comparable to the [context dependent] prior probabilities in the language model in ASR). As more matching evidence accrues, the activation of the word *increases*. This is contrary to what happens in conventional speech recognisers, where the log-likelihood of words *decreases* as more input frames are processed.

Recognising words before they are complete is very important in human-human communication, for example in turn-taking, and in minimising response latencies. It may also enhance the segmentation of the continuous stream of acoustic information into words, a process that should be easier if the end of words can be predicted. The capability of recognising words on the basis of their initial part certainly helps humans in detecting and processing self-corrections, broken words, repeats, etc. [2]. For these reasons, ‘early recognition’ is a key issue in the IST project COMIC [3].

In [4], we have presented a speech recognition system that is, in principle, capable of providing word activations during the recognition process, in addition to doing ‘normal’ speech recognition where the output is only known after the recognition process. The system proposed in [4] consists of two modules. The first one converts the speech signal into a phone graph; the second parses the graph to detect (sequences of) words. The latter, named SpeM, is a new implementation of *Shortlist*, the computational model of human word recognition proposed by Norris [5]. In this paper, we will show that these activations can be used to recognise polysyllabic words in continuous speech before the complete acoustic signal is available, a capability to which we will refer as ‘early recognition’. In addition, we will show that word activation makes for an attractive confidence measure.

This paper is organised as follows: In section 2, we describe the general architecture of our speech recognition system. Section 3 is devoted to a detailed explanation of how SpeM computes activations of words, and how it can ensure that word activations grow as long as compatible acoustic information enters the system. In section 4, we describe the task used to test SpeM in more detail. Here, we also explain the evaluation procedure developed to test SpeM’s performance in early recognition. Section 5 presents the results of our experiments, which are subsequently discussed in section 6. The paper ends with conclusions and suggestions for future research.

2. THE PROPOSED SYSTEM

The fact that the present implementation of our speech recognition system consists of two modules that operate in sequence is mainly because it originated from a project that investigates the relations between psycholinguistic models of human speech recognition (HSR) and ASR. The first module generates a graph of sub-word units (phones) that forms a symbolic representation of the acoustic signal. The labels on the arcs are enriched with acoustic scores and time instants for the symbols. The second module (SpeM) parses the graph to find the most likely (sequences of) words and converts the accumulated acoustic evidence for the words to activations. In the present implementation, the two modules work in sequence: the word activation module only starts after the graph generation module has processed a complete utterance. However, the theory underlying SpeM is not dependent on the availability of the full graph. Thus, it is straightforward to convert the system to a version that processes speech in a time-synchronous way.

In the present implementation, the sub-word units are context-independent phones. This choice is not essential for the underlying theory of SpeM. The module that generates word activations uses a custom built dynamic programming (DP) routine, but the underlying theory is compatible with standard versions of the dynamic programming algorithm.

Below, we give the relevant details of the automatic phone recogniser (APR) and the search algorithm that was implemented to enable SpeM to function as an automatic speech recogniser and at the same time also as a tool for psycholinguistic research.

2.1 The automatic phone recogniser

2.1.1 Training the APR

The APR is based on the Phicos automatic speech recognition system [6]. Acoustic features are 14 MFCCs (c0…c13), and their first order derivatives, i.e. 28 features. These vectors are based on 16 ms frames and a 10 ms frame shift. 37 context-independent phone models, one noise, and one silence model were trained on 25,104 utterances (81,090 words, corresponding to 8.9 hours of speech excluding leading, utterance internal, and trailing silent portions of the recordings) selected from the VIOS database that consists of telephone calls recorded with the public transport information system OVIS [7]. The speech is extemporaneous.

All phone models and the noise model have a linear left-to-right topology with three pairs of two identical states, one of which can be skipped. For the silence model, a single-state HMM is used. Training was initialised using a linear segmentation of the speech portions of the signal, followed by a number of Viterbi optimisation passes to further train the models. Ultimately, each state comprised a mixture of maximally 32 Gaussian densities [8].

The phone models were trained using a transcription generated by a straightforward look-up of the words in a lexicon of 1,415 entries, including entries for background noise and filled pauses. For each word, the lexicon contains a single unique phonemic representation, corresponding to the canonical (citation) pronunciation. So, pronunciation variation is not taken into account during the training.

2.1.2 The APR during recognition

The lexicon used for the phone recognition consists of all Dutch phones, plus one entry for background noise, and two entries for filled pauses yielding 40 entries in total. During recognition, the APR uses a uni- and bigram phonotactic model (PM) trained on the phonemic transcriptions of the training material.

2.2 SpeM

In SpeM [4], the best-matching sequence of words for a given input is defined as the cheapest path through the product graph of an input phone graph and a lexical tree. The input graph is an a-cyclic directed connected graph with one root node and one end node. Each arc carries a phone and its bottom-up evidence in the acoustic signal (acoustic cost) calculated by the APR. In the lexical tree, entries share common phone prefixes (called word-initial *cohorts* [1]), and each complete 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 total cost of each path is composed of a number of costs: (1) the acoustic cost of that path (i.e., the negative log likelihood determined by the APR), (2) costs of a mismatch between the input and the lexical tree due to phone insertions, deletions, and substitutions, (3) a word entrance penalty (the cost of starting a new word), and (4) a cost associated with the *Possible Word Constraint* (PWC) [9]. The PWC is related to whether a (sequence of) phone(s) occurring between the word and a boundary is phonotactically well formed (being a possible word) or not. Each of the costs can be tuned separately for the task at hand.

The present implementation of SpeM does not include a language model (LM), mainly because the original implementation of Shortlist did not have one. However, it is easy to incorporate an N-gram LM. The search implemented in SpeM is a Viterbi-like DP – time-synchronous and breadth-first.

A potential advantage of SpeM in an ASR task is its capability of giving a ranked list of the most likely words before the end of the word. At each point in time (a node in the input graph) such a list can be created. Other approaches, like Weighted Finite State Transducers [10] or conventional HMMs, only produce output after all input has been processed. A second potential advantage of SpeM is its use of an explicit cost for deletions, insertions, and substitutions, each of which can be tuned individually. In models where these costs are not explicit, it is difficult to obtain more insight in the modelling of these phenomena. However, the difference between SpeM and other approaches that is most important for this research is that SpeM computes *activations* for each word and each path. The next section deals with the notion of activation and how it is computed.

3. ACTIVATION

The measure of *word activation* implemented in SpeM was designed to simulate experimental results of human word recognition experiments. The way it is implemented is also closely related to the notion of confidence measures in ASR. To make for a useful measure, word activation must have a number of properties:

- The word that matches the input best must have the highest activation.
- The activation of a word that matches the input must increase while processing of the input proceeds.
- The activation must be appropriately normalised; the activation should be a measure that is meaningful when comparing multiple concurrent word candidates on the one hand, and words corresponding to different segmentations on the other.

In SpeM, the model for the activation of a word *W* is based on the conditional probability *P*(*W*|*X*), where *W* denotes a certain word and *X* denotes the speech signal. Following Bayes' rule, we obtain:

$$P(W \mid X) = \frac{P(X \mid W) \bullet P(W)}{P(X)} \qquad (1)$$

However, since we also want to deal with incomplete acoustic input, (1) is changed into (2):

$$P(W(n) \mid X(t)) = \frac{P(X(t) \mid W(n)) \bullet P(W(n))}{P(X(t))} \qquad (2)$$

where *W*(*n*) denotes a phone sequence of length *n*, and *X*(*t*) is the gated signal *X* until time *t* [corresponding to *W*(*n*)]. *P*(*W*(*n*)) denotes the prior probability of *W*(*n*); *P*(*X*(*t*)) denotes the prior probability of observing the gated signal *X*(*t*). *W*(*n*) may for example be /Amst@/, i.e. the word-initial cohort of the word 'amsterdam'.

In order to balance the weighting between acoustic scores and language model scores, ASR usually takes a language model factor γ into account, which turns equation (2) into (3):

$$P(W(n) \mid X(t)) = \frac{P(X(t) \mid W(n)) \bullet (P(W(n)))^{\gamma}}{P(X(t))} \qquad (3)$$

The conditional probability log(*P*(*X*(*t*)|*W*(*n*))), which is defined in (4) is delivered by the APR:

$$P(X(t) \mid W(n)) = e^{-a \bullet TPC} \qquad (4)$$

where *TPC* is the total path cost (i.e., the sum of the arc-cost and the costs associated with insertions, deletions, and substitutions) associated with the path starting from the beginning of the graph up to the node corresponding with instant *t*. The value of *a* determines the impact of the costs of time-aligned hypotheses on the word activation measure, and must be chosen to make P(*W*(*n*)|*X*(*t*)) a useful measure (see below).

In SpeM, the prior P(*X*(*t*)) in the denominator cannot be discarded, because hypotheses covering different numbers of input phones must be compared. The problem of normalisation across different paths is also relevant in other systems [11]. Instead of normalising (3) by the sum of the numerator over all paths, in SpeM, the denominator is estimated by

$$P(X(t)) = D^{\#nodes(t)} \qquad (5)$$

where *D* is a constant (0 < *D* < 1) and *#nodes*(*t*) denotes the number of nodes in the cheapest path up to the node associated with *t* in the input phone graph. In combination with *a*, and to a lesser extent with γ, *D* plays an important role in the behaviour over time of P(*W*(*n*)|*X*(*t*)). The values of *a* and *D* must be determined to make the word activation increase over time as more acoustic evidence for a word becomes available, and it must be useful to compare scores of different paths. Only the ratio is relevant for the functionality of SpeM. The values of *a* and *D*, -0.01 and 0.7 respectively, are determined on the basis of the behaviour of the total path cost of the best hypothesis for all sentences from a small training set.

The current implementation of SpeM makes use of a flat LM, in which each word is allotted the same unigram probability. This probability is implemented as a word entrance penalty that is directly applied during the search through the product lattice. On top of this, the LM is enriched with a penalty for entering a word after leaving a phone sequence that, according to a simple heuristic, cannot be a word. (This particular feature of SpeM is elaborated on in another paper [4]).

It is useful to emphasise that in principle SpeM can straightforwardly be endowed with a bigram LM; such an LM can easily be included both from a theoretical and an implementation point of view. The choice for a flat LM means that the parameter γ, which plays an important role in the mainstream ASR approach for balancing the acoustic and language model score, is currently in SpeM of a lesser importance. The most important effect of the LM in the current implementation of SpeM is to favour a parse with fewer (longer) words and cohorts over a parse with more (shorter) words.

Next to the –log probability scores, at each time instant SpeM also outputs a sorted list of word activations as defined by (2), which in turn will be used to determine the ranked list of most likely words.

4. EXPERIMENT

4.1 Method

We used a subset of the VIOS database to investigate SpeM's ability for early recognition of words in continuous speech. The APR created phone graphs, which were subsequently presented to SpeM. For each input node, SpeM created a list of the most likely sequences of words. At the top of this list is the sequence of words that best matched the phonemic representation of the acoustic signal. In this experiment, we only looked at the top two sequences of words at each node in the input graph (i.e. the local winner and the best competitor).

Prior to the experiment, 25 representative utterances were randomly selected as a development set, on which all (see Section 2.2) parameters of SpeM were simultaneously tuned by hand. The parameter settings maximising the number of target words that were correctly recognised by SpeM at the final state of the word were used for the experiment.

The lexicon used by SpeM in the test consisted of 981 entries, including one entry for garbage. This garbage entry matches all phones against the same cost. A sequence of garbage entries is treated as one word (i.e., a single word entrance penalty is added to the total path cost). For each word in the lexicon, only a unique canonical phonemic representation was available.

A set of 318 *target word* types (all polysyllabic station names) was selected for evaluating the system.

4.2 Test material

The corpus used for testing contained 922 utterances (3,299 words) independent from the training corpus. Each utterance ended in a target word. An utterance could contain from two to five words. The total number of target word tokens in the test set was 1,601; thus, some utterances contained multiple target words.

4.3 Evaluation and analysis

For each target word, the *recognition point* was determined; this is the node after which the target word always has the highest activation and therefore is recognised correctly. This node is expressed as a phone position in the phonemic representation of the word. When a word is not recognised correctly, the recognition point is undefined.

#Target tokens	**%At end**		**%Before end**	
	%	*#nb*	*%*	*#nb*
1,601	52.78	845	38.41	615

Table 1. The performance of SpeM on the VIOS test set.

To evaluate the performance of the recognition system, we determined the proportion of the target words that were recognised correctly, and the proportion of the latter subset where the recognition point lies before the end of the word. The recognition point will be analysed in relation to both the length of the canonical phonemic transcription and the uniqueness point of the word, i.e. the phone position in the word after which the word becomes unique given the lexicon. These analyses give us more insight in the latency of early recognition on the one hand, and the potential gain to be obtained from early recognition (in terms of predictability of the trailing portions of polysyllabic words) on the other hand.

Finally, we investigate whether the activation score can be used to identify words for which the recognition hypothesis is likely to be correct. To this end, distributions of the activations of the correctly recognised words at their recognition points and the activations of the incorrectly recognised words at the final state of the input were compared.

5. RESULTS

Table 1 shows the performance of SpeM on the VIOS test set. The first column gives the total number of target word tokens in the test set. Columns 2 and 3 show the percentage ('%') and the total number ('#nb') of correctly recognised words at and before the end of the word, respectively.

Table 1 shows that the performance of SpeM as an ASR is not yet very good. However, 73% (615 / 845) of the target words that were correctly recognised were recognised before the end of the word. We will return to this issue in the Discussion.

For each of the 845 target words that were correctly recognised at the end of the word, the recognition point was related to the uniqueness point and the total number of phones of the word. The results are shown in the form of a histogram in Figure 1. The frequency is given along the y-axis. 'UP+N' represents the distance (in phones) between the uniqueness point and the recognition point of the target words. The '0' indicates that the word was correctly recognised before or at the uniqueness point. A correct recognition before the uniqueness point means that

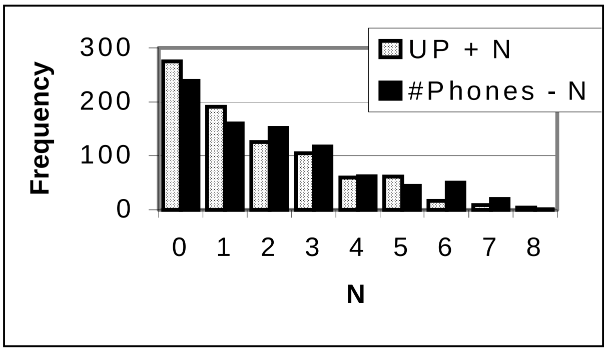

Figure 1. Recognition point related to the uniqueness point (UP+N) and the total number of phones in the word (#Phones-N) for the 845 correctly recognised target words.

a phone string has been recognised that corresponds to the word-initial cohort of the correct word. '#Phones-N' represents the distance (in number of phones (N)) between the last phone in the phonemic canonical transcription of the word and the recognition point. The '0' indicates that the word was correctly recognised at the end of the word.

To be able to interpret the information in Figure 1, it is necessary to know that the uniqueness point of 84% of the target words is at least two phones before the last phone in their phonemic representation. This is due to the fact that all target words are polysyllabic station names; less than 3% of the target words have their uniqueness point at the end of the word. The fact that a substantial proportion of the target words can be recognised up to eight phones before their end is due to the fact that many station names are relatively long, and have their uniqueness point very early in the word. An example of such a long word is 'stadspolders' (/stAtspOLd@Rs/) with its uniqueness point at phone position 3.

From Figure 1 it can be deduced that 55.1% of the total number of recognised target words were recognised before, at, or maximally one phone after the uniqueness point. This indicates that SpeM is able to take advantage of the redundancy caused by the fact that many words in the vocabulary are unique before they are complete.

Figure 2 shows the activation of the correctly recognised target words (solid line) at their recognition points and the activation of incorrectly recognised target words at the last state in the input graph (dotted line) *after* all input has been processed (and the correct word is thus known). The figure suggests that our measure for word activation is a useful measure to indicate the confidence with which words are recognised. However, in order to be able to really use word activation as a confidence value, further research is needed to compare the evolution of word activation in the *course* of the processing of words that end up recognised correctly and incorrectly.

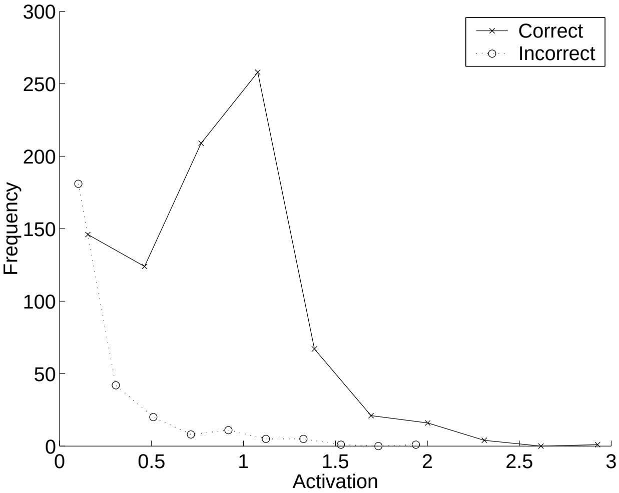

Figure 2. The activation of correctly recognised target words vs. incorrectly recognised target words.

6. DISCUSSION

The performance of the APR+SpeM system as an ASR system cannot be compared directly to results presented for the VIOS database in previous publications (e.g. [8]). There are various reasons for this. First of all, contrary to SpeM, the ASR systems used in previous experiments used bigram language models. Second, the subset of the VIOS test set used in the present study contains the longest utterances, which are most difficult to recognise, while previous results were obtained on the full test set, including a large number of yes/no answers that appear to boost performance substantially. Finally, the present model uses a two-step recognition procedure, while previous results were obtained with an ASR having direct access to the lexicon (and therefore is able to avoid analysing phone sequences that do not occur in the canonical representations of the words).

In the present study, no attempt has been made to maximise the performance of the APR. Quite probably, an APR that misses fewer phones, and perhaps even more importantly, computes more reliable acoustic likelihoods, should allow the combined system of APR+SpeM to reach a performance level comparable to a conventional ASR system. The results presented in [4] already show that SpeM's performance is comparable to that of an off-the-shelf ASR system with a LM in which all words are equally probable.

The results in [4] suggest that the lack of a suitable language model has a substantial impact on the performance of the APR+SpeM as an ASR system. Although one might think that high performance on ASR tasks is not crucial for SpeM's use in psycholinguistic research, this is not the case. SpeM can only simulate the results of psycholinguistic experiments when it is able to recognise (real and artificial) stimuli correctly.

The activation distributions of correctly and incorrectly recognised words in Figure 2 show that *word activation* is a promising concept, not only for early recognition of polysyllabic words, but also as a confidence measure (comparable to the approach taken in [1]). In order to be able to exploit this double asset in speech centric multimodal interaction it will be necessary to change the architecture of dialogue systems to allow for incremental operation. The elegant normalisation potential of word activation applied to a segmental representation might already appear to be a substantial advantage, as soon as the performance of the APR+SpeM system equals that of a conventional ASR system.

The concept of *word activation* can be applied to a number of ASR tasks in conventional architectures, most notably the spotting of key words. The fact that word activation doubles as a confidence measure might make it a very promising approach for searching index terms in spoken documents retrieval when the original recordings are represented in the form of phone graphs.

The concept of *word activation* proposed in this study opens the door towards alternatives for the integrated search that is used in almost all current ASR systems. A search based on word activations should be able to handle many spontaneous speech effects such as hesitations and repetitions that are problematic for integrated search.

7. CONCLUSION AND FUTURE RESEARCH

In this paper, we showed that the automatic speech recognition system consisting of an APR and SpeM is able to recognise words before the end of the word is available. In 72.8% of the correctly recognised target words, the use of local word activation allowed us to identify the word before its last phone was available, and 55.1% of those words was already recognised one phone after the uniqueness point.

The word activation score developed in this paper can also be used as a confidence measure for individual words.

The performance of the APR+SpeM system on this task is (not surprisingly) rather poor. Therefore, the next step in our research plan is to improve the performance of the system. This can be done by incorporating realistic N-gram language models in SpeM and also by improving the performance of the APR.

ACKNOWLEDGEMENT

Part of the research reported in this paper was supported by the IST project COMIC (IST-2001-32311).

REFERENCES

[1] Marslen-Wilson, W.D., Tyler, L., “The Temporal Structure of Spoken Language Understanding,” *Cognition 8*, 1-71, 1980.

[2] Stolcke, A., Shriberg, E., Tür, D., Tür, G., “Modeling the Prosody of Hidden Events for Improved Word Recognition,” *Proc.Eurospeech*, pp. 311-3314, 1999.

[3] den Os, E., Boves, L. “Towards Ambient Intelligence: Multimodal Computers that Understand Our Intentions,” *Proceedings of eChallenges*, Bologna, 22-24 October 2003.

[4] Scharenborg, O., ten Bosch, L., Boves, L., “Recognising ‘Real-life’ Speech with SpeM: A Speech-based Computational Model of Human Speech Recognition,” *Proc. Eurospeech*, pp. 2085-2088, 2003.

[5] Norris, D., “Shortlist: a Connectionist Model of Continuous Speech Recognition,” *Cognition 52*, 189-234, 1994.

[6] Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C., Geller, D., “The Philips Research System for Large-vocabulary Continuous Speech Recognition,” *Proc. Eurospeech*, pp. 2125-2128, 1993.

[7] Strik, H., Russel, A.J.M., van den Heuvel, H., Cucchiarini, C., Boves, L., “A Spoken Dialog System for the Dutch Public Transport Information Service,” *Int. Journal of Speech Technology*, Vol. 2, No. 2, 119-129, 1997.

[8] de Veth, J., “On Speech Sound Model Accuracy,” *Ph.D. Thesis*, University of Nijmegen, The Netherlands, 2001.

[9] 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.

[10] Mohri, M., Pereira, F., Riley, M., “Weighted Finite-State Transducers in Speech Recognition,” *Proceedings of the ITRW Automatic Speech Recognition: Challenges for the New Millenium*, Paris, France, 2000.

[11] Glass, J.R., “A Probabilistic Framework for Segment-based Speech Recognition,” *Computer Speech and Language 17*, 137-152, 2003.

Wessel, F., Schlueter, R., Macherey, K., Ney., H., “Confidence Measures for Large Vocabulary Continuous Speech Recognition,” *IEEE Trans. on Speech and Audio Proc. 9*, No 3., pp. 288-298, 2001.