Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 54–71
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

'Early recognition' of polysyllabic words in continuous speech

Odette Scharenborg *, Louis ten Bosch, Lou Boves

Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 23 September 2004; received in revised form 5 December 2005; accepted 19 December 2005
Available online 23 January 2006

Abstract

Humans are able to recognise a word before its acoustic realisation is complete. This in contrast to conventional automatic speech recognition (ASR) systems, which compute the likelihood of a number of hypothesised word sequences, and identify the words that were recognised on the basis of a trace back of the hypothesis with the highest eventual score, in order to maximise efficiency and performance. In the present paper, we present an ASR system, SpeM, based on principles known from the field of human word recognition that is able to model the human capability of 'early recognition' by computing word activation scores (based on negative log likelihood scores) during the speech recognition process.

Experiments on 1463 polysyllabic words in 885 utterances showed that 64.0% (936) of these polysyllabic words were recognised correctly at the end of the utterance. For 81.1% of the 936 correctly recognised polysyllabic words the local word activation allowed us to identify the word before its last phone was available, and 64.1% of those words were already identified one phone after their lexical uniqueness point.

We investigated two types of predictors for deciding whether a word is considered as recognised before the end of its acoustic realisation. The first type is related to the absolute and relative values of the word activation, which trade false acceptances for false rejections. The second type of predictor is related to the number of phones of the word that have already been processed and the number of phones that remain until the end of the word. The results showed that SpeM's performance increases if the amount of acoustic evidence in support of a word increases and the risk of future mismatches decreases.

1. Introduction

Most theories of human speech recognition (HSR; Gaskell and Marslen-Wilson, 1997; Luce et al., 2000; McClelland and Elman, 1986; Norris, 1994) assume that human listeners first map the incoming acoustic signal onto prelexical representations (e.g., in the form of phonemes or features) and that these resulting discrete symbolic representations are then matched against the words in an internal lexicon. In general terms, this is not unlike the way automatic speech recognition (ASR) systems operate, although most mainstream ASR systems avoid an explicit representation of the prelexical level to prevent early decisions that might incur

* Corresponding author. Tel.: +31 24 36 11644; fax: +31 24 36 12907.
E-mail address: O.Scharenborg@let.ru.nl (O. Scharenborg).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.12.001

irrecoverable errors. Human listeners are able to recognise (long and frequent) words reliably even before the corresponding acoustic signal is complete. According to theories of HSR, human listeners compute a word activation measure (indicating the extent to which a word is activated based on the speech signal and the context) as the speech comes in and presumably make a decision as soon as the activation of a word is high enough, often before all acoustic information of the word is available (Marslen-Wilson, 1987; Marslen-Wilson and Tyler, 1980; Radeau et al., 2000). This suggests that the lexical search performed by human listeners and ASR systems is organised quite differently. ASR systems postpone final decisions as long as possible (i.e., until additional input data can no longer affect the result), in order to avoid premature decisions, the results of which may affect the recognition of following words.

Marslen-Wilson (1987) coined the term *early selection* for the ''reliable identification of spoken words, in utterance contexts, *before* sufficient acoustic-phonetic information has become available to allow correct identification on that basis alone.'' He reviews a number of gating experiments (a word is being presented in segments of increasing duration, and subjects are asked to identify the word being presented and to give a confidence rating after each segment) and monitoring experiments (detection of a target sequence, which may be embedded in a sentence or list of words/nonwords, or in a single word or nonword) in the context of early selection. On the basis of the results of these experiments, he concluded that in normal speech recognition, content words heard in an utterance context can be selected and recognised earlier than would be possible if just the acoustic input was being taken into account.

Identifying and recognising words before their acoustic realisation is complete is important in human-human communication, for example for adequate turn-taking in a dialogue with minimal 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 (Marslen-Wilson, 1987). The capability of recognising words on the basis of their initial part certainly helps human listeners in detecting and processing self-corrections, broken words, repeats, etc. (Stolcke et al., 1999). In this paper, our goal is to model the capability of recognising a word before its acoustic off-set in an ASR system after human speech recognition.

This paper introduces the concept of '*early recognition*', i.e., the reliable identification of spoken words *before* the end of their acoustic realisation, but *after* the uniqueness point (UP) of the word (given the lexicon). The restriction to recognition at or after the uniqueness point allows us to focus on acoustic recognition, with only a small impact of a language model, which would be comparable – but certainly not identical – to the contexts used in human word recognition in Marslen-Wilson's definition of 'early selection'.

If one wants to model early recognition in ASR after human speech recognition, one needs to develop an ASR system that is able to produce a measure analogous to the word activation measure – as used by human listeners – that can be computed on-line, as additional speech comes in. In Scharenborg et al. (2003a,b, 2005), we have presented a speech recognition system called SpeM (SPEech based Model of human speech recognition) that is indeed capable of providing 'word activations' that are derived from the log-likelihood values in conventional ASR systems. Since the procedure that converts log-likelihoods into word activations is based on Bayes' Rule, we use the term 'Bayesian activation' along with the more general term 'word activation' (Section 3). The SpeM system consists of three modules: The first converts the speech signal into a phone graph; the second parses the graph to detect (sequences of) words; the third makes decisions about the recognition of words as more acoustic evidence comes in (Section 2). Furthermore, during the lexical search, SpeM provides a list of the most likely path hypotheses at every phone node in the phone graph. This enables SpeM to recognise and accept words *before* the end of an utterance or phrase.

In previous papers (Scharenborg et al., 2003b, 2005), we investigated the performance of SpeM as a standard speech recognition system, which makes decisions about the identity of the words (spoken mainly in isolation) it has recognised after the complete signal has been processed. In this paper, we extend this research by investigating SpeM's capability for early recognition of spoken words. For standard speech recognition, it suffices to search for the best-scoring path through the search space spanned by the language model, the lexicon, and the acoustic input. In early recognition, on the other hand, an additional decision procedure is needed for accepting a word as being recognised if its local word activation fulfils one or more criteria. The point at which this happens is referred to as the 'Decision Point' (DP; Section 6). In order to be able to put the results of SpeM on the task of early recognition into perspective, it is necessary to know how well SpeM performs as a standard ASR system. This issue is taken up in Section 5. In that section, we also define the crucial concept

of the 'Recognition Point' (RP) of a word, and we analyse the location of the RP in the focus words that were recognised correctly. The speech material used in the experiments is briefly described in Section 4.

Early recognition is dependent on the structure and the contents of the lexicon. If a lexicon contains many words that only differ in the last one or two phones, early recognition (on the basis of acoustic input) is more difficult than when the lexicon mainly consists of words which contain many different phone sequences after the lexical uniqueness point. At the same time, it is evident that making decisions on the basis of only a few phones at the beginning of a long word is more dangerous than deciding on the basis of a longer string of word-initial phones. Therefore, we will investigate the impact of the number of phones before and after the UP on the decision criteria that must be applied to the Bayesian activation in early recognition. This should allow us to draw conclusions about the feasibility of early recognition in an ASR system (Section 6). The results are discussed in Section 7.

2. The recognition system

SpeM was developed to serve as an experimental ASR system and at the same time also as a tool for research in the field of HSR. In fact, it is a new and extended implementation of the theory underlying *Shortlist*, the computational model of human word recognition developed by Norris (1994). Unlike Shortlist and most other computational models of HSR, which take handcrafted symbolic phoneme-like representations of the speech signal as input, SpeM starts from the actual acoustic signal.

SpeM consists of three modules the first two of which work in sequence; the third module is entered each time after a node in the phone graph is processed in the second module (see below and Fig. 1). The first module, the automatic phone recogniser (APR), generates a symbolic representation of the speech signal in the form of a (probabilistic) phone graph (Section 2.1). The second module, the word search module, parses the graph to find the most likely (sequence of) words, and computes for each word its activation based on, among others, the accumulated acoustic evidence for that word (Section 2.2). Below, we give the relevant details of the first two modules. The focus of this paper is on the third module (see Fig. 1), which makes decisions about the recognition of words as more acoustic evidence comes in. This module is explained in detail in Section 6.

The sequential operation of the first two modules should be considered as an implementation detail. It would be easy to change the phone-based architecture of SpeM in such a way that the search module would advance one step each time the APR adds a new node to the phone graph. The essential difference with ASR is that the search module in SpeM depends in a crucial manner on the availability of some kind of prelexical symbolic representation of the speech signal. Consequently, it is not straightforward to implement early recognition as presented here in conventional frame-based ASR systems, since in those systems a prelexical symbolic representation is lacking.

2.1. The automatic phone recogniser

The APR is based on the Phicos ASR system (Steinbiss et al., 1993), but it is easy to build an equivalent module using open source software, such as HTK (Young et al., 2002). For the experiments reported in this paper, 37 context-independent phone models, one noise model, and one silence model were trained on 25,104 utterances in Dutch (81,090 words, corresponding to 8.9 h 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 Dutch public transport information system OVIS (Strik et al., 1997). More details about the VIOS database are given in Section 4. 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

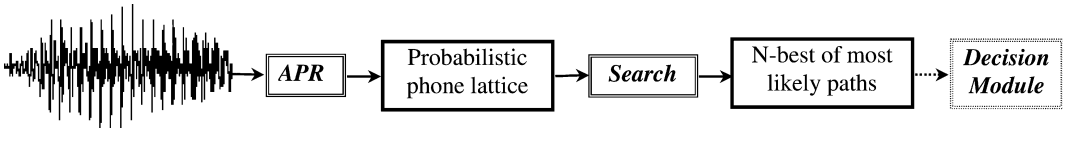

Fig. 1. Overview of the SpeM model and the additional decision module.

single-state hidden Markov Model (HMM) is used. Each state comprises a mixture of maximally 32 Gaussian densities. The phone models have been trained using a transcription generated by a straightforward look-up of the phonemic transcriptions of the words in a lexicon of 1415 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. No pronunciation variation is accounted for in the lexicon.

The 'lexicon' used for the phone recognition by the APR consists of 37 Dutch phones and one entry for background noise, yielding 38 entries in total (in the lexicon, no explicit entry for silence is needed). During recognition, the APR uses a bigram phonotactic model trained on the canonical phonemic transcriptions of the training material.

The APR converts the acoustic signal into a weighted probabilistic phone lattice without using lexical knowledge. Fig. 2 shows a simplified weighted phone lattice: The lattice has one root node ('B') and one end node ('E'). Each edge (i.e., connection between two nodes) carries a phone and its bottom-up evidence in terms of negative log likelihood (its acoustic cost). The acoustic cost is directly related to the probability that the acoustic signal X was produced given the phone ($P(X|Ph)$, in which Ph denotes a phone).

2.2. The search module

The input of the search module consists of the probabilistic phone lattice created by the first module and a lexicon represented as a lexical tree.

In the lexical tree, entries share common phone prefixes (called word-initial *cohorts*), and each complete path through the tree represents a pronunciation of a word. Fig. 3 shows an example of a graphical representation of the beginning of a lexical tree for Dutch. The lexical tree has one root node ('B') and as many end nodes as there are words in the lexicon. The hash '#' indicates the end of a word; the phonemic transcription in the box is the phonemic representation of the complete word. Each node in the lexical tree represents a word-initial cohort. The phonemic transcriptions belonging to the word-initial cohorts are not explicitly shown. Note that the word [ɑs] (English 'axle') is an example of an embedded word, since the node labelled with [ɑs] in the lexical tree (Fig. 3, node 2) has outgoing arcs (thus in this case the phonemic transcription [ɑs] also represents a word-initial cohort).

The search for the best-matching sequence of words is in effect the search for the cheapest path through the product graph of the input phone lattice and the lexical tree. It is implemented using dynamic programming (DP) techniques, and is time-synchronous and breadth-first. SpeM calculates scores for each path (the *total cost*), and also a score for the individual words on a path (the *word cost*). The total cost of a path is defined as the accumulation along the path arcs of the bottom-up acoustic cost (as calculated by the APR) and several

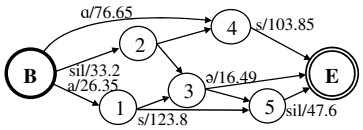

Fig. 2. A graphical representation of a weighted probabilistic input phone lattice. For the sake of clarity, not all phones and acoustic costs are shown.

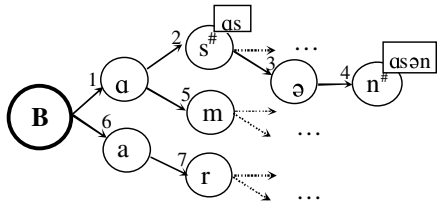

Fig. 3. A graphical representation of the beginning of a lexical tree. The hash '#' indicates the end of a word; the phonemic transcription in the box is the phonemic representation of the complete word. The words indicated in the figure are: [ɑs] ('as', English 'axle', see node 2), and [ɑsən] ('Assen', a Dutch city name, see node 4).

cost factors computed in the search module. SpeM has a number of parameters that affect the costs and that can be tuned individually and in combination. Most of these parameters, e.g., a word entrance penalty (the cost to start hypothesising a new word) and the trade-off between the weights of the bottom-up acoustic cost of the phones and the contribution of the language model, are similar to the parameters in conventional ASR systems. In addition, however, SpeM has two types of parameters that are not usually present in conventional ASR systems. The first novel parameter type is associated to the cost for a symbolic mismatch between the input lattice and the lexical tree due to phone insertions, deletions, and substitutions. Insertions, deletions, and substitutions have their own weight that can be tuned individually. Because the lexical search in SpeM is phone based, mismatches can arise between the phonemic representation of the input in the phone graph and the phonemic transcriptions in the lexical tree. It is therefore necessary to include a mechanism which explicitly adjusts for phone-level insertions, deletions, and substitutions. In mainstream ASR, however, the search space is usually spanned effectively by the combination of the pronunciation variants in the system's dictionary and the system's language model, so that explicit modelling of insertions, deletions, and substitutions on the phone-level is not necessary.

The second novel parameter type is associated to the *possible word constraint* (PWC, Norris et al., 1997). The PWC determines whether a (sequence of) phone(s) that cannot be parsed as a word (i.e., a lexical item) is phonotactically well formed (being a possible word) or not (see also Scharenborg et al., 2003b, 2005). In SpeM, the PWC is implemented using 'garbage' symbols, comparable to the 'acoustic garbage' models in ASR systems. The garbage symbol in SpeM matches all phones with the same cost (note that the acoustic costs of the phones themselves do vary) and is hypothesised whenever an insertion that is not word-internal occurs on a path. A garbage symbol (or an uninterrupted sequence of garbage symbols) is itself regarded as a word, so the word entrance penalty is added to the total cost of the path when garbage appears on that path. The PWC evaluation is applied only to paths on which garbage is hypothesised. Word onsets and offsets, plus utterance onsets and offsets and pauses, count as locations relative to which the viability of each garbage symbol (or sequence of symbols) is evaluated. If there is no vowel in the garbage sequence between any of these locations and a word edge, the parse is penalised and the PWC cost is added to the total cost of the path. For example, consider the utterance ''they met a fourth time'', where the last sound of the word *fourth* is pronounced as [f]. If *fourf* is not stored as a possible pronunciation in the lexicon, a potential parse by the recogniser in terms of lexical items is *they metaphor f time*. Since the phone '*f*' is not a possible word in English, the PWC mechanism penalises this parse, and if the cost of the substitution of [θ] by [f] is less than the PWC cost, the parse yielding the word sequence '*fourth time*' will win. At the same time, it is worth mentioning that the presence of the garbage phones enables SpeM to parse input with broken words and disfluencies, since it provides a mechanism for handling arbitrary phone input (cf., Scharenborg et al. (2005) for more information).

All parameters in SpeM are robust: Even if they are not optimised in combination, SpeM's output does not change significantly if the value of the parameter that was optimised with fixed values of other parameters is changed within reasonable bounds. In this study, the parameters were tuned on an independent tuning set (see Section 4), and subsequently used for processing the test corpus.

SpeM supports the use of unigram and bigram language models, which models the prior probability of observing individual words and of a word given its predecessor. In the experiments reported in this paper, only a unigram language model is used (see also Section 4).

During the recognition process, for each node in the input phone graph, SpeM outputs *N*-best lists consisting of hypothesised word sequences and word activation scores for each of the hypothesised words (see Section 3) on the basis of the phones in the phone graph (thus the stretch of the acoustic signal) that have been processed so far. The order of the parses in the *N*-best list is determined by the total cost of the parses (thus not by the word activation scores). Each parse consists of words, word-initial cohorts, phone sequences, garbage, silence, and any combination of these, except that a word-initial cohort can only occur as the last element in the parse. So, in addition to recognising full words, SpeM is able to recognise partial words. In the *N*-best list, no identical parses exist: Word sequences on different paths that are identical but have different start and end time of the words are treated as the same word sequence (thus timing differences are ignored). That is, we only take the *order* and *identity* of the words into account for pruning the *N*-best lists. The number of hypotheses in the *N*-best list is set at 10, so that SpeM will output the 10 most likely parses for each node in the input phone

graph. Subsequently, the N-best list with the word sequences and their accompanying word activation scores is sent to the decision module that makes decisions about early recognition.

3. The computation of word activation

The functionality of SpeM that is most important here is the computation of *word activation*. The measure of word activation in SpeM was originally designed to simulate experimental results of human word recognition experiments, which show how words are activated over time (Scharenborg et al., 2003a, 2005). In the computation of the word activation, the local negative log-likelihood scores for paths and words on a path are converted into activation scores that obey the following properties. These properties follow from the concept of word activation as it is used in HSR.

- The word that matches the input best, thus having the smallest *word cost* (see Section 2.2), must have the highest activation.
- The activation of a word that matches the input must increase each time an input phone is processed.
- The measure must be appropriately normalised. That is, word activation should be a measure that is meaningful, both for comparing competing word candidates, and for comparing words at different moments in time.

The way SpeM computes word activation is based on the idea that word activation is a measure related to the bottom-up evidence of a word given the acoustic signal: If there is evidence for the word in the acoustic signal, the word should be activated. Activation should also be sensitive to the prior probability of a word – even if this effect was not modelled in the original version of Shortlist (Norris, 1994). This means that the word activation of a word W is closely related to the probability $P(W|X)$ of observing a word W, given the signal X, the cost function maximised in virtually all ASR systems. Thus, it is reasonable to stipulate that the word activation $Act(W|X)$ is a function of $P(W|X)$, and apply the same Bayesian formulae that form the basis of virtually all theories in ASR to estimate $P(W|X)$. This is why we refer to $Act(W|X)$ as the 'Bayesian activation'. It is important to emphasise that the theory underlying word activation does not require that the sum of the activations of all active words should add to some constant (e.g., 1.0, as in probability theory). For the purpose of early recognition it suffices to normalise the activation value in such a manner that (possibly context dependent) decisions can be made. This is reminiscent of what happens in conventional ASR systems.

Following Bayes' Rule, we define the word activation $Act(W|X) = P(W|X)$, which can be written as

$$Act(W|X) = \frac{P(X|W)P(W)}{P(X)}, \tag{1}$$

Since we also want to deal with incompletely processed acoustic input (for early recognition of words), Eq. (1) is extended to

$$Act(W(n)|X(t)) = \frac{P(X(t)|W(n))P(W(n))}{P(X(t))}, \tag{2}$$

where $W(n)$ denotes a phone sequence of length n, corresponding to the *word-initial cohort* of n phones of W. Note that n is discrete because of the segmental representation of the speech signal. $X(t)$ is the gated signal X from the start of $W(n)$ until time t (corresponding to the end of the last phone included in $W(n)$). $P(X(t))$ denotes the prior probability of observing the gated signal $X(t)$. $P(W(n))$ denotes the prior probability of $W(n)$. $W(5)$ may, for example, be /ɑmstə/, i.e., the word-initial cohort of the word 'amsterdam'. In the experiments reported in this paper, $P(W(n))$ is exclusively based on the unigram probability of the word-initial cohorts and the words.

The (unnormalised) conditional probability $P(X(t)|W(n))$ in Eq. (2), is calculated by SpeM as

$$P(X(t)|W(n)) = \mathrm{e}^{-aTC}, \tag{3}$$

where TC is the total bottom-up cost associated with the word starting from the beginning of the word up to the node corresponding to instant t. TC includes not only the acoustic costs in the phone lattice, but also the

costs contributed by substitution, deletion, and insertion of symbols (like the acoustic cost calculated by the APR, TC is a negative log likelihood score). The definition of the total bottom-up cost is such that $TC > 0$. The value of a determines the contribution of the bottom-up acoustic scores to the eventual activation values. The a weights the relative contribution of TC to $Act(W(n)|X(t))$, and therefore balances the contribution of $P(X(t)|W(n))$ compared to $P(W(n))$, so it acts similar to the language model factor in standard ASR systems. To illustrate the effect of a, consider a cohort $W_1(n)$ on one path and a different cohort $W_2(n)$ on a competing path with the same history as $W_1(n)$ ($P(X(t)|\text{history})$ is identical) and an identical LM score ($P(W_1(n)) = P(W_2(n))$). The difference in word activation between $W_1(n)$ and $W_2(n)$ is now completely determined by the difference in acoustic scores $P(X(t)|W_1(n))$ and $P(X(t)|W_2(n))$ between the two words. a is a positive number; its numerical value is determined such that the three properties of word activation introduced at the start of this section will hold. The comparison of the results of HSR experiments and SpeM simulations (Scharenborg et al., 2003a, 2005), yielded a value $a = 0.01$. In the phone graphs generated of our test material by the APR, the average acoustic score (in terms of negative log likelihoods) of a matching phone is 25. In combination with $a = 0.01$, this amounts to $P(X(t)|W(n)) \approx \exp(-0.25) \approx 0.78$, if $W(n)$ is one phone long (i.e., if $n = 1$).

In SpeM, in contrast to conventional ASR systems, the prior $P(X(t))$ in the denominator of Eq. (2) 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 unconventional ASR systems (e.g., Glass, 2003). The denominator, then, is approximated by

$$P(X(t)) = D^{\#nodes(t)}, \tag{4}$$

where D is a constant ($0 < D < 1$) and $\#nodes(t)$ denotes the number of nodes in the cheapest path from the beginning of the word up to the node associated with t in the input phone graph. In combination with a, D plays an important role in the behaviour over time of $Act(W(n)|X(t))$. Once the value of a is fixed, the value of D follows from two constraints: (1) the activation on a matching path should increase; (2) the activation on any mismatching path should decrease. Then it follows from Eq. (2) and these two additional requirements that:

$$e^{-a(avgMismatchPhone+\mathrm{SubC})} \leqslant D \leqslant e^{-a(avgMatchPhone)}, \tag{5}$$

where $avgMismatchPhone$ is the average acoustic cost of a mismatching phone on a competing path, $SubC$ is the cost for a phone substitution, and $avgMatchPhone$ is the average acoustic cost of a matching phone on the first-best path. Because of the way the APR works, the average acoustic cost of a mismatching phone is only marginally smaller than the average acoustic cost of a matching phone. Thus, the difference between ($avgMismatchPhone + SubC$) and $avgMatchPhone$ is essentially determined by the value of $SubC$. The tuning experiments to be described in Section 4 yielded $SubC = 150$. The left-most term in Eq. (5), then, evaluates to $\exp(-0.01(26 + 150)) \approx 0.17$; the rightmost term evaluates to $\exp(-0.01 \cdot 25) \approx 0.78$. We set $D = 0.7$.

Our choice to normalise the Bayesian activation by the expression given by Eq. (4) is based on two considerations. Firstly, given the Bayesian paradigm, it seems attractive to use a measure with the property that logarithmic scores are additive along paths. Let X_1 and X_2 be two stretches of speech such that X_2 starts where X_1 ends, associated with two paths P_1 and P_2 in the phone lattice (such that P_2 starts where P_1 ends), then $\log(P(X_1)) + \log(P(X_2)) = \log(P(X_1{:}X_2))$ (where ':' means 'followed by'). This means that the lengths of X_1 and X_2 are assumed to be independent, which is a plausible assumption. Secondly, the normalisation as given by Eq. (4) is similar to the normalisation that has to be performed in the calculation of confidence measures (e.g., Bouwman et al., 2000; Wessel et al., 2001). In order to be able to compare confidence measures of hypotheses with unequal length, the normalisation must, in some way, take into account the duration of the hypotheses. Eq. (4) can be regarded as a normalisation in which the number of phones is the normalising factor, rather than the number of frames, that is, as a type of normalisation that is more phonetically oriented.

4. Material

There is a considerable phonological overlap among words, because of which any given word is likely to begin and end in the same way as several other words (Luce, 1986). In addition, longer words are likely to

have shorter words embedded within them (McQueen et al., 1995). Consequently, short words are likely to have a UP that is not before the end of the word, making it impossible to recognise the word before its acoustic offset. Furthermore, Grosjean (1985) pointed out that especially function words and short infrequent content words may not even be identified by human listeners until the word following it has been heard. Therefore, in our evaluation of SpeM's ability for early recognition, we focus on polysyllabic content words.

The VIOS training and test corpus consists of utterances taken from dialogs between customers and an automatic timetable information system (Strik et al., 1997). We decided to define a set of 318 polysyllabic station names as *focus* words. From the VIOS database, 1106 utterances (disjoint from the corpus used for training the acoustic phone models) were selected to tune and test SpeM. Each utterance contained two to five words, at least one of which was a focus word (708 utterances contained multiple focus words). 885 utterances of this set (80% of the 1106 utterances) were randomly selected and used as the independent test corpus. The total number of focus words in the test corpus was 1463; 563 utterances contained multiple focus words. The remaining 221 utterances were used as development test set and served to tune the parameters of SpeM (see also Section 2.2). The parameter settings yielding the lowest word error rate (WER) on the development test set were used for the experiment. The WER is defined as

$$\mathrm{WER} = \frac{\#\ \text{insertions} + \#\ \text{deletions} + \#\ \text{substitutions}}{N} \cdot 100\%, \tag{6}$$

The insertions, deletions, and substitutions in Eq. (6) concern words (different from the phone insertions, deletions, and substitutions discussed in the previous sections); N denotes the number of words in the reference transcription.

The lexicon used by SpeM in the test consisted of 980 entries: the 318 polysyllabic station names, additional city names, verbs, numbers, and function words. There are no out-of-vocabulary words. For each word in the lexicon, one unique canonical phonemic representation was available. A unigram language model (LM) was trained on the VIOS training data – the same data that was used for training the acoustic models and the bigram phonotactic model for the APR.

5. Early recognition in SpeM

In this section, we first present the results of an experiment designed to establish the performance of SpeM as a standard ASR system (Section 5.1). The results are presented in terms of WER (for all words in the test set, thus not only the focus words) by taking the best matching sequence of words as calculated by SpeM after processing the entire input and comparing it with the orthographic transcriptions of the test corpus.

Second, we investigate how many of the focus words have a *recognition point* (RP) that is before the end of the word, and thus can, in principle, be recognised before their acoustic offsets during the recognition process (Section 5.2). The RP is defined as the node after which the activation measure of a correct focus word exceeds the activation of all competitors, and remains the highest until the end of the word (after the offset of a word, the word's activation does not change), and is expressed as the position of the corresponding phone in the phonemic (lexical) representation of the word. To determine the RP, we use the Bayesian word activation score to rank path and word hypotheses. In the analysis, we first determine the proportion of the focus words that were recognised correctly at the end of the utterance. Obviously, a word that is not recognised correctly does not have an RP. In our analysis, the RP will be related to both the length of the canonical phonemic representation and the lexical uniqueness point of the word. Prior to the UP, multiple words (in the word-initial cohort) share the same lexical prefix, and therefore cannot be distinguished on the basis of the acoustic evidence.

5.1. The performance of SpeM as a standard speech recognition system

To get an idea of the difficulty of the task that SpeM is facing, we determined the average depth of the input phone graphs. For all graphs, the average depth was computed by dividing the number of arcs by the number of nodes. The sum of these averages was then divided by the total number of phone graphs in the test set. The average depth of all input graphs was 6.3. Thus, on average, SpeM evaluates 6.3 arcs (or phones) at any point in time. Or in other words, each node in the input graph has on average 6.3 outgoing arcs.

Of the 1463 focus words, 64.0% (936 focus words) were recognised correctly at the end of the word. An analysis of the phone graphs revealed that 309 utterances (34.9% of the test utterances) did not contain a path that matched exactly with the canonical representation of the sequence of spoken words. For 95 of these utterances (30.7%), SpeM was able to correctly recognise the (one or more) focus word(s). Thus, SpeM is able to 'repair' part of the deficiencies in the output of the APR.

The WER obtained by SpeM on *all* words in the test material was 40.4%. This result is certainly worse than the best performance of other ASR systems on the VIOS database observed in previous experiments (i.e., WERs in the range of 11.6–16.0% in Kessens et al. (1999, 2003); a WER of 9.9% in Wester (2003)). However, the performance of SpeM as an ASR system cannot be compared directly to results presented for the VIOS database in previous publications. There are a number of reasons for this. First of all, contrary to SpeM, the ASR systems used in previous experiments used bigram language models, while SpeM only used a unigram language model. 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 in which the APR generated many phone sequences that do not occur in the canonical representations of the words in the lexicon of SpeM. In contrast, previous results were obtained with an ASR in which the acoustic signal could be directly matched against the lexicon (and therefore avoided considering 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 acoustic model set of the APR. Quite probably, an APR that computes more accurate acoustic likelihoods should allow SpeM to reach a performance level comparable to a conventional ASR system. The results presented in Scharenborg et al. (2003b, 2005) show that SpeM's performance is comparable to that of an off-the-shelf ASR system (with an LM in which all words are equally probable) when the acoustic model set used to construct the phone graph is optimised for a specific task.

Despite the mediocre performance of SpeM as an ASR system, and although there is still room for improvement of SpeM's performance as a standard ASR system, it is possible to use SpeM to investigate early recognition, since there are a sufficiently large number of words recognised correctly.

5.2. Recognition point analysis

Of the focus words that were recognised correctly, 81.1% had their RP *before* the end of the word (759 of 936 correctly recognised focus words; 51.9% of all focus words).

For the 936 focus words that were ultimately correctly recognised, we related the RP to the UP and to the total number of phones of the word. The results are shown in the form of two histograms in Fig. 4. The frequency is given along the y-axis. In the left panel, the x-axis represents the distance (in phones) between the UP and the RP of the focus words. $N = 0$ means that the word activation exceeded all competitors already at the UP. In the right panel, the x-axis represents the position of the RP (in number of phones (N)) relative to the last phone in the canonical representation of the word. Here, $N = 0$ means that the word activation exceeded the competitors only at the last phone of the word.

For the interpretation of the information in Fig. 4, the phonemic structure of the words in the set of 936 correctly recognised focus words and the position of the UP of the words must be known. This information is shown in Table 1. The first column shows the distance in number of phones between the UP and the end of the word. 'Total-UP' = 0 means that the UP is at the end of the word: The word is embedded in a longer word. Columns 2 and 3 show the number of focus word types and tokens with 'Total-UP' phones between the end of the word and the UP. From Table 1 it can be deduced that the UP of 85.0% of all focus word tokens (1243/1463) is at least two phones before the end of the word; only 2% of the focus word tokens (30/1463) have their UP at the end of the word. The high frequency in the case of $N = 3$ in the right panel of Fig. 4 is due to an idiosyncratic characteristic of the data, which is irrelevant for the task. As can be seen in Table 1, there is a large set of words that have their UP three phones before the end of the word (450).

Combining the information in Fig. 4 and Table 1 reveals that although only 2% of the focus words have their UP at the end of the word, 19.8% (185/936, see right panel of Fig. 4) of the words were only recognised at the end of the word. Apparently, SpeM is not always able to recognise a word before its acoustic offset,

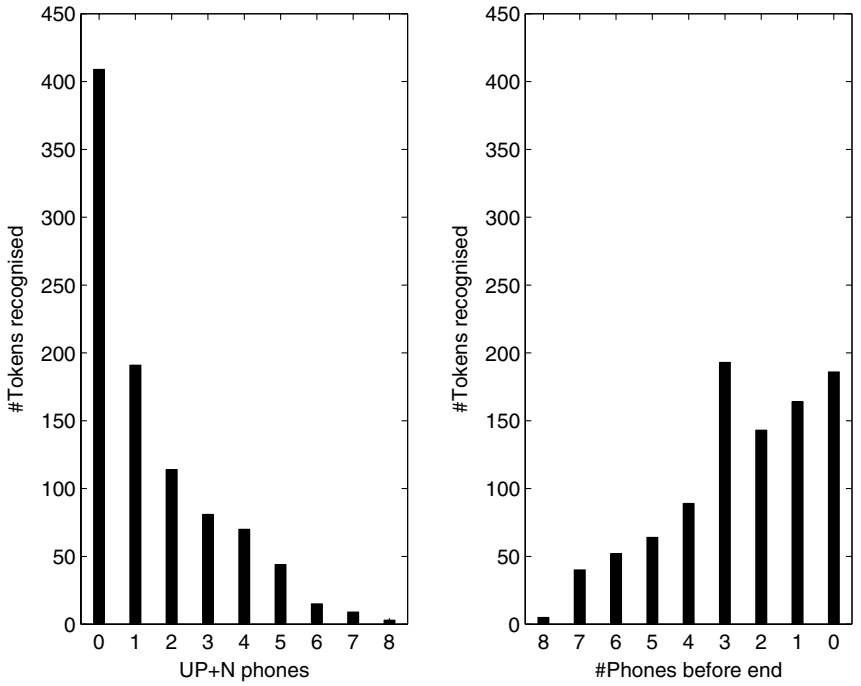

Fig. 4. In the left panel, a histogram relating the recognition point to the uniqueness point ('UP + N phones'); in the right panel, a histogram relating the recognition point to the total number of phones in the word ('#Phones before end') for the 936 correctly recognised focus words.

Table 1
The distribution (in # Types and # Tokens) in number of phones between the UP and the length of a word (Length-UP); Cumulative: # focus word tokens that could in principle be recognised at position length-UP

Length-UP	# Types	# Tokens	Cumulative
0	10	30	1463
1	44	190	1433
2	50	182	1243
3	63	450	1061
4	50	271	611
5	39	186	340
6	38	82	154
7	17	57	72
8	3	11	15
9	2	4	4

despite the fact that the UPs in the set of words were almost always at least one phone before the end of the word. More interestingly, however, from Fig. 4 it can also be deduced that 64.1% (sum of $N = 0$ and $N = 1$, see left panel of Fig. 4) of the total number of recognised focus words were already recognised at, or maximally one phone after the UP. Taking into account that 85.0% of the focus words have at least two phones after their UP, 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.

However, it is obvious that the UP and RP do not coincide for all focus words that were recognised. This can be explained by the fact that a focus word that is correctly recognised at the end of an utterance may not match perfectly with the phone sequence in the phone graph. As indicated before, for 34.9% of the utterances, the canonical phone transcription of the utterance was not present in the phone graph. For these focus words, phone insertion, deletion, and substitution penalties are added to the total score of the word and the path.

Competing words may have a phonemic representation that is similar to the phonemic representation of the correct word sequence. In these cases, it may happen that the best matching word can only be determined after all information of all competing words is available.

6. Predictors for reliable on-line early recognition

The experiment presented in the previous section showed that the word activation of many polysyllabic content words exceeds the activation of all competitors already before the end of the words. However, this does not imply that word activation can be safely used to perform early recognition. If we want to use word activation as a basis for deciding whether a word is considered as recognised before the end of its acoustic realisation, we must develop a decision procedure. To that end, we have experimented with a combination of absolute and relative values of word activation. In addition, we have investigated whether the reliability of early decisions is affected by the number of phones of the word that have already been processed and the number of phones that remain until the end of the word.

In Section 6.1, we explain the decision module that we implemented. The performance of that module will be evaluated in terms of *precision* and *recall*:

Precision. The total number of *correctly* recognised focus words relative to the total number of recognised focus words. Precision measures the trade-off between correctly recognised focus words and false acceptances.

Recall. The total number of *correctly* recognised focus words divided by the total number of focus words in the input. Recall represents the trade-off between correctly recognised focus words and false rejects.

As usual, there is a trade-off between precision and recall. Everything else being equal, increasing recall tends to decrease precision, while increasing precision will tend to decrease recall. We are not primarily interested in optimising SpeM for a specific task in which the relative costs of false acceptances and false rejections can be established, since we are mainly interested in the feasibility of early recognition in an ASR system. Therefore, we decided to refrain from defining a total cost function that combines recall and precision into a single measure that can be optimised.

6.1. Decision module

The input of the decision module consists of the *N*-best list with the word sequences and their accompanying word activation scores as created by the search module at a certain point in time. The decision module only makes a decision about early recognition for focus words. For a focus word to be recognised by SpeM, the following three conditions have to be met:

1. The phone sequence assigned to the focus word is at or beyond the focus word's UP.
2. The *quotient* of the word activation of the focus word on the best-scoring path and the word activation of its closest *competitor* (if present) exceeds a certain threshold (θ).
3. The value of the word activation of the focus word itself should exceed a certain *minimum activation* (Act_{min}).

The second requirement implies that we do not want SpeM to make a decision as long as promising competitors are still alive. The notion that there must be a sufficiently large difference between the first best hypothesis and its runner-up has also been used for a long time in various types of ASR systems to compute a kind of confidence measure (e.g., Brakensiek et al., 2003).

In the SpeM search, two words are said to be in competition if the paths they are on contain an identical sequence of words, except for the word under investigation. Recall that we only look at the *order* and *identity* of the words (see Section 2.2). Thus, two word sequences on two different paths that are identical, but have a different start and end time of the words, are treated as the same word sequence, and cannot be in competition. Furthermore, we only look at the current word; it does not matter whether the paths on which the two competing words lie combine again later on. Fig. 5 illustrates this with an example where the first-best path: [a:vɔnt vo:rbYr*] competes with the path: [a:vɔnt xu:dəm*]. The competitor of [vo:rbYr*] is thus [xu:dəm*]. The asterisk indicates that the processing of a word has not yet reached its last phone.

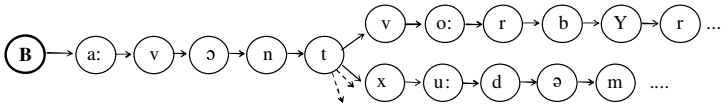

Fig. 5. Two focus words in competition.

Given our definition of 'competitor' it is not guaranteed that all words always have a competitor, because it is possible that all paths in the N-best list are completely disjunct – and so do not share the same history, as is required for being competitor. Absence of a competitor makes the computation of θ impossible. To prevent losing all words without competitors due to a missing value, we accept all focus words without a competitor that appear at least five times (at the same position in the word sequence) in the N-best list. In the experiments described below, we tested various values for θ.

The threshold for the minimum activation (Act_{min}) is introduced because we do not want SpeM to accept a word that happens to have the highest activation, irrespective of the absolute value of the activation. Act_{min} is reminiscent of the graph-based confidence measures introduced in Wessel et al. (2001). In the experiments described below, we tested various values for Act_{min}.

The process of early recognition is schematically depicted in Fig. 6. The word activation of words grows over time as matching evidence is added. Before the word's UP, several words are consistent with the phone sequence; the difference in activation of the individual words in the cohort is caused by the influence of the LM. After a word's UP, it has its own word activation. For the purpose of the experiments in this section, we define the *decision point* (DP) as the point at which a word on the first best path meets the decision criteria described in this section.

6.2. θ and Act_{min} as predictors of early recognition

In this section, we investigate the Bayesian activation as a predictor of early speech recognition as a function of θ and Act_{min}. The value of Act_{min} was varied between 0.0 and 2.0 in 20 equal-sized steps. The value of θ was varied between 0.0 and 3.0 in six equal-sized steps of 0.5. Fig. 7 shows the relation between precision (y-axis) and recall (x-axis) for the combinations of three values of θ, where $\theta = 0.5$, 1.5, 2.5, and 21 values of Act_{min}. For the sake of clarity, Fig. 7 is limited to three values of θ; all other values of θ show the same trend. The left-most symbol on each line corresponds to $Act_{min} = 2.0$; the right-most one corresponds to $Act_{min} = 0.0$.

The results in Fig. 7 are according to our expectation. Recall should be an inverse function of θ: The smaller θ becomes, the less it will function as a filter for words that have a sufficiently high activation, but which still have viable competitors. Similarly for Act_{min}: For higher values of Act_{min}, fewer focus words will have an activation that exceeds Act_{min}, and thus fewer words are recognised. These results indicate that the Bayesian activation can be used as a predictor for the early recognition of polysyllabic words.

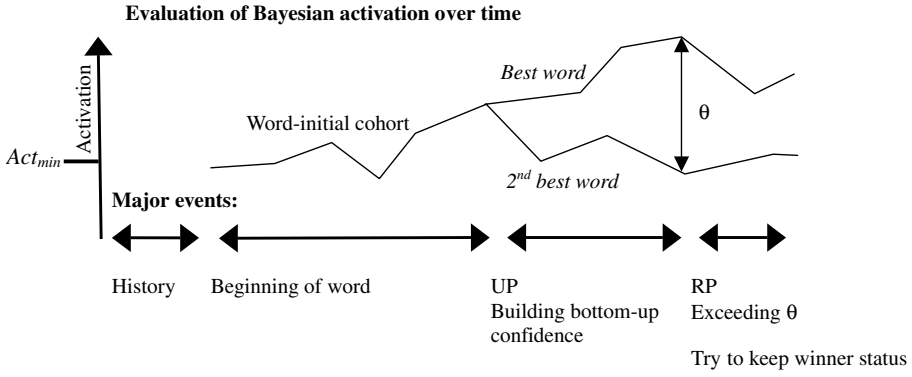

Fig. 6. Schematic illustration of the process of (on-line) early recognition.

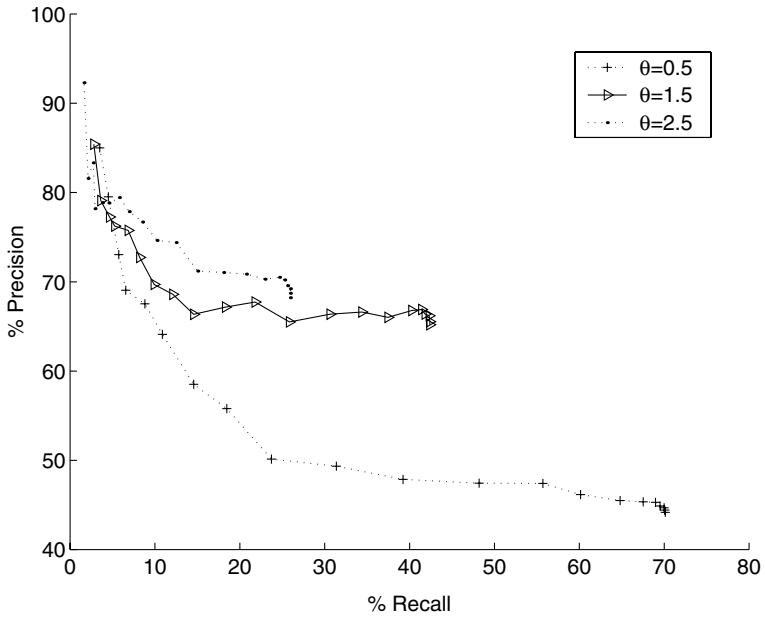

Fig. 7. For three values of θ, the precision and recall of all 21 values of Act_{min} are plotted.

6.3. The effect of the length of the word

As pointed out before, in our definition of early recognition, a word can only be recognised at or after its UP. Thus, words that have an early UP can fulfil the conditions to be recognised while there is still little evidence for the word. This raises the question whether the amount of evidence in support of a word (the number of phones between the start of the word and the DP), or the 'risk' (in the form of the number of phones following the DP until the end of the word) can be helpful in increasing precision and recall. This is the focus of the analyses described in this section. The value for Act_{min} is set to 0.5, a value that guarantees that we are on the plateau shown in Fig. 7; the value of θ was varied between 0.0 and 2.0 in 80 equal-sized steps.

We are interested in the number of words that could in principle be recognised correctly at a certain point in time. Therefore, for calculating precision and recall, only the number of focus word tokens that in principle could be recognised correctly should be taken into account. The column 'Cumulative' in Table 1 shows the number of focus word tokens that could in principle be recognised correctly at 'Length-UP' phones before the end of a word. For instance, at 8 phones before the end of the word, the only words that could in principle be recognised correctly are those that have a distance of 8 or more phones between the end of the word and the UP. At 0 phones before the end of a word, all words could in principle be identified correctly. For calculating recall, the total number of correctly recognised focus words is divided by the total number of focus words that could in principle have been recognised correctly. Precision is calculated in the same manner: The total number of correctly recognised focus words *so far* is divided by the total number of recognised focus words *so far*. The effect of the amount of evidence is investigated in a similar fashion. Precision and recall are computed as a function of the number of phones between the start of the word and the DP, and again, only the number of focus word tokens that in principle could be recognised correctly is taken into account.

The contour plots in Fig. 8 show the relation between the number of phones separating the DP from the end of the word and precision and recall for the different values of θ. On the y-axis, the value of θ is shown; the x-axis shows the number of phones between the DP and the end of the word. The lines in the plots are the equal-percentage lines for the cumulative precision (upper panel) and the cumulative recall (lower panel). Precision and recall of a point between two equal-percentage lines can be estimated using the distance of the point to the two neighbouring equal-percentage lines. For instance, for $\theta = 1.0$ and a distance of four phones between the DP and the end of the word, precision is about 39%. Fig. 8 suggests that precision and recall

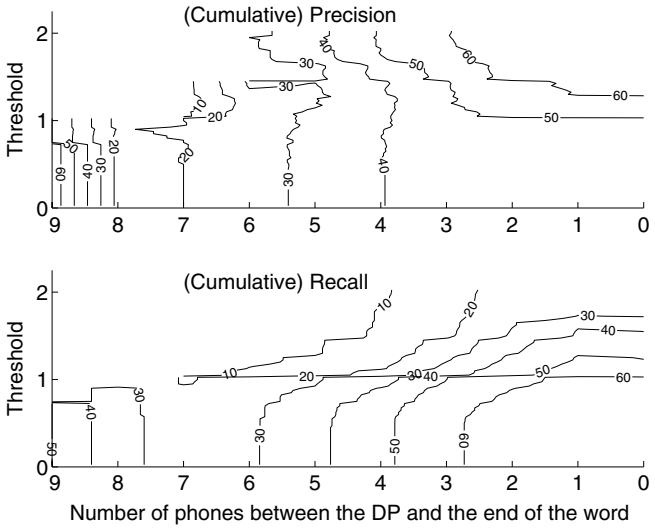

Fig. 8. The x-axis shows the number of phones between the DP and the end of the word; the y-axis shows the value of θ. The upper panel shows the precision; the lower shows the recall.

at DPs where there is a high number of phones separating the DP from the end of the word can be rather high (see the bottom left part of Fig. 8). However, this is an artefact caused by the special characteristics of the 15 focus words that happen to be unique already 8 phones before the end of the word. Precision and recall of distances between 8 and 5 are rather low. However, distances of 4 phones or less show a clear increase in both precision and recall.

The contour plots in Fig. 9 show the relation between the number of phones between the start of the word and its DP and precision and recall for different values of θ. In other words, Fig. 9 shows the effect of the amount of information available for a word on precision and recall. The results shown in Fig. 9 reveal – not surprisingly – that when there is yet little evidence available for the word, recall is rather low. The more phones have been processed, the higher recall is. The high precision for the situation where only two phones have been processed and high values of θ is an artefact of the data (see top left part of Fig. 9) – there are only a few words that exceed the threshold θ.

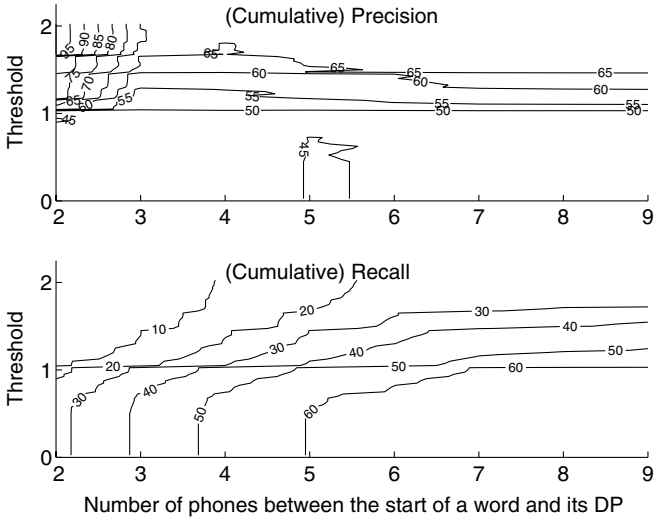

Fig. 9. The x-axis shows the number of phones between the start of the word and its DP; the y-axis shows the value of θ. The upper panel shows the precision; the lower shows the recall.

To clarify the effects of an increasing number of phones between the DP and the end of the word and an increasing number of phones between the start of the word and its DP, the precision and recall are plotted for a single θ and Act_{min} value, viz. $\theta = 1.625$ and $Act_{min} = 0.5$. Fig. 10 shows on the x-axis the number of phones between the DP and the end of the word; the y-axis shows the percentage recall (solid line) and precision (solid line with crosses +), respectively, while Fig. 11 shows on the x-axis the number of phones between the start of

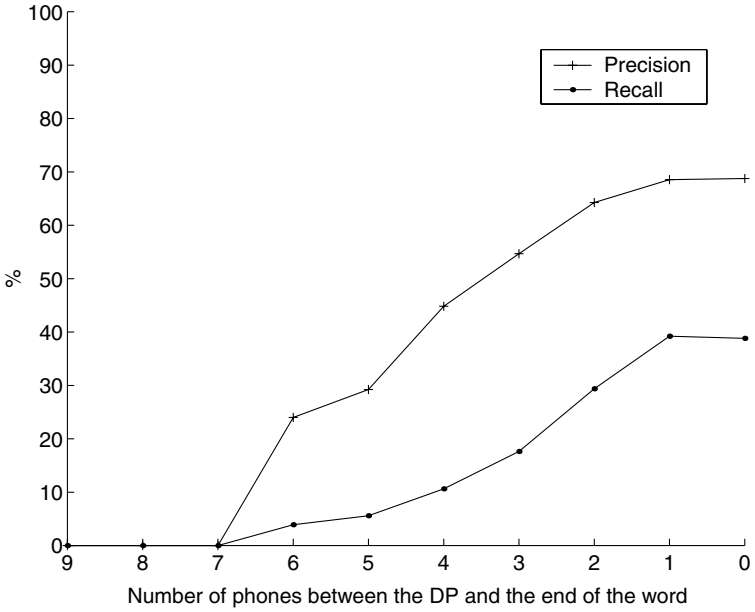

Fig. 10. The x-axis shows the number of phones between the DP and the end of the word; the y-axis shows for a $\theta = 1.625$ and $Act_{min} = 0.5$, the percentage recall (solid line) and precision (solid line with crosses +), respectively.

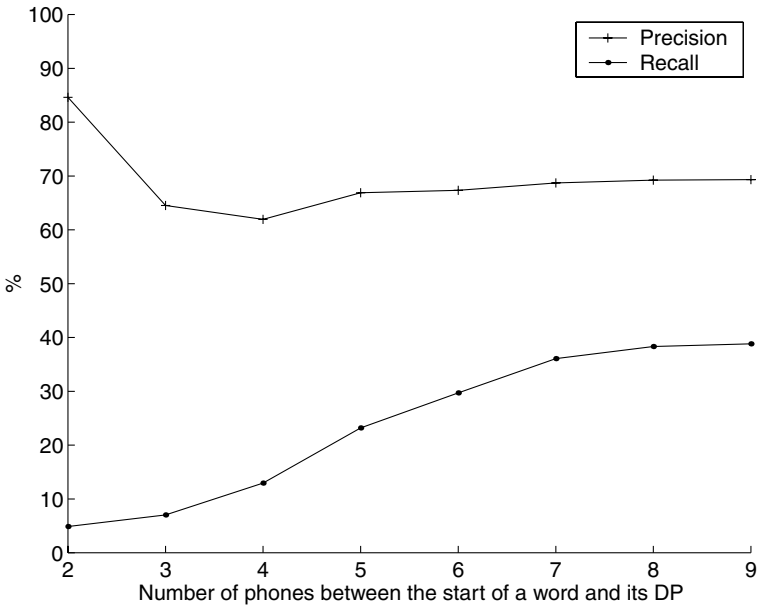

Fig. 11. The x-axis shows the number of phones between the start of the word and its DP; the y-axis shows for a $\theta = 1.625$ and $Act_{min} = 0.5$, the percentage recall (solid line) and precision (solid line with crosses +), respectively.

the word and its DP; the *y*-axis again shows the percentage recall (solid line) and precision (solid line with crosses +), respectively.

Figs. 10 and 11 clearly show what was already (implicitly) shown in Figs. 8 and 9, respectively. Precision and recall increase if the number of phones remaining after the DP is smaller (Figs. 8 and 10). This is easy to explain, since mismatches in the part of the word that is as yet unseen cannot be accounted for in the activation measure, but the risk that future mismatches occur will be higher if more phones remain until the end of the word. At the same time, performance – in terms of recall – increases if the DP is later, so that more information in support of the hypothesis is available (Figs. 9 and 11). This too makes sense, since one may expect that a high activation measure that is based on more phones is statistically more robust than a similarly high value based on a small number of phones. What should be noted, however, is that Figs. 9 and 11 further suggest that precision is not dependent on the number of phones between the start of a word and its DP: The trade-off between the false acceptances and the correctly recognised focus words does not change much.

6.4. Summary

We investigated two types of predictors for deciding whether a word is considered as recognised before the end of its acoustic realisation. The first type of predictor is related to the absolute and relative values of the word activation, Act_{min} and θ, respectively. The results showed that both predictors function as filters: The higher the values for both predictors, the fewer words are recognised, and vice versa. In this paper, we only presented the results in a form equivalent to ROC curves. Selecting the best possible combination of the two predictors is straightforward once the costs of false accepts and false rejects can be determined.

The second type of predictor is related to the number of phones of the word that have already been processed and the number of phones that remain until the end of the word. Not surprisingly, the results showed that SpeM's performance increases if the amount of evidence in support of a word increases and the risk of future mismatches decreases. These results clearly indicate that early recognition is indeed dependent on the structure and the contents of the lexicon. If a lexicon contains many (long) words that have an early UP, decisions can be made while only little information is known, at the cost of increasing the risk of errors. It is an obvious issue for follow-up research to investigate whether the decision thresholds for θ and Act_{min} can be made dependent on the phonemic structure of the words on which decisions for early recognition must be made.

Summarising, we observed that a word activation score that is high and based on more phones with fewer phones to go predicts the correctness of a word more reliably than a similarly high value based on a small number of phones or a lower word activation score.

7. General discussion and conclusion

Human listeners are able to reliably identify polysyllabic content words before the end of the acoustic realisation (e.g., Marslen-Wilson, 1987). Human listeners not only use acoustic-phonetic information, but also contextual constraints to make a decision about the identity of a word. This makes it possible for human listeners to recognise content words even before their uniqueness point. In the research presented in this paper, we investigated an alternative ASR system that is able to recognise words *during* the speech recognition process, called SpeM, for its ability for recognising words before their acoustic offset, a capability that we dubbed 'early recognition'. We define early recognition as the reliable identification of spoken words *before* the end of its acoustic realisation, but *after* the uniqueness point of the word (given the lexicon). The restriction to recognition at or after the uniqueness point allowed us to focus on acoustic recognition only, and minimise the impact of contextual constraints. One might wonder whether an advanced statistical language model would be able to emulate the context effects that enable humans to recognise words even before their uniqueness point. This would make SpeM's recognition behaviour more like human speech recognition behaviour.

In our analyses, we investigated the Bayesian word activation and the contents and structure of the lexicon as predictors for early recognition. The results in Section 6 indicate that the Bayesian word activation can be used as a predictor for the on-line early recognition of polysyllabic words if we require that the quotient of the activations of the two hypotheses with the highest scores (θ) and the minimum activation (Act_{min}) both exceed a certain threshold. There is, however, a fairly high percentage of false alarms. In the subsequent analysis, we

found an effect of the amount of evidence on the performance. If the DP was later in the word, thus with increasing acoustic evidence in support of a word, the performance in terms of precision and recall improved. Furthermore, the risk of future mismatches decreases with fewer phones between the end of the word and the DP, which also improves the performance. The predictors we have chosen have their parallels in the research area that investigates word confidence scores. For instance, the predictor θ is identical to the measure proposed in Brakensiek et al. (2003) for scoring a word's confidence in the context of an address reading system, while θ and Act_{min} are reminiscent of the graph-based confidence measure introduced in Wessel et al. (2001). The definition of word activation in SpeM resembles the calculation of word confidence measures (e.g., Bouwman et al., 2000; Wessel et al., 2001) in that both word activation and word confidence require a mapping from the non-normalised acoustic and language-model scores in the search lattice to normalised likelihoods or posterior probabilities. Conceptually, both word activation and word confidence scores are measures related to the 'probability' of observing a word given a certain stretch of speech (by the human and automatic speech recogniser, respectively). However, most conventional procedures for computing confidence measures only provide the scores after the end of an utterance.

The incremental search, used by SpeM to recognise a word before its acoustic offset, in combination with 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. An incremental search combined with word activations will be able to spot potential problems such as restarts, hesitations, and repetitions. This will be beneficial for speech-centric interaction applications.

In conclusion, we showed that SpeM, consisting of an APR and a lexical search module, is able to recognise words before the end of the word is available. In other words, the results presented in this paper showed that early recognition in an ASR system is feasible. This property of SpeM is based on the availability of a flexible decoding during the word search and on the availability of various scores along the search paths during the expansion of the search space. The early recognition process is comparable to the early selection procedure human listeners perform while decoding everyday speech. However, there is still ample room for improvement. First, the performance of SpeM as a standard ASR system is mediocre. SpeM obtained a WER of 40.4% on a set of 1106 utterances with lengths between two and five words, while the WER for the 1463 focus words was 36%. We can think of several ways for improving this performance. It has been shown that optimising the performance of the APR helps to improve the performance of SpeM as an ASR system. The same holds for the addition of an N-gram language model to the lexical search module. The search can also be improved by making the insertion, deletion, and substitution penalties dependent on the identity of the phones. For example, substitutions between the phones /t/ and /d/, which differ only in one phonetic feature, could be made smaller than the substitution of /t/ for /ʃ/, where the number of different features is higher.

For 81.1% of those 936 correctly recognised focus words (51.9% of all focus words), the use of local word activation allowed us to identify the word before its last phone was available, and 64.1% of those words were already recognised one phone after the uniqueness point. However, the straightforward predictors that we derived from the Bayesian word activation appeared to yield relatively many false accepts. Yet, we are confident that the predictive power of measures derived from word activation can be improved, if only by making decision thresholds dependent on knowledge about the words that are being hypothesised. Also, we believe that improvements in the APR will have a positive effect on the difference in word activation between the correct words and their competitors.

Acknowledgements

The authors thank Bart Kerkhoff, Tom Evers, Bram Vonk, and Joran Kapteijns, Computer Science students of the Radboud University Nijmegen, for their new implementation of SpeM. Furthermore, the authors thank two anonymous reviewers for useful comments on an earlier version of this manuscript.

References

Bouwman, G., Boves, L., Koolwaaij, J., 2000. Weighting phone confidence measures for automatic speech recognition. In: Proceedings of the COST249 Workshop on Voice Operated Telecom Services, Ghent, Belgium, pp. 59–62.

Brakensiek, A., Rottland, J., Rigoll, G., 2003. Confidence measures for an address reading system. In: Proceedings of the IEEE International Conference on Document Analysis and Recognition (CDROM).

Gaskell, M.G., Marslen-Wilson, W.D., 1997. Integrating form and meaning: a distributed model of speech perception. Language and Cognitive Processes 12, 613–656.

Glass, J.R., 2003. A probabilistic framework for segment-based speech recognition. Computer Speech and Language 17, 137–152.

Grosjean, F., 1985. The recognition of words after their acoustic offset: evidence and implications. Perception & Psychophysics 28, 299–310.

Kessens, J.M., Wester, M., Strik, H., 1999. Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. Speech Communication 29, 193–207.

Kessens, J.M., Cucchiarini, C., Strik, H., 2003. A data-driven method for modeling pronunciation variation. Speech Communication 40, 517–534.

Luce, P.A., 1986. A computational analysis of uniqueness points in auditory word recognition. Perception & Psychophysics 39, 155–158.

Luce, P.A., Goldinger, S.D., Auer, E.T., Vitevitch, M.S., 2000. Phonetic priming, neighborhood activation, and PARSYN. Perception & Psychophysics 62, 615–625.

Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word recognition. Cognition 25, 71–102.

Marslen-Wilson, W.D., Tyler, L., 1980. The temporal structure of spoken language understanding. Cognition 8, 1–71.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychology 18, 1–86.

McQueen, J.M., Cutler, A., Briscoe, T., Norris, D., 1995. Models of continuous speech recognition and the contents of the vocabulary. Language and Cognitive Processes 10, 309–331.

Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234.

Norris, D., McQueen, J.M., Cutler, A., Butterfield, S., 1997. The possible-word constraint in the segmentation of continuous speech. Cognitive Psychology 34, 191–243.

Radeau, M., Morais, J., Mousty, P., Bertelson, P., 2000. The effect of speaking rate on the role of the uniqueness point in spoken word recognition. Journal of Memory and Language 42 (3), 406–422.

Scharenborg, O., McQueen, J.M., ten Bosch, L., Norris, D., 2003a. Modelling human speech recognition using automatic speech recognition paradigms in SpeM. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 2097–2100.

Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., 2005. How should a speech recognizer work? Cognitive Science: A Multidisciplinary Journal 29 (6), 867–918.

Scharenborg, O., ten Bosch, L., Boves, L., 2003b. Recognising 'real-life' speech with SpeM: a speech-based computational model of human speech recognition. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 2285–2288.

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., 1993. The Philips research system for large-vocabulary continuous speech recognition. In: Proceedings of Eurospeech, Berlin, Germany, pp. 2125–2128.

Stolcke, A., Shriberg, E., Tür, D., Tür, G., 1999. Modeling the prosody of hidden events for improved word recognition. In: Proceedings of Eurospeech, Budapest, Hungary, pp. 311–314.

Strik, H., Russel, A.J.M., van den Heuvel, H., Cucchiarini, C., Boves, L., 1997. A spoken dialog system for the Dutch public transport information service. International Journal of Speech Technology 2 (2), 119–129.

Wessel, F., Schlueter, R., Macherey, K., Ney, H., 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9 (3), 288–298.

Wester, M., 2003. Pronunciation modeling for ASR – knowledge-based and data-derived methods. Computer Speech and Language 17 (1), 69–85.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK book (for HTK version 3.2). Technical Report, Cambridge University, Engineering Department.