# ‘On-line Early Recognition’ of Polysyllabic Words in Continuous Speech

**Odette Scharenborg, Lou Boves, and Louis ten Bosch**

Radboud University Nijmegen,
The Netherlands
{O.Scharenborg, L.Boves, L.tenBosch}@let.ru.nl

## Abstract

In this paper, we investigate the ability of SpeM, our recognition system based on the combination of an automatic phone recogniser and a word-search module, to determine as early as possible during the word recognition process whether a word is likely to be recognised correctly (this we refer to as ‘on-line’ early word recognition). We present two measures that can be used to predict whether a word is correctly recognised: the Bayesian word activation and the amount of available (acoustic) information for a word. SpeM was tested on 1,463 polysyllabic words in 885 continuous speech utterances. The investigated predictors indicated that a word activation that is 1) high (but not too high) and 2) based on more phones is more reliable to predict the correctness of a word than a similarly high value based on a small number of phones or a lower value of the word activation.

## 1. Introduction

Human listeners and automatic speech recognition (ASR) systems both are able to recognise speech, but there are clear differences between their competence levels. Human listeners outperform ASR systems on all types of recognition tasks, varying from the recognition of words spoken in isolation to the recognition of continuous speech [1]. Another difference is that human listeners are capable of recognising a word well before its acoustic realisation is complete [2,3], while ASR systems yield final results only after the end of a complete utterance (e.g., [4]).

According to mainstream psycholinguistic theory on human speech perception, humans seem to compute an on-line activation measure for words as the speech comes in (and presumably make a decision as soon as the activation of a word is high enough). Conventional ASR systems, on the other hand, compute the likelihood of a number of hypothesised word sequences, and identify the words that were recognised on the basis of the hypothesis with the highest score at the end of the utterance.

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 simplify 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. [5]. Thus, it is worthwhile to investigate how ASR systems should be adapted to be able to perform early recognition: i.e., recognising a word before the end of its acoustic realisation is complete.

In [6] and [7], we have presented an end-to-end speech recognition system called SpeM (SPEech based Model of human speech recognition) that is, in principle, capable of providing ‘word activations’ that are derived from the log-likelihood values such as used in conventional ASR systems. (Since the procedure that converts log-likelihoods into activations is based on Bayes rule, we use the term ‘Bayesian activation’ along with the more general term ‘word activation’.)

In [6], we investigated the performance of SpeM as a speech recognition system that makes decisions about the words (mainly spoken in isolation) it has recognised after the complete signal has been processed. In other words, SpeM was used as a standard ASR system. In [8], we extended this research to investigate SpeM’s capability for doing continuous speech recognition, and we showed that the Bayesian activation could be used to recognise a word before the end of its acoustic realisation. However, we should note that no perfect recognition result was obtained.

Even in SpeM, early recognition is a different task than ‘normal’ speech recognition. In the latter task, it suffices to search for the best scoring path through the search space spanned by the language model and the acoustic input. In the case of early recognition as described in [8], this best-scoring path was analysed to investigate the position in a word at which the activation

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. 

Accepted after full paper review

of that word became and remained the highest until the end of the input. In the case of *on-line* early recognition, however, we need an additional decision procedure for accepting a word as being recognised if its local activation fulfils one or more criteria.

Early recognition is a task with many aspects. The task is clearly 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 phonemes, early recognition on the basis of acoustic input is more difficult than when the lexicon mainly consists of words which contain many phonemes after the lexical uniqueness point (UP). At the same time, it is evident that making decisions on the basis of only a few phonemes at the beginning of a long word is more dangerous than making a decision on the basis of a longer string of phonemes. Therefore, we will investigate the decision criteria in early recognition as a function of the structure of the lexicon.

In the next section, we will describe our ASR system SpeM, including an explanation on how Bayesian word activations are derived from log-likelihood values, and the materials used in the experiments and analyses. Section 3 describes the general procedure used throughout the experiment, the criteria a word must comply with in order to be recognised by SpeM, and how these were implemented in the current study. In section 4, the Bayesian activation is introduced in the context of on-line early recognition, and the optimal decision criteria for on-line early recognition based on word activation as a function of the contents and structure of the lexicon are investigated. Finally, in section 5, we will discuss our results, and the most important findings are summarised.

## 2. Materials

### 2.1. The recognition system (SpeM)

SpeM was implemented to function as an automatic speech recognition system and at the same time as a tool for research in the field of human speech recognition (HSR). It is a new and extended implementation of *Shortlist*, the computational model of human word recognition developed by Norris [9].

SpeM consists of two modules that operate in sequence. The first module, the automatic phone recogniser (APR), generates a symbolic representation of the speech signal in the form of a probabilistic phone graph. 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 the accumulated acoustic evidence for that word. In the present set-up, the word search module only starts after the first module has processed a complete utterance. However, the search module is able to analyse a phone graph that grows incrementally as more speech becomes available. Below, we give the relevant details of the two modules.

#### *2.1.1. The automatic phone recogniser*

For the APR, 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 [10]. The speech style is extemporaneous. The phonemic transcriptions of the training material consisted of the citation forms of the words; thus no pronunciation variation was taken into account.

The ‘lexicon’ used for the phone recognition consists of all Dutch phones, 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 trained on the phonemic transcriptions of the training material.

The focus of the present study is on ways in which word recognition can be predicted during the subsequent search process; we did not invest effort in optimising the performance of the APR.

#### *2.1.2. The search module*

The most important functionalities of the search module of SpeM used in this experiment have been described in detail in [6]. For the current paper, a new implementation was used, which provides more parameters to control the search. The biggest improvement is that the new version supports unigram and bigram language models.

In SpeM, the search for the best-matching sequence of words is the search for the cheapest path through the product graph of an input phone lattice and a lexical tree. The search is implemented as a time-synchronous and breadth-first DP. In the lexical tree, entries share common phone sequence prefixes (word-initial cohort), and each complete path through the tree represents a pronunciation of a word. SpeM is programmed to output the list of the N-best paths through the product graph for each node in that product graph, together with the activation scores of the paths and the words on these paths. This capability enables SpeM to provide the activation values that can be used in the decision module performing on-line early recognition.

SpeM has a number of parameters that can be tuned individually and in combination. Most of these parameters (e.g., a word entrance penalty and the trade-off between the bottom-up acoustic cost of the phones calculated by the APR and the contribution of the language model) are similar to the parameters in conventional ASR systems. In addition, however, SpeM has two parameters that are not usually applied in



Accepted after full paper review

conventional ASR systems. The first novel parameter is associated to the cost for a symbolic mismatch between the input lattice and the lexical tree due to phone insertions, deletions, and substitutions (comparable to the Levenshtein distance). Insertions, deletions, and substitutions have their own weight that can be tuned individually. The second novel parameter is associated to the *Possible Word Constraint* (PWC; [11]). The PWC checks 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 [6]). Phone sequences that do not conform to the PWC are penalised, while sequences that do not violate the constraint are not.

#### *2.1.3. Word activation*

The measure of *word activation* in SpeM was originally designed to simulate experimental results of human word recognition experiments [7]. 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:

- The word that matches the input best 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 [9]). 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 all ASR systems. For more details on the implementation of the Bayesian word activation, the reader is referred to [8].

### 2.2. Data

In our evaluation of SpeM's ability for early recognition, we focus on polysyllabic words. Given the characteristics of the training corpus and test corpus, which consists of utterances taken from dialogs between customers and an automatic timetable information system, we decided to define a set of 318 polysyllabic station names as *target* words. From the VIOS database, 1,106 utterances (disjoint from the training corpus) were selected to tune and test SpeM. Each utterance contained two to five words, at least one of which was a target word (708 utterances contained multiple target words).

885 utterances of this set (80% of the 1,106 utterances) were randomly selected and used as the independent test corpus. The total number of target words in the test corpus was 1,463; 563 utterances contained multiple target words. The remaining 221 utterances were used as development test set and served as a tuning corpus on which the parameters of SpeM were tuned. The parameter settings yielding the lowest Word Error Rate (WER) were used for the experiment. The WER is defined as the number of (word) insertions, deletions, and substitutions divided by the total number of words in the reference transcriptions times 100%.

The lexicon used by SpeM in the test consisted of 980 entries: the 318 polysyllabic station names, additional city names, verbs, numbers, function words, etc. 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 uni- and bigram phonotactic models for the APR.

## 3. Experimental design

### 3.1. General procedure and evaluation

For each utterance, the APR module created a phone lattice, which was subsequently presented to the search module of SpeM. For each input node, a list of the 10 most likely sequences of words was created (10-best list). At the top of this list is the sequence of words that best matches the acoustic signal. For each target word, the *recognition point* (RP) was determined; this is the node at which the target word is recognised by SpeM. The next section explains the decision module that we implemented that decides when a word is said to be recognised.

In order to use word activation as a basis for deciding whether a word is considered as recognised before the end of its acoustic realisation, a decision procedure is needed that takes absolute and relative values of the Bayesian activation into account, perhaps conditioned on the number of phonemes of the word that have already been processed and the number of phonemes that remain until the end of the word. The performance of the decision module will be evaluated in terms of *precision* and *recall*:

*Precision:* The total number of correctly recognised target words relative to the total number of recognised target words. Precision gives an impression of the trade-off between correctly recognised target words and false accepts.

*Recall:* The total number of correctly recognised target words divided by the total number of target words



Accepted after full paper review

in the input. Recall gives an impression of the trade-off between correctly recognised target words and false rejects.

Since we are not primarily interested in optimising SpeM for a specific task in which the relative costs of false accepts and false rejects can be established, we decided to refrain from defining a total cost function that combines recall and precision into a single measure that can be optimised.

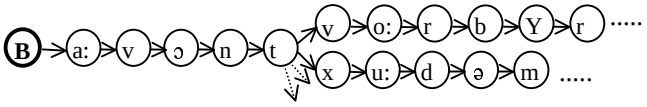


*Figure 1*. Two target words in competition.

### 3.2. Criteria for word recognition

For a target word to be recognised by SpeM, the following three conditions have to be met:

1) The recognised phone sequence assigned to the target word is at or beyond the target word's UP.
2) The quotient of the word activation of the target word on the first-best path and the word activation of its closest *competitor* (if present) exceeds a certain *threshold* (θ). Thus, we do not want SpeM to make a decision as long as promising competitors are still alive. 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. Figure 1 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 the last phone. It is not guaranteed that all words always have a competitor. It is possible that all paths in the N-best list are completely disjunct. According to our definition no word can have a competitor in this case. In the experiments described below, we have varied θ between 0 and 4 in 100 equal-sized steps. Absence of a competitor makes the computation of θ impossible. To prevent losing all words without competitors due to a missing value, we define target words without a competitor as being above the quotient by assigning the arbitrary value of 5 to these words.
3) The value of the Bayesian activation of the target word should exceed a certain *minimum activation* ($Act_{min}$). Thus, SpeM does not just accept the word with the highest activation, irrespective of the absolute value of the activation. In the experiments described below, the value of $Act_{min}$ was varied between 0.0 and 10.0 in 20 equal-sized steps.

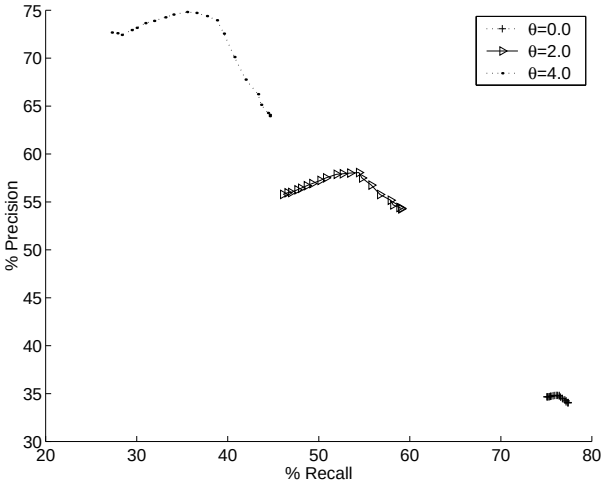


*Figure 2*. For three values of θ, the precision and recall of all values of $Act_{min}$ are plotted.

## 4. Analyses and results

### 4.1. The effect of the Bayesian activation

The effect of the Bayesian activation on the recognition performance is investigated in terms of θ and $Act_{min}$. Figure 2 shows the relation between precision (y-axis) and recall (x-axis) for a number of combinations of the two thresholds. The symbols on the lines in Figure 2 represent the values of $Act_{min}$ for three different values of θ. The left-most symbol on each line corresponds to $Act_{min}$=10.0; the right-most one corresponds to $Act_{min}$=0.0. For the sake of clarity, Figure 2 is limited to three values of θ; all other values of θ show the same trend, i.e., a maximal precision for $Act_{min}$ values in the middle of the range between 0.0 and 10.0. Moreover, lower values of θ always yield a higher recall.

For the most part the results in Figure 2 are according to expectation. Recall should be an inverse function of θ: the smaller θ becomes, the less will it function as a filter for words that have a sufficiently high activation, but which still have viable competitors. For higher values of $Act_{min}$, fewer target words will have an activation that exceeds $Act_{min}$, and thus fewer words are recognised. However, it is puzzling why precision should decrease if $Act_{min}$ becomes bigger than about 5.0. Apparently, it is possible for words to have very high local activation values, despite the fact that they are not present in the speech signal. At the same time, these high values seem to be non-trustworthy.

### 4.2. The effect of the structure of the lexicon

As pointed out before, 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



Accepted after full paper review

(the number of phones between the start of the word – or alternatively the UP – and the RP in a word), or the 'risk' (in the form of the number of phones following the RP until the end of the word) can be helpful in increasing the precision and the recall. This is the focus of the analyses described in this section. The value for $Act_{min}$ is set to 0.5.

| **L-UP** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|
| **#types** | 10 | 44 | 50 | 63 | 50 | 39 | 38 | 17 | 3 | 2 |
| **#tokens** | 30 | 190 | 182 | 450 | 271 | 186 | 82 | 57 | 11 | 4 |
| **#Cum.** | 1463 | 1433 | 1243 | 1061 | 611 | 340 | 154 | 72 | 15 | 4 |

*Table 1*. The distribution (in #types and #tokens) in number of phones between the UP and the length of a word (L(ength)-UP); #Cum.: #target word tokens that could in principle be recognised at position L(ength)-UP.

Since a word can only be recognised at or after its UP, we first would like to know the distribution in number of phones between the UP and the length of a word. This is shown in Table 1. The first row shows the distance in number of phones between the UP and the end of the word. A 'L(ength)-UP' of 0 means that the UP lies at the end of the word: the word is embedded in a longer word. Rows 2 and 3 show the number of target word types and tokens, respectively. The row '#Cum(ulative)' shows the number of target word tokens that could in principle be recognised correctly at 'L(ength)-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 length of the word and the UP. At 0 phones before the end of a word, all words could in principle be recognised.

For calculating the recall, the total number of correctly recognised target words is divided by the total number of target words that could in principle have been recognised correctly. In the same manner, the precision is calculated: The total number of correctly recognised target words *so far* are divided by the total number of recognised target words *so far*.

The effect of the amount of evidence is investigated in a similar fashion. The precision and recall are computed as a function of the number of phones between the start of the word and the RP, and again, only the number of target word tokens that in principle could be recognised correctly is taken into account.

The contour plots in Figure 3 show the relation between the number of phones between the RP and the end of the word and the precision and recall for different values of θ. The cumulative precision is plotted in the upper panel; the cumulative recall is plotted in the lower panel. On the y-axis, the value of θ is shown; the x-axis shows the number of phones between the RP and the end of the word. The lines in the plots are the equal-percentage lines for the precision (upper panel) and the recall (lower panel). The 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 θ=1.0 and a distance of two phones between the RP and the end of the word, the precision is about 38%. Similarly, the contour plots in Figure 4 show the relation between the number of phones between the start of the word and its UP and the precision and recall for different values of θ. In other words, Figure 4 shows the effect of the amount of information available for a word on the precision and recall.

Figure 3 suggests that precision and recall at RPs where little evidence is yet available can be rather high. However, this is an artifact caused by the special characteristics of the 15 target words that happen to be unique already 8 phones before the end of the word. The precision and recall of distances between 7 and 5 are rather low. However, distances of 4 phones or less show a clear increase in both precision and recall. The results shown in Figure 4 reveal – not surprisingly – that when there is yet little evidence available for the word (the number of phones processed is lower than 4), the precision and recall are rather low. The more phones have been processed, the higher the precision and recall are.

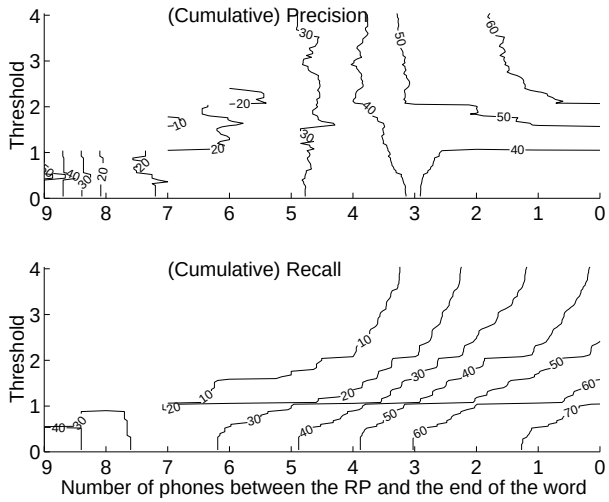


*Figure 3*. The x-axis shows the number of phones between the RP and the end of the word; the y-axis shows the value of θ. The upper panel shows the precision; the lower shows the recall.

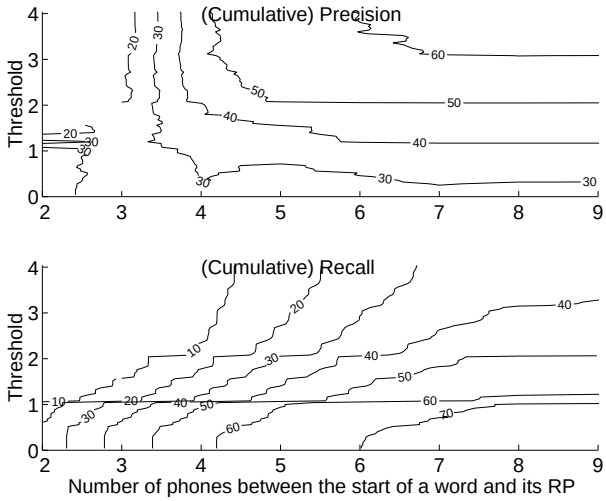


*Figure 4*. The x-axis shows the number of phones between the start of a word and its UP; the y-axis shows the value of θ. The upper panel shows the precision; the lower shows the recall.

The results show that precision and recall decrease if the number of phones remaining after the RP is larger. 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 increases if the RP is later, so that more information in support of the hypothesis is available. This too makes sense, since one may expect that a high activation measure that is based on more phones is more reliable than a similarly high value based on a small number of phones.

## 5. Discussion and Conclusion

In our analyses, we investigated the effect of the Bayesian word activation and the effect of the structure of the lexicon on the performance during on-line early recognition. The results in section 4 indicate that the Bayesian activation in terms of θ and $Act_{min}$ 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 high percentage of false alarms. In the subsequent analysis, we found an effect of the amount of evidence on the performance. In the case the RP was far before the end of the word, the word could not be reliably recognised. On the other hand, the fewer phones there are between the end of the word and the recognition point, the more reliable the recognition of a word became.

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 [12] for scoring a word's confidence in the context of an address reading system.

## 6. Acknowledgements

Part of the research reported in this paper was supported by the IST project COMIC (IST-2001-32311). Furthermore, the authors would like to thank Gies Bouwman for his help in this research.

## 7. References

[1] Lippman, R. (1997). Speech recognition by machines and humans. *Speech Communication 22* (*1*), 1-15.

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

[3] 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.

[4] Lee, C.-H., Soong, F. K., Paliwal, K. K. (1996). *Automatic Speech and Speaker Recognition*. Boston: Kluwer Academic Publishers.

[5] 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.

[6] Scharenborg, O., ten Bosch, L., Boves, L. (2003a). Recognising 'Real-life' Speech with SpeM: A Speech-based Computational Model of Human Speech Recognition. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 2085-2088.

[7] Scharenborg, O., McQueen, J. M., ten Bosch, L., Norris, D. (2003b). Modelling Human Speech Recognition using Automatic Speech Recognition Paradigms in SpeM. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 2097-2100.

[8] Scharenborg, O., ten Bosch, L., Boves, L. (2003c). 'Early Recognition' of Words in Continuous Speech. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, US Virgin Islands (CDROM).

[9] Norris, D. (1994). Shortlist: a Connectionist Model of Continuous Speech Recognition. *Cognition 52*, 189-234.

[10] 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.

[11] 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.

[12] 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).