19

Early decision making in continuous speech

Odette Scharenborg, Louis ten Bosch, Lou Boves
CLST, Department of Linguistics/Language and Speech, Radboud University Nijmegen
The Netherlands

1. Introduction

In everyday life, speech is all around us, on the radio, television, and in human-human interaction. Communication using speech is easy. Of course, in order to communicate via speech, speech recognition is essential. 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 corresponding symbolic representations of the words in an internal lexicon. Psycholinguistic experiments have shown that listeners are able to recognise (long and frequent) words reliably even before the corresponding acoustic signal is complete (Marslen-Wilson, 1987). According to theories of HSR, 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 can make a decision as soon as the activation of a word is high enough, possibly before all acoustic information of the word is available (Marslen-Wilson, 1987; Marslen-Wilson and Tyler, 1980; Radeau et al., 2000). 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" is referred to as *early selection* by Marslen-Wilson (1987).

In general terms, automatic speech recognition (ASR) systems operate in a way not unlike human speech recognition. However there are two major differences between human and automatic speech recognition. First of all, most mainstream ASR systems avoid an explicit representation of the prelexical level to prevent premature decisions that may incur irrecoverable errors. More importantly, ASR systems postpone final decisions about the identity of the recognised word (sequence) as long as possible, i.e., until additional input data can no longer affect the hypotheses. This too is done in order to avoid premature decisions, the results of which may affect the recognition of following words. In more technical terms: ASR systems use an integrated search inspired by basic Bayesian decision theory and aimed at avoiding decisions that must be revoked due to additional evidence. The competition between words in human speech recognition, on the other hand, is not necessarily always fully open; under some conditions an educated guess is made about the identity of the word being spoken, followed by a shallow verification. This means that the winning word might be chosen before the offset of the acoustic realisation of the word, thus while other viable competing paths are still available. Apparently, humans are willing to take risks that cannot be justified by Bayesian decision theory.

At a higher level, i.e., the level of human-human interaction, early decision making plays an eminent role as well. This is shown by the extremely efficient turn-taking processes, in which listeners take the turn after predicting the moment when turn-switching can take place (Garrod & Pickering, 2004), resulting in dialogues with minimal latencies between successive turns. Since current ASR systems do not recognise words before their acoustic offset, let alone that they would predict the end of an utterance, latencies in turn-taking in human-computer interaction are unavoidable, resulting in unnatural dialogues.

If one wants to build an ASR system capable of early decision making, one needs to develop an algorithm 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 speech is coming in. It is important to note that since early recognition involves making decisions before all potentially relevant information is available, it introduces the risk of making errors (i.e., false alarms of other words than were actually spoken).

This chapter introduces a novel approach to speech decoding that enables recognising polysyllabic words before their acoustic offset. The concept behind this novel approach is '*early recognition*', i.e., the reliable identification of spoken words *before* the end of their acoustic realisation, but *after* the uniqueness point (UP)[1] of the word (given a lexicon). The restriction to recognition at or after the uniqueness point allows us to focus on acoustic recognition, with the same impact of a language model as in conventional ASR systems, which would be comparable – but certainly not identical – to the contexts used in human word recognition in Marslen-Wilson's definition of 'early selection'.

Early recognition is dependent on the structure and the contents of the lexicon. If a lexicon contains many words that have a UP very late in the word (i.e., 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 have an early UP (i.e., contain long 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 for early recognition. This chapter will present experiments conducted to optimise the performance of a procedure for making decisions before all acoustic information is available and discuss the results.

2. The recognition system

For conventional speech recognition, it suffices to search for the best-scoring path in 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. In Scharenborg et al. (2003, 2005), we presented a speech recognition system called SpeM (Speech based Model of human speech recognition), based on *Shortlist* (Norris, 1994), that is

[1] In a lexicon organised in the form of a prefix-tree, the uniqueness point is the phoneme after which a path does not branch anymore.

capable of providing on-line dynamically changing 'word activations' derived from the log-likelihood values in conventional ASR systems. SpeM was originally developed to serve as a tool for research in the field of HSR.

SpeM consists of three modules. The first module, the automatic phone recogniser (APR), generates a symbolic prelexical 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, among others, the accumulated acoustic evidence for that word. During the lexical search, SpeM provides a list of the most likely path hypotheses at every phone node in the phone graph. The third –decision- module is entered each time after a node in the phone graph is processed in the second module. This enables SpeM to recognise and accept words before the end of an utterance or phrase. The focus of this paper is on the third module, which makes decisions about accepting a word as being recognised if its local word activation fulfils one or more criteria.

The most important difference between SpeM and conventional ASR systems 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 deliberately lacking. This is not to say that computing on-line dynamically varying word activation scores is fundamentally impossible in decoders that avoid an explicit prelexical representation, but doing so would require a class of algorithms that differ very much from SpeM.

2.1. Material and evaluation

In our evaluation of SpeM's ability for early decision, we focus on polysyllabic content words. The reasons for this are twofold. Firstly, function words and short content words that are not easy to predict from the (linguistic) context may not be identified by human listeners until the word following it has been heard (Grosjean, 1985). Secondly, short words are likely to have a UP that is not before the end of the word, since they are often embedded in longer words (McQueen et al., 1995), making it a priori impossible to recognise the word before its acoustic offset on the basis of only acoustic evidence.

The training and test data are taken from the VIOS database, which consists of utterances taken from telephone dialogs between customers and the Dutch public automatic transport timetable information system (Strik et al., 1997). The material to train the acoustic models of the APR (AM training material) consists of 25,104 utterances in Dutch (81,090 words, corresponding to 8.9 hours of speech excluding leading, utterance internal, and trailing silent portions of the recordings).

A set of 318 polysyllabic station names is defined as *focus* words. From the VIOS database, 1,106 utterances (disjoint from the AM training corpus) were selected. 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 1,106 utterances) were randomly selected and used as the independent test corpus. The total number of focus words in this test corpus was 1,463 (563 utterances contained multiple focus words). The remaining 221 utterances were used as development set and served to tune the parameters of SpeM (see also Section 2.3). 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:

$$WER=(\#I+\#D+\#S)/N \cdot 100\% \quad (1)$$

where $\#I$ denotes the number of word insertions; $\#D$ the number of word deletions, $\#S$ the number of word substitutions, and N denotes the total 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 in the test. For each word in the lexicon, one unique canonical phonemic representation was available. A unigram language model (LM; see also Section 2.3) was trained on the AM training data. This implies that the SpeM decoder only knew about the relative frequency of the 980 lexical entries (words), but that it had no means for predicting words from the preceding linguistic context – which is good for making the word competition as fair as possible.

2.2. The automatic phone recogniser

The APR used in this study was 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). 37 context-independent phone models, one noise model, and one silence model were trained on the VIOS training set. All phone models and the noise model have a linear left-to-right topology with three pairs of two identical states, one of which can be skipped. For the silence model, a single-state hidden Markov Model 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 1,415 entries (a superset of the 980 words in the recognition lexicon), including entries for background noise and filled pauses. For each word, the lexicon contained only the unique canonical (citation) pronunciation. Thus, potential pronunciation variation in the training corpus was ignored while training phone models.

The 'lexicon' used for the phone decoding 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 decoding, the APR uses a bigram phonotactic model trained on the canonical phonemic transcriptions of the AM training material.

The APR converts the acoustic signal into a probabilistic phone lattice without using lexical knowledge. The lattice has one root node and one end node. 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). This acoustic cost is directly related to the probability $P(X|Ph)$ that the acoustic signal X was produced given the phone Ph.

2.3. 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 the pronunciation of a word. The lexical tree has one root node and as many end nodes as there are pronunciations in the lexicon. Nodes that are flagged as end nodes but also have outgoing edges indicate embedded words.

Like a conventional ASR system, SpeM searches for the best-scoring or 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 cost factors computed in the search module.

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 (see Section 3) for each of the hypothesised words 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 in terms of phone symbols, 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 to 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.

The current implementation of SpeM supports the use of unigram and bigram LMs, which model the prior probability of observing individual words and of a word given its predecessor. In the experiments reported in this paper, only a unigram LM is used. SpeM has a number of parameters that affect the total cost 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 LM, 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 symbols in the input 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, on the other hand, the search space is usually spanned by a 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., 2003, 2005). The PWC evaluation is applied only to paths that do not consist solely of complete words. Word onsets and offsets, utterance onsets and offsets, and pauses count as locations relative to which the viability of symbol sequences that are no words (i.e., lexical items) are evaluated. If there is no vowel in the sequence between any of these locations and a word edge, 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]. Because *fourf* is not stored as a word in the lexicon, a potential parse by the recogniser is *they metaphor f time*. Since *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 PWC enables SpeM to parse input with broken words and disfluencies, since it provides a mechanism for handling arbitrary phone input (see 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 the independent development set (see Section 2.1), and subsequently used for processing the test corpus.

3. The computation of word activation

An essential element for early decision making is the computation of word activation. The measure of word activation in SpeM was originally designed to simulate the way in which word activations evolve over time in experiments on human word recognition (Scharenborg et al., 2005, 2007). In the computation of the word activation, the local negative log-likelihood scores for complete paths and individual words on a path are converted into word activation scores that obey the following properties, which follow from the concept of word activation as it is used in HSR:

- The word that matches the input best, i.e., the word with the smallest *word cost* (see Section 2.3), must have the highest activation.
- The activation of a word that matches the input must increase each time an additional matching input phone is processed.
- The measure must be appropriately normalised: 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 and some kind of (probabilistic) LM, which is precisely the cost function that is maximised in conventional ASR systems. Thus, it is reasonable to stipulate that the word activation $Act(W/X)$ is equal to $P(W|X)$, 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 $Act(W|X)$ (or $P(W|X)$) 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 too is reminiscent of what happens in conventional ASR systems.

Since we also want to deal with incompletely processed acoustic input (for early recognition of words), Bayes' Rule is applied to *Act*(*W* | *X*) in which both *W* and *X* are evolving over time *t*, and *t*-steps coincide with phone boundaries:

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

where *W*(*n*) denotes a phone sequence of length *n*, corresponding to the word-initial cohort of *n* phones of *W*. So, *W*(5) may, for example, be /ɑmstə/, i.e., the prefix (or word-initial cohort) of the word 'amsterdam' (but also of other words that begin with the same prefix). *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*).

As said before, in the experiments reported in this chapter, *P*(*W*(*n*)) is exclusively based on the unigram probability of the words and the word-initial cohorts (the unigram probabilities for word-initial cohorts are determined by summing over the unigram probabilities of all words in the cohort). The (unnormalised) conditional probability *P*(*X*(*t*) | *W*(*n*)) in equation 2, is calculated by SpeM as:

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

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*)) and *P*(*W*(*n*)). Thus, *a* is similar to the 'language model factor' in standard ASR systems. *a* is a positive number; it's numerical value is determined such that the three properties of word activation introduced at the start of this section hold (for a more detailed explanation of *a*, see Scharenborg et al., 2007).

In contrast to conventional ASR systems, in SpeM, the prior *P*(*X*(*t*)) in the denominator of equation 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). Furthermore, the normalisation needed in SpeM 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. In normalising equation 2, we followed the procedure for normalising confidence measures. However, instead of the number of frames, the number of phones is the normalising factor, resulting in a type of normalisation that is more phonetically oriented. The denominator of equation 2, 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 $Act(W(n) \mid X(t))$. 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 (for a more detailed explanation of D, see Scharenborg et al., 2007).

Our choice to normalise the Bayesian activation by the expression given by equation 4 is also based on another consideration. Given the Bayesian paradigm, it seems attractive to use a measure with the property that logarithmic scores are additive along paths. If X_1 and X_2 are 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'). By doing so the lengths of X_1 and X_2 are assumed to be independent, which is a plausible assumption.

4. Early recognition in SpeM

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

In order to assess SpeM's ability for early decision, it is essential to know the upper-bound of its performance: First of all, if a word is not correctly recognised, it will be impossible to analyse its *recognition point* (RP); secondly, only if the RP of a word lies before the end of that word, it can, in principle, be recognised before its acoustic offset during the recognition process. In this paper, 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). The RP necessarily lies after the UP (since 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), and is expressed as the position of the corresponding phone in the phonemic (lexical) representation of the word.

In a first step, the performance of SpeM as a standard ASR system was investigated. The WER on the full test set and on the focus words was calculated 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. The WER obtained by SpeM on all words in the test material was 40.4%. Of the 1,463 focus words, 64.0% (936 focus words) were recognised correctly at the end of the word. Despite the mediocre performance of SpeM as an ASR system, we believe it is still warranted to investigate SpeM on the task of early decision, since there is a sufficiently large number of correctly recognised focus words. It should be noted that in this study no attempt has been made to maximise the performance of the acoustic model set of the APR. However, the results presented in Scharenborg et al. (2003, 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. It is thus quite probable that improving the performance of the APR should allow SpeM to reach an ASR performance level comparable to a conventional ASR system on the VIOS data set (see Scharenborg et al., 2007).

Length-UP	#Types	#Tokens	Cumulative
0	10	30	1,463
1	44	190	1,433
2	50	182	1,243
3	63	450	1,061
4	50	271	611
5	39	186	340
6	38	82	154
7	17	57	72
8	3	11	15
9	2	4	4

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.

4.2. Analysis of the recognition point

In a second step, to determine SpeM's upper-bound performance on the task of early decision, we investigate how many of the focus words have an RP that lies before the end of the word. Table 1 shows the distribution of the distance in number of phones between the UP and the end of a focus word. 'Length-UP' = 0 means that the UP is at the end of the word: Either the word is embedded in a longer word or the words only differ in their last phoneme. Columns 2 and 3 show the number of focus word types and tokens with 'Length-UP' phones between the end of the word and the UP. The column 'Cumulative' 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 a word, the only words that can 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. From Table 1 it can be deduced that the UP of 85.0% of all focus word tokens (1,243/1,463) is at least two phones before the end of the word; only 2% of the focus word tokens (30/1,463) have their UP at the end of the word.

In our analysis of the RP, we only took those focus words into account that were recognised correctly, since, obviously, a word that is not recognised correctly does not have an RP. First, the path and word hypotheses were ranked using the Bayesian word activation score. Subsequently, for each correctly recognised focus word, the node after which the Bayesian word activation exceeds the Bayesian word activation of all its competitors, and remains the highest until the end of the word is determined. 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).

To understand how much evidence SpeM needs to make an early decision about 'recognising' a word, the RP of all 936 correctly recognised focus words was related to the UP and the total number of phones of that word. The results are shown in the form of two histograms in Fig. 1. 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. The high frequency in the case of $N = 3$ in the right panel of Fig. 1 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. 1 and Table 1 shows 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. 1) 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, 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. 1 it can also be deduced that 64.1% (sum of $N = 0$ and $N = 1$, see left panel of Fig. 1) 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.

As pointed out at the start of this chapter, psycholinguistic research (Marslen-Wilson, 1987) has shown that listeners are able to recognise long and frequent words before their acoustic offset. However, this does not imply that this always happens and for all words. There are still words (including frequent and long words) that can only be recognised by a listener after some of the following context has been heard. The SpeM results showed that the UP and RP do not coincide for all focus words that were recognised. SpeM, like listeners, occasionally needs information from the following context to make a decision about the identity of a word. 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. An analysis (see Scharenborg et al., 2007) showed that 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.

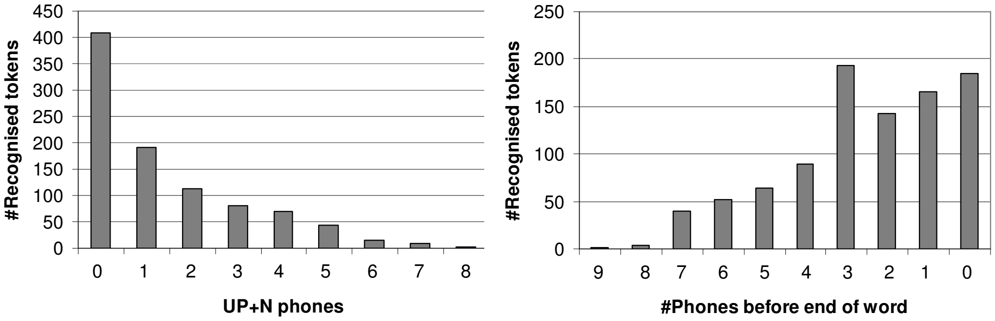

Fig. 1. *Left panel:* histogram relating the RP to the UP. *Right panel:* histogram relating the RP to the total number of phones in the word for the 936 correctly recognised focus words.

5. Predictors for reliable on-line early decision making

In the previous analyses, the RP and thus SpeM's ability of early decision was investigated *after* the recognition process had taken place. This was done to determine the upper-bound of SpeM's performance on the task of early decision: 936 focus words were recognised correctly, while 2% of the focus words have their UP at the end of the word. It can be deduced from the right panel of Fig. 1 that, as an upper-bound, 751 (80.2%) correctly recognised focus words can be recognised before the acoustic offset of the word. Of course, in order to use the concept of early decision in an operational system in order to speed up for instance human-computer interaction, a procedure needs to be developed that accurately and reliably decides whether a word is considered as recognised before the end of its acoustic realisation.

As stated above, in comparison to integrated search approaches as used in mainstream ASR systems, early decision making introduces an additional decision problem that introduces additional errors and thus additional risks. Intermediate results are also computed in integrated search and therefore might be made available at the output interface of the search module, but because these results can still change later on in the recognition process, this is not usually done. One exception to this rule are dictation systems that show word hypotheses on the screen that are subsequently revised as the search progressed. In the case of early decision making as defined in this study, however, a decision made during the recognition process cannot be adapted, and is thus final. Early decision making is thus not synonymous with *fast* decision making; early decision making *predicts* the future.

The analyses presented in the previous section showed that the Bayesian word activation of many polysyllabic content words exceeds the activation of all competitors before the end of the words. However, this does not imply that the Bayesian word activation can be safely used to perform early decision. We created a decision procedure on the basis of the Bayesian word activation and experimented with a combination of absolute and relative values of the Bayesian word activation. Additionally, we investigated whether the reliability of early decision making 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. 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. Thus, *precision* measures the trade-off between correctly recognised focus words and *false alarms*.
- *Recall*: The total number of correctly recognised focus words divided by the total number of focus words in the input. Thus, *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 interested in optimising SpeM for a specific task in which the relative costs of false alarms and false rejections can be established, since in this paper 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.

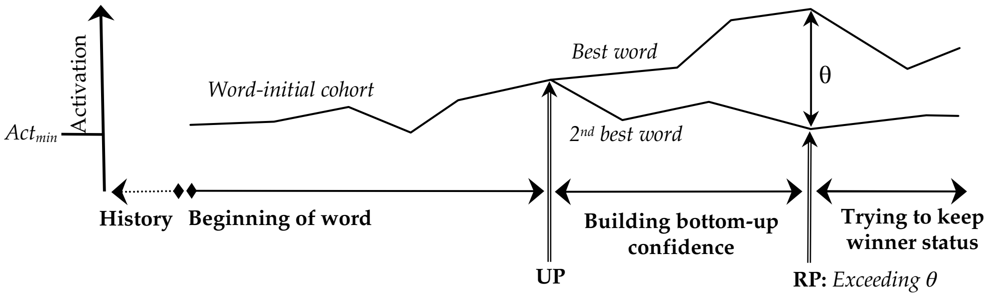

Fig. 2. Schematic illustration of the process of early decision making.

5.1. Decision Module

For the task of early decision making, the SpeM model is expanded with a decision module. The input of the decision module consists of the *N*-best list with the word sequences and their accompanying Bayesian word activation scores as created by the search module at each point in time when a new symbol is added to the phone graph. 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. We do not want SpeM to accept a word that happens to have the highest activation irrespective of the absolute value of the activation. Therefore, the value of the Bayesian word activation of the focus word itself must exceed a certain *minimum activation* (*Act$_{min}$*). In the experiments described below, various values for *Act$_{min}$* were tested.
3. Since we do not want SpeM to make a decision as long as promising competitors are still alive, the *quotient* of the Bayesian word activation of the focus word on the best-scoring path and the Bayesian word activation of its closest *competitor* (if present) must exceed a certain threshold (θ). In the experiments, various values for θ were tested.

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 2.3). 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 so do not compete with each other. (Remember that 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.) 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 in our definition of the term. 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.

Fig. 2 schematically depicts the process of early recognition. The Bayesian 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, each word has its own Bayesian 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 three decision criteria described above.

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

To determine the effect of the variables introduced in decision criteria 2 and 3, experiments were carried out in which their respective values were varied: 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. 3 shows the relation between precision (y-axis) and recall (x-axis) for a number of combinations of θ and 21 values of Act_{min}. For the sake of clarity, Fig. 3 is limited to three values of θ, viz. θ = 0.5, 1.5, 2.5; 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. 3 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 absolute and relative values of Bayesian activation that were defined as decision criteria seem to work as predictors for the early recognition of polysyllabic words.

5.3. The effect of the amount of evidence for a word on precision and recall

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 while there is still little evidence for the word. This raises the question what the effect is of the amount of evidence in support of a word (the number of phones between the start of the word and the DP) or of the 'risk' (in the form of the number of phones following the DP until the end of the word) on precision and recall. In the following analysis this question is investigated. For fixed values for Act_{min} and θ, precision and recall are calculated for different amounts of evidence, thus different 'risk' levels, as a function of the number of phonemes between the start of the word and its DP and number of phones between the DP and the end of the word.

The value for Act_{min} is set to 0.5, a value that guarantees that we are on the plateau shown in Fig. 3; θ was set at 1.625 (on the basis of results in Scharenborg et al., 2007). In these analyses, we are interested in the number of words that could in principle be recognised correctly at a certain point in time. The definitions of precision and recall are therefore adapted, such that they only take into account the number of focus word tokens that in principle could be recognised. 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 at that position in the word (accumulating to 1,463 focus words). 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*.

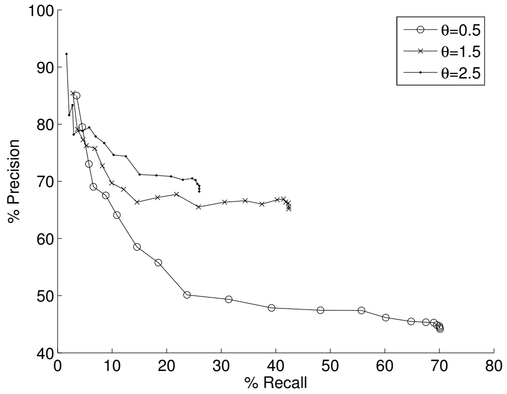

Fig. 3. For three values of θ, the precision and recall of 21 values of Act_{min} are plotted.

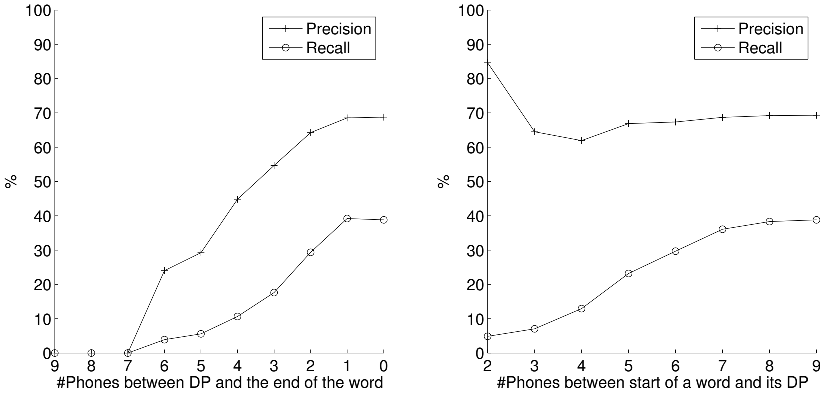

Fig. 4.The x-axes show the number of phones between the DP and the end of the word (left panel) and between the start of the word and its DP (right panel); the y-axes show for θ = 1.625 and Act_{min} = 0.5 the percentage recall (solid lines with crosses +) and precision (solid lines with circles o), respectively.

Fig. 4 shows the results. In the left panel, the x-axis shows the number of phones between the DP and the end of the word; the y-axis shows the percentage recall (line with circles o) and precision (line with crosses +), respectively. The right panel of Fig. 4 shows on the x-axis the number of phones between the start of the word and its DP; the y-axis shows the percentage recall (line with circles o) and precision (line with crosses +), respectively. The left panel of Fig. 4 clearly shows that precision and recall increase if the number of phones remaining after the DP decreases. 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, recall increases if the DP is later, so that more information in support of the hypothesis is available (see right panel of Fig. 4). 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. It should be noted, however, that the right panel of Fig. 4 suggests 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 alarms and the correctly recognised focus words does not change much.

5.4. Summary

We investigated a predictor related to the absolute and relative values of the word activation, Act_{min} and θ, respectively, for deciding whether a word is considered as recognised before the end of its acoustic realisation. The results showed that the predictor functions as a filter: The higher the values for the predictor, 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 values of the predictor is straightforward once the costs of false alarms and false rejects can be determined. In the subsequent analyses, the effect on precision and recall of the amount of evidence for a word, in terms of 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 was investigated. 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 if there are fewer phones left until the end of the word. 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 left to 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.

6. Discussion

In the laboratory, listeners are able to reliably identify polysyllabic content words before the end of the acoustic realisation (e.g., Marslen-Wilson, 1987). In real life, 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 listeners to guess the identity of content words even before their uniqueness point. In the research presented here, we investigated an alternative ASR system, called SpeM, that is able to recognise words *during* the speech recognition process for its ability for recognising words before their acoustic offset – but after their uniqueness point – a capability that we dubbed 'early recognition'. 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. The probability theory underlying SpeM makes it possible for an advanced statistical LM to emulate the context effects that enable humans to recognise words even before their uniqueness point. Such an LM would make SpeM's recognition behaviour even more like human speech recognition behaviour.

In our analyses, we investigated the Bayesian word activation as predictor for early recognition. The results in Section 5 indicate that the Bayesian word activation can be used as a predictor for on-line early recognition of polysyllabic words if we require that the quotient of the activations of the two hypotheses whose scores with first and second rank (θ) and the minimum activation (Act_{min}) of the word with the highest activity score both exceed a certain threshold. There was, however, a fairly high percentage of false alarms. In the subsequent analysis, we found that the amount of evidence supporting a decision affects the performance. If the decision point was later in the word, thus based on more 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 decision point, which also improves the performance. The predictor we chose has its parallels in the research area that investigates word confidence scores. For instance, 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 LM 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 ASR, respectively). However, in contrast to the early decision paradigm presented in this chapter, most conventional procedures for computing confidence measures are embedded in an integrated search; therefore, they only provide the scores at (or after) a point in an utterance when no new data are available that might revise the original scores.

The capability of recognising words on the basis of their initial part helps listeners in detecting and processing disfluencies, such as self-corrections, broken words, repeats, etc. (Stolcke et al., 1999). The integrated search used in ASR systems makes it difficult to adequately deal with these disfluencies. The incremental search, however, 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 detect and process potential problems such as disfluencies more accurately and faster. Furthermore, if an incremental search would be incorporated in a speech-driven application, the time needed to respond to a speaker can be much shorter. This will be beneficial for ease of use of speech-centric interaction applications.

7. Conclusions and future work

In this chapter, we showed that SpeM, consisting of an automatic phone recogniser, a lexical search module, and an early decision mechanism is able to recognise polysyllabic words before their acoustic offset. In other words, the results presented in this chapter showed that early decision making in an ASR system is feasible. This early decision making 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 that can be properly normalised to support decision making. The early recognition process is comparable to what human listeners do while decoding everyday speech: Making guesses and predictions on the basis of incomplete information.

For 81.1% of the 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 alarms. 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.

Finally, the reason for starting the research on early recognition to begin with was the potential benefits that early recognition promises for improving the speed and naturalness of human-system interaction. So far, the results of our work are promising. However, our experiments have shown that substantial further research is needed to better understand the impact of all the factors that affect and support the ‘informed guessing’ that humans perform in day-to-day interaction, and that allows them to predict what their interlocutor is going to say and when (s)he will reach a point in an utterance where it is safe to take the turn.

8. Acknowledgements

The work of Odette Scharenborg was supported by a Veni grant from the Netherlands Organization for Scientific Research (NWO).

9. References

Bouwman, G., Boves, L., Koolwaaij, J. (2000). Weighting phone confidence measures for automatic speech recognition. *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. *Proceedings of the IEEE International Conference on Document Analysis and Recognition*, (CDROM).

Garrod, S. & Pickering, M.J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*, 8-11.

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

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., Goldinger, S.D., Auer, E.T., Vitevitch, M.S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics*, *62*, 615-625.

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

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

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 Processe*s, *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., ten Bosch, L., Boves, L. (2003). Recognising 'real-life' speech with SpeM: A speech-based computational model of human speech recognition. *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 2285-2288.

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. (2007). 'Early recognition' of polysyllabic words in continuous speech. *Computer Speech and Language*, *21* (*1*), 54-71.

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. *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. *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 & 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.