Modeling the use of durational information in human spoken-word recognition

Odette Scharenborg[a)]
Centre for Language and Speech Technology, Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

(Received 5 March 2009; revised 12 February 2010; accepted 8 March 2010)

Evidence that listeners, at least in a laboratory environment, use durational cues to help resolve temporarily ambiguous speech input has accumulated over the past decades. This paper introduces Fine-Tracker, a computational model of word recognition specifically designed for "tracking" fine-phonetic information in the acoustic speech signal and using it during word recognition. Two simulations were carried out using real speech as input to the model. The simulations showed that the Fine-Tracker, as has been found for humans, benefits from durational information during word recognition, and uses it to disambiguate the incoming speech signal. The availability of durational information allows the computational model to distinguish embedded words from their matrix words (first simulation), and to distinguish word final realizations of [s] from word initial realizations (second simulation). Fine-Tracker thus provides the first computational model of human word recognition that is able to extract durational information from the speech signal and to use it to differentiate words. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3377050]

PACS number(s): 43.71.Sy, 43.71.An, 43.71.Es [JES] Pages: 3758–3770

I. INTRODUCTION

We, as listeners, are continually confronted with novel utterances that speakers may generate on the spot, and usually we encounter little to no difficulty in recognizing and understanding them. Word beginnings and endings are often clearly separated in written text in languages that use an alphabetic script. However, in spoken language, clear boundaries are often absent. This can occasionally lead to ambiguity in the speech signal (e.g., Gow and Gordon, 1995), which can be illustrated with the following example (taken from Norris, 1994). Take the phonemic representation of the phrase "ship inquiry:" [ʃipiŋkwaiəri]. This phoneme sequence contains many embedded words, such as "ink" and "choir" in "inquiry," but also words that straddle the word boundary such as "shipping" and "pink." While the speech signal unfolds over time, all these possible words will become activated and compete with one another (e.g., Allopenna *et al.*, 1998; Gow and Gordon, 1995). However, this temporary ambiguity is usually quickly solved by the listeners, and the intended word sequence is recognized.

Perceptual studies provide a clue to how listeners are able to disambiguate the incoming speech signal without a delay. There is now a vast amount of evidence, accrued over the past decades, that has shown that listeners can use subtle phonetic information, such as acoustic cues due to coarticulation and assimilation processes (e.g., Gaskell and Marslen-Wilson, 1996; Gow, 2002; Tanenhaus *et al.*, 2000) and durational and prosodic cues (e.g., Andruski *et al.*, 1994; Cho *et al.*, 2007; Davis *et al.*, 2002; Denes, 1955; Gow and Gordon, 1995; Kemps *et al.*, 2005; Salverda *et al.*, 2003, 2007; Shatzman and McQueen, 2006a, 2006b; Strange *et al.*, 1983; for a review of early work on the importance of duration in identifying speech sounds, syllables, and words, see Lehiste, 1970), for the disambiguation of temporarily ambiguous stretches of speech. Subtle phonetic detail helps to resolve the temporary ambiguity present in the speech signal by reducing the activation of words that have mismatching phonetic detail and by reducing the number of activated words during the process of the recognition of spoken words. For instance, Gow (2002) showed that the [raip] in "right berries" where the /p/ is assimilated to a [p] is not identical to the [raip] in "ripe berries." Humans show priming of the word "right" but not of "ripe" when the [raip] derived from "right" was presented. Apparently, the assimilation process preserves perceptible acoustic-phonetic evidence about the unassimilated form of the word. Davis *et al.* (2002) and Salverda *et al.* (2003, 2007) showed that listeners can make the distinction between the two interpretations of an ambiguous sequence in the case of initially embedded words, such as "ham" in "hamster", even before the acoustic end of the first syllable *ham*. Using an eye-tracking paradigm, they showed that an embedded word was more activated, i.e., attracted more eye fixations, when the ambiguous sequence came from a monosyllabic word than when it came from the longer word in which it was embedded. Salverda *et al.* (2003) concluded that a longer sequence tends to be interpreted as a monosyllabic word more often than a shorter one, and that the lexical interpretation of temporarily ambiguous sequences is influenced by duration.

Durational cues also seem to play a pivotal role in resolving temporary ambiguities that straddle word boundaries (e.g., Gow and Gordon, 1995; Shatzman and McQueen, 2006a, 2006b, and references therein). Gow and Gordon (1995) investigated the recognition of lexically ambiguous sequences that could either be interpreted as a single longer word or as two shorter words (e.g., "tulips" vs "two lips"). They found priming effects for "lips" when the participant

[a)]Electronic mail: o.scharenborg@let.ru.nl

 0001-4966/2010/127(6)/3758/13/$25.00

had just heard "two lips," but not after hearing "tulips." Analyses of the stimuli showed that the word-initial consonants (here, [l]) were longer in duration than word-internal consonants. Shatzman and McQueen (2006a) showed in two eye-tracking studies that listeners use segment duration to decide whether a speaker said "eens pot" *(once jar)* or "een spot" (*a spotlight*). They concluded that the duration of individual speech sounds is used as a cue for online word segmentation of continuous speech. Segment, syllable, and word durations are influenced by various mechanisms, such as word-initial lengthening, polysyllabic shortening, accentual lengthening, and syllable ratio equalization (e.g., Cho *et al.*, 2007; Klatt, 1976; Salverda *et al.*, 2003; Turk and Shattuck-Hufnagel, 2000).

Although there now is an abundance of evidence that durational cues play a role in resolving temporarily ambiguous stretches of speech, it is still unclear whether durational information is indeed the crucial factor, or whether listeners also use other acoustic cues, such as spectral changes (which can occur due to durational changes), formant frequency information, assimilation cues, or relative durations within the span of the syllable, to differentiate between possible interpretations of an ambiguous speech signal. Shatzman and McQueen (2006a) investigated whether other cues played a role. They showed that there were indeed other differences between the two recording contexts of their stimuli besides the duration of the [s]: namely, the duration of the closure before the stop, the duration of the target word (excluding the stop), the root mean squared (rms) energy of the [s], and the rms energy of the stop. However, it was shown that listeners only used the duration of the [s] segment to disambiguate the signal. Furthermore, Salverda *et al.* (2003) showed that when removing the durational differences between the monosyllabic word and the first syllable of the polysyllabic word, this also removes the possibility for listeners to differentiate between the two. They concluded that the production of a monosyllabic word or of the initial portion of a longer word does not always contain acoustic cues that can resolve the ambiguity, and that the duration of the ambiguous sequence, more than the word it originates from, thus determines its lexical interpretation (Davis *et al.*, 2002; Salverda *et al.*, 2003). As, so far, durational information has been the only cue shown to help the disambiguation process, this work focuses on the role of durational information for resolving the temporary ambiguity in the speech signal due to lexical embedding.

The role of subtle phonetic information is problematic for computational models of spoken-word recognition that assume a discrete, abstract prelexical level between the acoustic input and the lexicon, such as TRACE (McClelland and Elman, 1986), Shortlist (Norris, 1994), and the distributed cohort model (Gaskell and Marslen-Wilson, 1997). When confronted with an input such as [ʃipiŋkwaiəri], all words that (partly) match the input will be activated and compete with each other. However, as phonemic prelexical representations do not provide an adequate means to capture subtle phonetic detail, this results in spurious activated words in these models. Crucially, the recognition of an embedded word can only occur after its offset, resulting in a slower disambiguation of temporarily ambiguous parses for the models than for humans. There now is, as Gow and McMurray (2004) point out, a move toward using input representations that capture aspects of phonetic detail [e.g., in TRACE (although TRACE uses a limited set of abstract representations, there is the possibility of incorporating some aspects of phonetic detail) McClelland and Elman, 1986; Gaskell, 2003]. Nevertheless, computational models that are sensitive to subtle phonetic detail, take the acoustic signal as input, and in which subtle phonetic variation can be represented as some sort of continuous features do not yet exist. This paper introduces and tests such a computational model: Fine-Tracker.

Fine-Tracker is a novel computational model of human spoken-word recognition specifically designed for "tracking" subtle, or fine, phonetic information in the speech signal and using it for word recognition. Fine-Tracker takes the acoustic speech signal as its input, and therefore can be tested with exactly the same stimulus materials as used in behavioral studies, instead of using some idealized form of input representation as is done by other models of human word recognition.

Eye-tracking studies have shown (Davis *et al.*, 2002; Salverda *et al.*, 2003, 2007; Shatzman and McQueen, 2006a, 2006b) that listeners are able to extract phonetic detail from the acoustic signal and use it during the word recognition process, i.e., "on-line." Listeners thus do not need an explicit segmentation of the speech signal to use durational information (note, "duration" can only be obtained after segmentation). We investigate whether subtle phonetic detail, more specifically durational cues, in the speech signal can be automatically detected in the acoustic speech signal and used during word recognition by a computational model, without the need for segmentation of the speech signal. The first half of this paper is devoted to introducing Fine-Tracker. The second half of the paper focuses on testing Fine-Tracker with respect to modeling the human ability to detect and use durational cues during spoken-word recognition. We investigate whether durational information is beneficial for Fine-Tracker, as has been found for listeners, in two sets of simulations. In the first set of simulations, Fine-Tracker is tested on its ability to distinguish monosyllabic words from the longer words in which they are embedded, using the original acoustic stimuli of Salverda *et al.* (2003). To investigate Fine-Tracker's simulation performance, Fine-Tracker's output in terms of word activation over time is correlated with the duration of the stimuli. The second set of simulations focuses on the differences in durations of a single segment. For this set of simulations, we use the acoustic stimuli from Shatzman and McQueen (2006a). The effect of durational information on Fine-Tracker is investigated by correlating its word activations over time to the segment durations. To investigate the effect of durational information, Fine-Tracker is tested in two conditions: one in which Fine-Tracker was not able to use the durational cues in the speech signal and one where durational information was incorporated in the model. Given the accumulated evidence that listeners use durational information to resolve temporary ambiguity in the speech signal, it is to be expected that not being able to use dura-

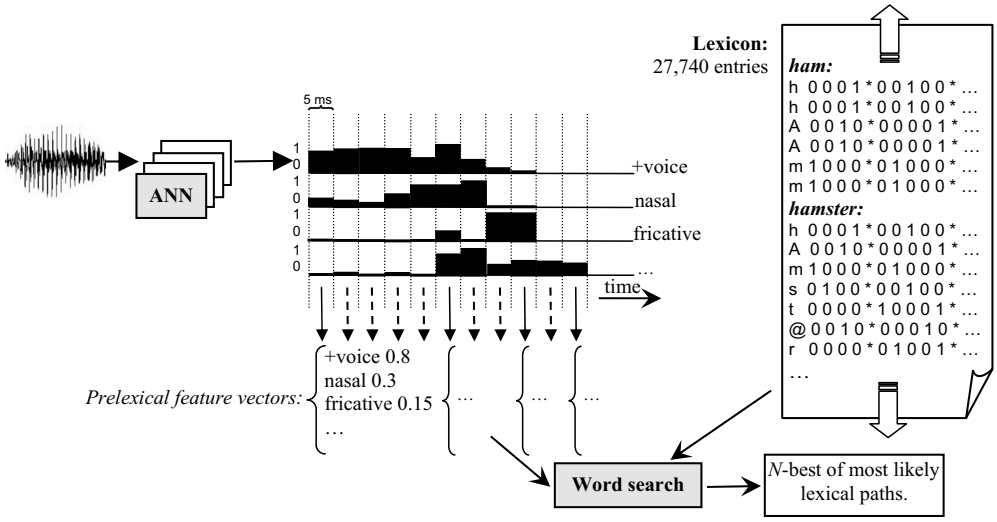

FIG. 1. Overview of Fine-Tracker: the output of the prelexical level, consisting of a set of ANNs, is the input to the "Word search" module at the lexical level.

tional information will result in a failure of Fine-Tracker to distinguish monosyllabic words from the longer words in which they are embedded and a failure to distinguish longer from shorter segments. If introducing durational information into the model proves to be beneficial to word level disambiguation, it will allow us to further investigate the effect of other subtle phonetic information on spoken-word recognition in a computational model.

II. FINE-TRACKER

Fine-Tracker is developed as part of a research line aimed at building a complete *end-to-end* computational model of human spoken-word recognition (Scharenborg *et al.*, 2003, 2005), i.e., a model that takes acoustic recordings of speech as its input. To that end, Fine-Tracker is built using techniques from the field of automatic speech recognition (ASR), and as such is part of a growing line of research aimed at bridging the research fields of psycholinguistics and ASR (for an overview, see Scharenborg, 2007).

Like its predecessor SpeM (Scharenborg *et al.*, 2005), Fine-Tracker is based on the theory underlying Shortlist (Norris, 1994). This theory holds that the human speech recognition process consists of two levels. First, listeners map the incoming acoustic signal onto so-called prelexical representations at the prelexical level. Second, at the lexical level, all lexical representations are stored in the form of sequences of prelexical units, and those lexical representations that (partly) match the prelexical representations are activated. The flow of information from the prelexical level to the lexical level is unidirectional. This means that the processing at the prelexical level is totally unaffected by lexical level processing. All words that have a good match with the input enter the lexical competition phase, where word hypotheses that overlap in time compete with each other. The result of this competition is a sequence of nonoverlapping words, usually identical to the sequence of words actually produced by the speaker. The competition phase thus resolves the temporary ambiguity of overlapping words competing with one another, and results in the optimal segmentation of the input.

Figure 1 shows an overview of Fine-Tracker; implementing the processing at two levels. The prelexical level consists of a set of artificial neural nets (ANNs) which convert the continuous acoustic signal into feature vectors with a time resolution of 5 ms. At the lexical level, the feature vectors are used as input to the word search module, which is responsible for finding the word sequence (note, a word sequence can also consist of a single word) that corresponds to the best path through the search space spanned by the prelexical feature vectors and the lexical representations. The output of Fine-Tracker is an *N*-best list of most likely lexical paths with word scores for each word on each path. The details of Fine-Tracker will be explained below.

A. The prelexical level

Prelexical representations provide a means of capturing the acoustic-phonetic information in the speech signal in terms of a limited set of predefined sub-word units. The exact form of the representations at the prelexical level is still the topic of research and debate (McQueen, 2005). In the absence of a clear answer, different models make different assumptions about the form that prelexical representations take, for example, acoustic-phonetic features (TRACE, McClelland and Elman, 1986), features (DCM, Gaskell and Marslen-Wilson, 1997), context sensitive allophones (PARSYN, Luce *et al.*, 2000), and phonemes in Shortlist and SpeM. Fine-Tracker deviates from SpeM, which also takes real speech as its input, in that it uses "articulatory features" (AFs) as prelexical representations. Fine-Tracker is therefore able to model subtle phonetic information in its lexical representations, whereas SpeM is not.

Articulatory features describe acoustic correlates of articulatory properties of speech sounds. One of the benefits of using AFs is that they are able to change asynchronously, which makes them suitable to describe the variation occurring in natural speech arising from effects such as coarticulation and assimilation. Table I shows an overview of the AFs used by Fine-Tracker. Note that *fr(ont)-back*, *round*, *height*, and *dur(ation)-diph(thong)* only apply to vowels.

TABLE I. Specification of the AFs, their types, and the number of hidden nodes in the ANNs.

AF	AF type	No. of hidden nodes
Manner	Plosive, fricative, nasal, glide, liquid, vowel, retroflex, silence	300
Place	Bilabial, labiodental, alveolar, velar, glottal, nil, silence	200
Voice	+voice, −voice	100
Fr-back	Front, central, back, nil	200
Round	+round, −round, nil	200
Height	High, mid, low, nil	250
Dur-diph	Long, short, diphthong, silence	200

For each of the seven AFs, one artificial neural net was trained for all its AF *types* (see Table I) using the NICO Toolkit (Ström, 1997). The ANNs were trained on 3410 randomly selected utterances from the manually transcribed read speech part of the Spoken Dutch Corpus [Corpus Gesproken Nederlands (CGN), Oostdijk *et al.*, 2002]. The speech files were parameterized with 12 Mel frequency cepstral coefficients (MFCCs) and log energy and augmented with first and second temporal derivatives resulting in a 39-dimensional acoustic feature vector. The MFCCs were computed using 25 ms analysis windows with a 5 ms shift. The section analyzed at every 5 ms is referred to as a "frame." To train the ANNs, the training material was labeled at the frame level. To that end, the training data were segmented at the phone level using a forced alignment with a set of 37 hidden Markov monophone models, each consisting of three emitting states, which were trained on the read speech part of the Spoken Dutch Corpus. Next, all frames belonging to a phoneme segment received the AF type labels belonging to the phoneme. During training, each ANN's performance was calculated at regular time intervals on a validation set of 379 utterances randomly taken from the CGN (disjoint from the test and training sets). The performance was evaluated using a "frame classification" task in which each ANN was forced to make one AF type decision for each frame. These frame decisions were compared to the canonical frame labels. Training was terminated when the validation set's frame classification error rate began to increase, as this indicates that the optimal ANN has been reached.

Each ANN consisted of three layers: an input, hidden, and output layer. The architecture of the ANNs was the same for all AFs, with the exception of the number of hidden nodes and number of output nodes. At the input layer, sequences of 11 frames were used with the frame to be classified in the sixth position. The output layer estimates the posterior probability of the AF type given the input. The number of output nodes is identical to the number of AF types (see Table I). The hidden layer had hyperbolic tan transfer functions and a different number of nodes[1] depending upon the AF. The optimal number of hidden units was determined through tuning experiments and is listed in the third column of Table I.

The output of the prelexical level serves as the input of the lexical level of Fine-Tracker. For each frame, each ANN creates a "soft" decision, i.e., a continuous value between 0 and 1, for each of its AF types. This numeric value can be regarded as an activation measure of this AF type over time. Per frame, the soft decisions for each of the AF types are combined into a feature vector (see Fig. 1), whose length is equal to the total number of AF types (33), resulting in a sequence of AF feature vectors with a time-spacing of 5 ms, which is fed as input to the lexical level of Fine-Tracker.

B. The lexical level

1. The lexicon

Fine-Tracker's lexicon contains all words that could potentially be recognized. The lexical representations of the words are based on the prelexical representations, so each word in the lexicon is represented in terms of AF vectors. Pronunciation variation is dealt with by adding multiple pronunciations for the specific word to the lexicon. Lexical feature vectors have the same dimension as the prelexical feature vectors, 33, and each AF type in the lexical feature vectors takes a value between 0 and 1, where 0 corresponds to the absence of the AF and 1 to the presence of the AF. The extremes of this scale can be regarded as "canonical" realizations of the AF. Intermediate values result in lexical feature vectors that are less canonical. These can be used to encode speech phenomena such as coarticulation, assimilation, and nasalization of vowels in a gradual continuous way instead of a binary decision. Note that using intermediate values might result in lexical feature vectors that are more similar, resulting in less differentiation between lexical feature vectors.

Figure 1 shows an example of the lexical feature representations of the words "ham" and "hamster" in the lexicon of Fine-Tracker. Note that the phone labels at the start of each line representing a lexical feature vector are only present for clarity purposes. These labels are not used during the word search. It is possible to assign an "unspecified" value to an AF type—this is indicated with an asterisk in the lexical feature representations in Fig. 1. During the calculation of the "goodness of fit" (see next section) between the lexical representation and the prelexical feature vector, this AF type is ignored. In this way, a match between lexical and prelexical feature vectors can deal with underspecification.

Essential in Fine-Tracker is the fact that the number of feature vectors can be set in the lexicon for each lexical item separately. An example is shown in Fig. 1 where each of the phonemes of "ham" is represented using two identical feature vectors, whereas there is only one feature vector per phoneme for the first syllable of "hamster." Fine-Tracker's word search module is able to deal with the resulting subtle differences in lexical representations. Currently, the number of lexical feature vectors is set by hand.

The lexicon is internally represented as a tree of feature vectors. When a node in the lexical tree is accessed, all words in the corresponding word-initial cohort, i.e., all words that start with the same sequence of lexical feature vectors, are equally activated. An example of the start of a lexical tree is depicted on the left hand side of Fig. 2. The "B" indicates the start of the lexical tree; each node depicts a

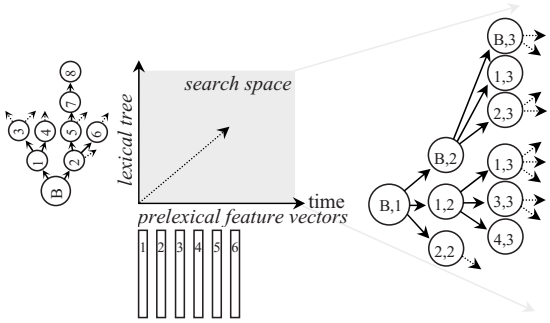

FIG. 2. The left hand side depicts a graphical representation of the lexical search implemented in Fine-Tracker: the *y*-axis shows the lexicon in the form of a lexical tree; the *x*-axis shows time in terms of prelexical feature vectors. The right hand side shows a subset of the search space nodes that are created during the word search: the first index represents the node in the lexical tree; the second index represents the number of the prelexical feature vector.

lexical feature vector. Continuous word recognition is implemented through a loop over the lexical tree: once the end of a word has been reached, the search algorithm jumps back to the start of the lexical tree.

2. The activation and competition process

There is considerable evidence that multiple candidate words are "activated" simultaneously during human word recognition (e.g., Allopenna *et al.*, 1998; Gow and Gordon, 1995; Luce *et al.*, 2000), the evaluation of which is assumed to be handled by a competition process (for an overview, see McQueen, 2005). In Fine-Tracker, following Shortlist B (Norris and McQueen, 2008) and SpeM (Scharenborg *et al.*, 2005), this process is implemented as a probabilistic word search.[2] This process is depicted in the left hand side of Fig. 2. Multiple activation of words is implemented through the use of word-initial cohorts in which all words are equally activated. As explained above, the *y*-axis shows the lexicon in the form of a lexical tree; the *x*-axis shows time in terms of the prelexical feature vectors (see also Fig. 1). The right hand side of Fig. 2 shows the search space nodes that are created during the word search. Each node refers to a position in the lexical tree (left index) and the number of the prelexical feature vector (right index). The word search algorithm is time-synchronous and breadth-first: all search space nodes at a given time (i.e., at a prelexical feature vector) are expanded before their "child" search space nodes are created. The word search algorithm starts in the root node of the lexical tree and the first prelexical feature vector, i.e., search space node (B,1). Subsequently, the child nodes of (B,1) are created using two mechanisms: (1) make a "step" in the input, but not in the lexical tree, this results in child node (B,2); (2) make a step in both the input and the lexical tree, this results in the child nodes (1,2) and (2,2). Note that this process can result in the creation of duplicate search space nodes, which is shown in Fig. 2 by two (1,3) nodes. The top (1,3) node is a child node of (B,2) resulting from making a step in the input and lexical tree; the bottom (1,3) node is a child node of (1,2) resulting from making only a step in the input. A sequence of search space nodes is called a path.

The search algorithm allows multiple 5 ms prelexical feature vectors to be mapped onto a single lexical feature vector. An example of this is the path (B,1), (B,2), (B,3), which maps the first three prelexical feature vectors onto the start of the lexical tree. This so-called "many-to-one mapping" results in chunks of speech input (or to be more precise sequences of prelexical feature vectors) mapped onto a single lexical feature vector. The goal of the word search process is to find the cheapest path, i.e., the path with the lowest *total cost* (see below), through the search space spanned by the lexical tree and the prelexical input feature vectors. This process necessarily results in the most likely word sequence given the input. Once the most likely word sequence is found, the exact mapping of lexical feature vectors onto lexical feature vectors is given at the output of Fine-Tracker. There therefore is no explicit segmentation process that chunks the input and maps it onto the lexical feature vectors. This also implies that if there is no evidence for the presence of a lexical feature vector (note: not phone as the lexical feature vectors strictly speaking do not have phone labels) in the input, a word containing that lexical feature vector can still be recognized. In that case, the number of prelexical feature vectors mapped onto that specific lexical feature vector is just one.

The cheapest path through the search space is determined by evaluating each search space node by calculating the goodness-of-fit between the prelexical and lexical feature vectors using a distance measure (DM). The present implementation of Fine-Tracker uses the averaged squared distance (ASD) between the prelexical and lexical feature vector. There is however an option in Fine-Tracker to implement other distance measures. The ASD was chosen as this measure is very similar to (Euclidean) measures that are frequently and successfully used in search mechanisms in ASR systems, and thus has proven to be successful in dealing with speech (cf. Cha. 7; Jurafsky and Martin, 2000). The relative weight of the distance measure is determined by a parameter α. The ASD is defined as follows:

$$\mathrm{ASD} = \frac{\Sigma_{\mathrm{AdmComp}}(\mathrm{lexval} - \mathrm{preval})^2}{\#\mathrm{Admissable}}. \quad (1)$$

First, the difference in raw values of each "admissible" AF type in the incoming prelexical (preval) and lexical feature vector (lexval) is determined and squared. The "admissible" AF types are those AF types without the "unspecified" marker in the lexical feature vector. Next the sum of all squared differences of all admissible AF types is normalized by dividing it by the number of admissible AF types, yielding a single ASD value between 0 and 1 measuring the dissimilarity between the two feature vectors.

The ASD value is part of the *word score* of the word in which the lexical feature vector occurs. The word score is the score from the beginning of the word up to that prelexical feature vector and corresponds to the degree of match of the word to the already processed prelexical feature vectors. It is defined as follows:

$$\text{word\_score} = \sum_{\text{prelex\_feature\_vectors}} \text{SV} + \alpha\text{DM}, \quad (2)$$

where the DM is the above-explained ASD and *step value* (SV) is either the *step-in-input* (SI) or the *step-in-input-and-lexicon* (SIL) value, depending on the current path.

- SI: value associated with making a step in the input but not in the lexicon.
- SIL: value associated with making a step in both the input and the lexicon. This is indicated with the dotted arrow in the gray box, indicating the search space, on the left hand side of Fig. 2.

The input can contain multiple words. To accommodate for word sequences, the path on which each word lies is assigned a score. The path that has the lowest *total cost* is said to have the best fit with the input. The total cost of the path then is the current word score accumulated with:

- *Word entrance penalty* (WEP): cost to start a new word, i.e., the algorithm goes through the start of the lexical tree. A higher WEP results in fewer hypothesized words, instead the algorithm will favor longer words; a lower WEP results in more and shorter words.
- *Word-not-finished penalty* (WNF): at the end of the input, i.e., when all prelexical feature vectors have been processed, all activated cohorts that do not correspond to words get a penalty. This is to penalize incomplete word hypotheses at the end of the acoustic input.
- *History*: the cost of the cheapest path from the beginning of the input up to the current search space node.

All parameters (i.e., SI, SIL, WEP, and WNF) can be tuned separately. For the simulations (and tuning experiments) presented in the current paper, the optimal parameter settings were obtained through word recognition experiments in which the evaluation criteria were, in order of importance: (1) the highest number of correctly identified words within the N-best list, where $N=50$ and where the list is obtained at the end of the utterance; (2) the number of times the correct word was on the best path; (3) the position at which the correct word was found in the N-best list (if found).

To restrict the search space, a maximum number of search space nodes, containing only the most likely candidate words and paths, are kept in memory during the word search. Furthermore, there are no duplicate paths: of identical word sequences, only the cheapest path is kept. At any moment in time the word search module can produce a ranked N-best list of alternative parses, each with its associated total cost. Each path, or parse, contains words, word-initial cohorts, silences, and any combination of these and the word score for each of these constituting items, with the restriction that a word-initial cohort can only occur as the last element in the parse. If a certain word sequence becomes more likely after the processing of more input, this word sequence will move up in the N-best list as processing proceeds; Fine-Tracker does not have to revise or recompute its parses.

In order to relate the output of a computational model to behavioral data, an important assumption of any model is a measure of how easy each word will be for subjects to respond to in a listening experiment, this measure is usually referred to as "word activation." Scharenborg *et al.* (2005) presents a way to directly compute word activations from the word scores as output by Fine-Tracker and SpeM. Since word activations can be directly computed from the word scores, we used the raw word scores in the subsequent simulations, as these give identical results to word activations.

III. SIMULATION I: LEXICAL EMBEDDING

In the first set of simulations, we investigate whether durational information is beneficial for Fine-Tracker for distinguishing fully embedded words, such as "ham," from the words in which they are embedded (i.e., the matrix word), such as "hamster," as has been found for humans (e.g., Davis *et al.*, 2002; Salverda *et al.*, 2003, 2007). The acoustic stimuli and behavioral data used are taken from the eye-tracking study referred to as "Experiment 1A" in Salverda *et al.* (2003; henceforth referred to as "Salverda"). In this study, participants were presented with manipulated Dutch sentences over headphones. The crucial difference between the sentences was the way the "target word" in each sentence was constructed. The target word is a polysyllabic word of which the first syllable also constitutes a monosyllabic word (e.g., "hamster" contains the embedded word "ham"). In constructing the target words, the first syllable was either cross-spliced from a monosyllabic word (e.g., "ham"; the MONO condition) or from the first syllable from another recording of that target word ("hamster"; the CARRIER condition). In total, 28 target words were used (see Table II).

The participants were asked to click on the picture of the target word mentioned in the sentence. The target word was represented by one of four pictures presented on a computer screen. The other three pictures consisted of two distractors and, crucially, the embedded word (here "ham"). During the experiment, participants' eye movements were monitored. Analysis of the eye movements showed that there were significantly more fixations to pictures representing monosyllabic words if the first syllable of the target word had been replaced by a recording of the monosyllabic word than when it came from a different recording of the first syllable of that target word. Although there might be multiple acoustic differences between the first syllable of the target word and a monosyllabic realization of that first syllable, Salverda *et al.* (2003) only found a significant effect for durational information to explain their findings. They concluded that listeners use durational information to distinguish between the embedded word and its matrix word.

A. Setup of the simulations

Fine-Tracker was tested in two conditions: with and without the ability to use durational information. The task set to Fine-Tracker is to reproduce the finding that pictures representing monosyllabic words attract more fixations when the first syllable of the target word has been replaced by a recording of the monosyllabic word than when it comes from a different recording of the target word. Tanenhaus *et al.* (2000) demonstrated that eye-tracking studies provide a sensitive measure of the time course of lexical activation in

TABLE II. The first column indicates the embedded and target word; in case of orthographic differences between the embedded and target word, the target word is given in full after the forward slash. The condition in which the embedded word has the highest word activation over time is indicated for the canonical lexicon and the duration lexicon; Diff syllable duration (ms)[a] shows the difference in duration of the first syllable of the target word between the two conditions; Effect size human data shows the difference in the average proportions of eye fixations to the embedded word between the two conditions.

Embedded and [target] word	Lexicon		Diff syllable duration (ms)	Effect size human data
	Canonical	Duration		
bij/[beitel]	CARRIER	MONO	20	0.19
blik[sem]	CARRIER	CARRIER	17	0.01
bok[ser]	MONO	MONO	16	0.12
ei[kel]	CARRIER	CARRIER	17	0.01
ham[ster]	CARRIER	CARRIER	11	0.01
hen[del]	CARRIER	MONO	18	−0.30
kan[delaar]	CARRIER	MONO	28	0.22
kei/[kijker]	CARRIER	MONO	20	−0.01
knip[sel]	MONO	MONO	14	−0.03
koe[kepan]	CARRIER	CARRIER	47	0.23
kok/[cocktail]	CARRIER	CARRIER	11	0.07
kom/[compact-disk]	MONO	MONO	19	0.08
la[ma]	CARRIER	MONO	28	0.05
lam[pekap]	MONO	MONO	22	−0.08
lei[ding]	CARRIER	CARRIER	15	−0.02
man[tel]	MONO	MONO	25	0.09
pan[da]	CARRIER	CARRIER	10	0.03
pen/[panty]	MONO	MONO	15	−0.03
pin[da]	MONO	MONO	21	−0.10
ree/[regenton]	CARRIER	MONO	23	0.14
roos[ter]	MONO	MONO	30	−0.01
schil[der]	MONO	MONO	24	0.20
sla[ger]	CARRIER	CARRIER	21	0.10
snor[kel]	CARRIER	MONO	12	0.00
tak/[taxi]	CARRIER	CARRIER	7	0.26
thee/[tegel]	CARRIER	CARRIER	13	0.13
tor[so]	CARRIER	MONO	19	−0.02
zee/[zebra]	CARRIER	CARRIER	5	0.12
Total	MONO: 9	MONO: 17		MONO: 18

[a]The syllable duration and human data are kindly provided by A.P. Salverda and are presented here with his kind permission. The window used to calculate the human data is identical to the window used in the original analyses by Salverda *et al.* (2003).

continuous speech, and that a simple "linking hypothesis" provides a good mapping of pattern and timing of eye fixations onto the underlying lexical activation. Following this, if we consider the amount of fixations of Salverda's participants as a degree of the word activation during word recognition, the output of the computational model can be compared with the behavioral data. We, then, expect the activation of the embedded word in the MONO condition to be higher than the word activation of the embedded word in the CARRIER condition.

Fine-Tracker is evaluated by comparing the word activations of the embedded words over time in the MONO and CARRIER conditions. If the word activations of the embedded words in the MONO condition are higher than those in the CARRIER condition, this is regarded as a correct simulation. The effect of durational information is investigated by comparing the simulation results of the conditions with and without durational information, and by correlating Fine-Tracker's word activations over time with the difference in duration of the stimuli. Finally, Fine-Tracker's results are compared to the behavioral data.

One way to code durational differences between words is in the lexical entries, which is the implementation chosen for Fine-Tracker. In the condition where Fine-Tracker is not able to use the durational cues (canonical lexicon condition), the lexical feature representations of the embedded word and the first syllable of the matrix word were kept identical. Each phoneme in the lexical representation of the words was represented by a single feature vector. In the condition where durational information was taken into account, the lexical representations of the monosyllabic words and the first syllable of the matrix words were different (the duration lexicon condition). Acoustic measurements using PRAAT (www.praat.org) showed that syllables were on average 232 ms long in the CARRIER condition, and 249 ms in the MONO condition. This 17 ms durational difference is equal to a difference of three frames at the prelexical level. To accommodate for this durational difference, each phoneme in the lexical representation of the monosyllabic word was represented by two identical feature vectors, whereas for the first syllable of the matrix word each phoneme was represented by a single feature vector (see also Fig. 1).

B. Materials

For the simulations, the speech files from Salverda's experiment are first cut manually such that the cut-out stimulus consists of the target word. When the stimulus did not allow for a clean cutting point at the start of the target word, the stimulus is cut before the target word's preceding article or adverb. Subsequently, the stimuli were parameterized with 12 MFCC coefficients and log energy and augmented with first and second temporal derivatives resulting in a 39-dimensional feature vector. The features were computed using 25 ms windows shifted by 5 ms per frame. The MFCC feature vectors were used as input to the ANN module at the prelexical level. The output of the prelexical level is then used as input to the search module at the lexical level of Fine-Tracker. The parameter settings for Fine-Tracker were optimized on the MONO test set (there was not enough data to create an independent tuning set), and subsequently tested on the CARRIER test set to ensure maximum performance on both sets. The parameter settings were the same in both conditions.

The lexical feature vector representations were obtained by substituting all phonemes of a word's canonical phonemic representation with its canonical AF vectors. The lexicon used in the simulations consists of 27 740 entries. To guide Fine-Tracker's word search, we applied priors to the 61 words that occurred in the stimuli such that they were far more likely than the other words in the lexicon. Thus, words that do not receive a prior but are in the same word-initial cohort as words that do receive the prior are far less activated. All words in a particular word-initial cohort that receive the prior are, however, equally activated.

C. Results and discussion

A prerequisite for a correct simulation is that Fine-Tracker is able to correctly identify the target and embedded words. For the canonical lexicon, for both the MONO and CARRIER condition, all 28 target and 28 embedded words were found in the 50-best list output by Fine-Tracker. For the MONO condition, the target word was first best 20 times, whereas this was the case 21 times in the CARRIER condition. For the duration lexicon, for both conditions, all 28 target and 28 embedded words were found in the 50-best list. For the MONO condition, the target word was the first best 15 times, for the CARRIER condition it was the first best 16 times.

In order to investigate the strength of Fine-Tracker's modeling ability and the effect of the ability to use durational information, we compared the word activations over time of the embedded words in the MONO and the CARRIER conditions for both lexicon conditions. To that end, the word scores for the embedded words are automatically extracted from the 50-best lists. The durational differences between the stimuli in the two conditions (and thus different numbers of prelexical feature vectors mapped onto the word-initial cohorts) means that it is not trivial to plot the word activations over time for the embedded words in the MONO and the CARRIER condition. Instead, Table II indicates in which condition (MONO or CARRIER) the embedded word had the highest word activation over time, for both the canonical and the duration lexicon. This decision was based on a comparison of the patterns of the word scores over time. The condition that had the highest word activation for the largest part of the stimulus was regarded as the "winner."

Table II shows that for the canonical lexicon, the embedded word had a higher word activation in the MONO than in the CARRIER condition for 9 of the 28 stimuli. This number increased substantially to 17 when using a lexicon that takes durational information into account. This improvement was shown to be significant ($p<0.005$) according to a one-tailed McNemar Test for related samples. The small number of stimuli decreases the certainty that the data has a normal distribution. Therefore, all statistical tests reported in this paper are nonparametric. In the McNemar test, the stimuli were pairwise compared (i.e., canonical vs. durational lexicon), where a "win" by the MONO condition was marked as "1" and a win by the CARRIER condition as "0."

We further investigated the effect of durational information on Fine-Tracker. We expect the best modeling results for Fine-Tracker for those stimuli where the difference in durational information is greatest between the monosyllabic word (i.e., MONO condition) and the first syllable in the polysyllabic word (CARRIER condition). This assumption was tested by correlating the difference in duration between the monosyllabic word and the first syllable of the polysyllabic word and the strength of the effect shown by Fine-Tracker for all 28 stimuli. As the patterns of word scores over time, which are used to make the decisions in Table II, cannot easily be used to calculate the correlation, these word score patterns were smoothed such that they were represented by a single value. The durational differences are shown in ms in the column "Diff syllable duration" in Table II for each stimulus. A positive number indicates a longer duration for the first syllable of the target word in the MONO condition. A one-tailed (bivariate) Spearman's rho test was used to investigate the correlation. The test, indeed, showed a significant positive correlation (Spearman's rho=0.448, $p<0.01$): bigger durational differences between the MONO and CARRIER conditions correlated with better modeling results of Fine-Tracker.

Fine-Tracker's preference to map longer signals onto the embedded words when using the duration lexicon, thus mapping the speech signal onto a lexical representation that consists of a doubling of the lexical feature vector for each phone, can be attributed to two aspects: the acoustic differences between the longer and the shorter signals, which is reflected in different results for the computation of the *distance measure*, and the settings of the parameters that guide the search in the word search module, more specifically, the SI and SIL. This can be clarified as follows. Imagine two signals, one consisting of 12 and one of 18 prelexical vectors, and a lexical representation consisting of six lexical feature vectors. Since the lexical representations of the embedded words are identical in both the MONO and CARRIER conditions, the SIL parameter is applied an equal number of times in both conditions. The difference thus lies in the application of the SI parameter. This parameter needs to be applied more often for a longer signal (usually the MONO condition). In order to compare the word scores of the embedded words across the two conditions at a specific point in time, a normalization needs to be carried out, i.e., the embedded word's word score at that point in time is divided by the number of prelexical feature vectors associated with the embedded word at that point in time. According to the tuning experiments, the value of SIL was to be set higher than the value of SI. With a higher number of input feature vectors, the relative contribution of the higher value for the SIL parameter to the normalized word score is less than in the case of a lower number of prelexical feature vectors, resulting in a lower average word score, thus a higher word activation (Scharenborg *et al.*, 2005), thus a better match of longer signals onto the embedded words compared to the shorter signals.

In short, subtle acoustic variation can be coded in the lexicon in Fine-Tracker, as is done for the durational differences between the embedded and target words resulting in the embedded and target words necessarily being in different word-initial cohorts. This makes it possible for Fine-Tracker to distinguish between embedded words and the first syllable of the target words. Furthermore, the SI and SIL parameters make Fine-Tracker sensitive to the subtle phonetic detail in the acoustic signal. These features allow Fine-Tracker to use durational information during word recognition. It is these features that set Fine-Tracker apart from other existing computational models of human word recognition, such as TRACE and Shortlist, which are not able to represent durational differences nor are able to use durational information during the word recognition process.

As is clear from Table II, the effect of durational information was not the same for all stimuli. This is in line with

Salverda *et al.* (2003). They also found that their main effect was not equally strong for all target words. The column "Effect size human data" shows the size of the effect in the behavioral study as the difference in the average proportions of eye fixations, calculated over the window [300–900] ms after onset of the target word, to the embedded word between the MONO and the CARRIER conditions. A positive difference means that there were on average more fixations to the embedded word in the MONO condition. As Table II shows, an effect was found for 18 of the stimuli for the listeners (Fine-Tracker: 17 stimuli). There is an overlap between Fine-Tracker and the human data in the stimuli for which an effect was found, but there were also differences. To investigate whether Fine-Tracker and the listeners showed similar behavior, the strength of the effect shown by Fine-Tracker (represented by the single value also used for the correlation with the durational differences, see above) and the human data was correlated. A one-tailed (bivariate) Spearman's rho test showed a nonsignificant correlation (Spearman's rho= −0.285, $p=0.071$). This nonsignificant correlation was further investigated by correlating the human data with the syllable duration differences between the MONO and CARRIER conditions. If this correlation is not significant, this suggests that humans use other cues besides durational information to resolve temporarily ambiguous stretches of speech. Indeed, this correlation proved not to be significant (Spearman's rho=0.111, $p=0.286$) indicating that bigger durational differences between the MONO and CARRIER conditions did not result in bigger differences in average proportion of eye fixations to the embedded word between the MONO and CARRIER conditions. These results seem to suggest that humans might use other cues, besides durational information, for disambiguation.

IV. SIMULATION II: SEGMENT DURATIONS AS A CUE TO WORD BOUNDARIES

We further investigate Fine-Tracker's ability to detect and use durational information during word recognition; this time with respect to differences in durations of a single segment. We use the results from Shatzman and McQueen (2006a). They presented listeners in an eye-tracking study with ambiguous Dutch sentences. For instance, two subsequent words could either be interpreted as "eens pot" (*once jar*) or "een spot" (*a spotlight*). The sentences were constructed such that the final [s] of "eens" and the target word (in this example) "pot" was constructed either through identity-splicing (the IDENT condition), where the [s] of "eens" and the target word were spliced from another recording of that same target-bearing sentence, or through cross-splicing (the CROSS condition), where the "eens" target word sequence was spliced from a phonemically identical sentence but where the [s] of "eens" was produced as the first segment of an [s]-plosive cluster, in our example "spot." Shatzman and McQueen (2006a) showed that the crucial difference between the two types of constructed sentences was the duration of the [s].

The participants of the study were asked to click on the picture of the target word mentioned in the sentence. The target word was represented by one of four pictures presented on a computer screen. The other three pictures consisted of two distractors, and crucially, a picture of a competitor word that had the same initial two-consonant cluster as the first segment of the target word preceded by the [s] of "eens," in our example the competitor word would start with [sp] [Shatzman and McQueen (2006a) used "spin," *spider*]. Analysis of the eye movements showed that participants used the duration of [s] as a cue for placing the word boundary. Participants made fewer fixations and were slower to fixate on the picture of the target word when the [s] in the ambiguous sequence was long, thus taken from a recording of the cluster-initial word "een spot," than when it was spliced from another recording of the "eens pot" reading of the sentence.

A. Setup of the simulations

In this simulation, we test Fine-Tracker on its ability to detect durational cues that distinguish word final from word onset [s] realizations, and use these cues to place the word boundaries. The task set to Fine-Tracker is to reproduce the findings that listeners are slower to fixate on the picture of the target word when the duration of the [s] in the ambiguous sequence is longer, and that listeners make fewer fixations to the target picture in the CROSS condition than in the IDENT condition. Considering the amount of eye fixations as a degree of the word activation, we expect the activation of the target word in the CROSS condition to be lower than in the IDENT condition, and at least the word activation to be lower at the start of the target word compared to the IDENT condition.

The setup of the simulation is that as used in the previous simulations. A simulation is correct when the target word's activation in the CROSS condition is lower than in the IDENT condition, at least at the start of the target word. Fine-Tracker's performance is evaluated by correlating the word activations over time with the difference in [s] duration in the IDENT and CROSS condition. Finally, the output of Fine-Tracker is compared with the behavioral data on a per stimulus basis.

Similarly to the previous simulation, each phoneme is represented by a single feature vector, in the canonical lexicon. In the duration lexicon, each phoneme in the canonical lexical representation of the words was represented by a single feature vector, apart from the word-initial [s]. Acoustic measurements, carried out using PRAAT, showed that the mean [s] duration was 95 ms in the IDENT condition and 110 ms in the CROSS condition, which results in a durational difference of three 5 ms frames. Taking this durational difference into account, the word-initial [s] was represented by three feature vectors, in the lexicon.

B. Materials

The stimuli consisted of 20 Dutch sentences each containing one stop-initial target word, the stop either being a [t] or a [p], preceded by the word "eens," taken from Shatzman and McQueen's (2006a) study. Again, they were cut manually such that the cut-out stimulus consisted of the "eens" followed by the target word sequence. Subsequently, the

stimuli were parameterized with 12 MFCC coefficients and log energy and augmented with first and second temporal derivatives resulting in a 39-dimensional feature vector. The features were computed on 25 ms windows shifted by 5 ms per frame. The MFCC feature vectors were used as input to the ANN module. The parameter settings for Fine-Tracker were similar to those of the previous simulations. Finally, as in the previous set of simulations, we applied priors to the 42 words in our stimuli (20 target words, 20 words that had the form [s]+target word, "een," and "eens").

C. Results and discussion

For the canonical lexicon, for the IDENT condition all 20 target words were found in the 50-best list; however, in the CROSS condition, only 18 were found. For the IDENT condition the target word was first best six times, and seven times in the CROSS condition. For the duration lexicon, for both conditions, all 20 target words were found in the 50-best list. For the IDENT condition, the target word was the first best only once, whereas it was twice first best in the CROSS condition. The word activation over time of the target words in the IDENT and the CROSS condition for both lexicon conditions were then compared following the procedure described in Sec. III C.

Table III shows, for the canonical lexicon and the duration lexicon separately, in which condition (IDENT or CROSS) the target word had the highest word activation over time, derived using the same procedure as in the previous simulation. For the canonical lexicon, for 8 out of 20 stimuli, the target word had the highest word activation in the IDENT condition. For an additional two stimuli, indicated with the asterisk in Table III, the word activation of the target word was initially lower in the CROSS condition than in the IDENT condition, although the word activation of the target word in the CROSS condition was eventually higher than in the IDENT condition. The word activation of the target word increased more slowly in the CROSS condition than in the IDENT condition, as was found for the listeners in Shatzman and McQueen's (2006a) study. For the duration lexicon, this number increased to 13 of the 20 stimuli, while for an additional two stimuli, the word activation of the target word was initially lower in the CROSS condition than in the IDENT condition. This improvement was shown to be significant ($p<0.05$) according to the one-tailed McNemar Test for related samples, in which the stimuli were pair-wise compared for the two lexicon conditions.

Subsequently, the difference in [s] duration in the IDENT and CROSS condition was correlated with the strength of the modeling effect of Fine-Tracker to test whether the best modeling results for Fine-Tracker could be found for those stimuli where the [s] duration difference is greatest. The durational differences are shown in the column 'Diff [s] duration (ms)' in Table III. A positive number indicates a longer [s] duration in the CROSS condition. Following the procedure described in Sec. III C, a one-tailed (bivariate) Spearman's rho correlation test was carried out. This correlation showed a significant positive correlation (Spearman's rho=0.620, $p<0.005$): a larger per stimulus difference in [s] duration between the IDENT and CROSS conditions correlated with a stronger modeling effect of Fine-Tracker, and vice versa.

TABLE III. The condition in which the target word has the highest word activation over time is indicated for the canonical and duration lexicon separately; Diff [s] duration (ms) shows the difference in duration of the [s] between the IDENT and the CROSS condition; Effect size human data[a] shows the difference in the average proportions of eye fixations to the embedded word between the two conditions. See the text for an explanation of the asterisk.

Target word	Lexicon		Diff [s] duration (ms)	Effect size human data
	Canonical	Duration		
pan	CROSS	IDENT	17	0.03
peen	IDENT	IDENT	25	0.22
peer	IDENT	IDENT	7	0.16
pier	CROSS	CROSS	2	−0.04
pijl	CROSS	IDENT	19	0.20
pil	IDENT*	IDENT*	16	0.13
pin	CROSS	CROSS	9	0.05
pion	CROSS	IDENT	27	0.18
pit	CROSS	IDENT	21	0.15
pot	CROSS	IDENT	17	0.07
prei	IDENT*	IDENT	16	0.01
taart	IDENT	IDENT	−2	−0.02
tand	CROSS	CROSS	7	0.05
tang	IDENT	IDENT	16	−0.02
teen	CROSS	CROSS	7	−0.10
teil	IDENT	IDENT	13	0.34
tempel	IDENT	IDENT*	12	0.19
thee	IDENT	IDENT	19	−0.20
tol	CROSS	CROSS	5	0.03
tulp	IDENT	IDENT	30	0.04
Total	IDENT: 10	IDENT: 15		IDENT: 15

[a]The human data are kindly provided by K. Shatzman and J. M. McQueen and are presented here with their kind permission. The window used to calculate the human data is identical to the window used in the original analyses by Shatzman and McQueen (2006a).

The column "Effect size human data" in Table III shows the size of the effect in the Shatzman and McQueen (2006a) study, calculated over the window [300–1200] ms after onset of the [s] by Shatzman and McQueen's (2006a), as the difference in the average proportions of eye fixations to the target word between the IDENT and the CROSS conditions. A positive difference in the average proportion means that there were on average more fixations to the target word in the IDENT condition. In this behavioral study, the main effect was not found for all stimuli. For 15 (Fine-Tracker: also 15) of the 20 stimuli Shatzman and McQueen's (2006a) found an effect, i.e., listeners were slower to fixate on the picture of the target word when the duration of the [s] in the ambiguous sequence was longer (CROSS condition) compared to when the [s] was shorter (IDENT condition). To investigate whether Fine-Tracker and the listeners showed similar behavior, the strength of the effect shown by Fine-Tracker and the human data was again correlated. Following the procedure described in Sec. III C, a one-tailed (bivariate) Spearman's rho correlation was carried out. Similarly to the simulation of the Salverda study, a nonsignificant correlation was found (Spearman's rho=0.323, $p=0.082$). In order to

investigate this nonsignificant correlation and to further investigate the hypothesis that human listeners might use other acoustic cues besides durational information for disambiguation, the human data was correlated with the differences in [s] durations between the IDENT and CROSS conditions. This correlation was also not significant (Spearman's rho $=0.349$, $p=0.66$) indicating that larger differences in [s] duration between the IDENT and CROSS conditions did not result in larger differences in average proportion of eye fixations between the IDENT and CROSS conditions. This again seems to suggest that human listeners might use other cues besides durational information.

V. GENERAL DISCUSSION

The current investigation introduced "Fine-Tracker," a computational model of human spoken-word recognition specifically designed for "tracking" subtle phonetic information in the acoustic speech signal and using it during word recognition. Two simulations, using the acoustic material from the original behavioral studies, were carried out. As has been found for humans, durational information is beneficial during word recognition in Fine-Tracker. The results for the simulations where durational information was included were significantly better than those without durational information. Durational cues allowed Fine-Tracker to distinguish embedded words from their matrix words (first set of simulations), and to distinguish word final realizations of [s] from word initial realizations (second set of simulations). Furthermore, Fine-Tracker's word activations over time correlated significantly with the durational differences between the test conditions in both simulations. These results show that Fine-Tracker is sensitive to durational cues and is able to use durational cues that distinguish whole syllables but also single segments to disambiguate temporarily ambiguous stretches of speech.

Fine-Tracker's results were also compared to the human data in order to investigate its ability to model the behavioral data. Correlation analyses between Fine-Tracker's results and the behavioral data proved to be nonsignificant. An analysis was carried out to investigate these nonsignificant correlations. The results showed that, although there is a significant correlation between Fine-Tracker's results and the durational differences in the stimuli, no such significant correlation was found between the behavioral data and the durational differences. This nonsignificant correlation suggests that human listeners employ additional acoustic cues, and perhaps additional strategies, that are not used by Fine-Tracker, for resolving temporarily ambiguous stretches of speech. However, more research is needed to investigate which acoustic cues, besides durational cues, play a role in the disambiguation process during spoken-word recognition. Incorporation of these possible other cues into Fine-Tracker is likely to result in an improvement in the modeling of the behavioral data. Despite the poor correlation between Fine-Tracker's results and the behavioral data, there are two clear advantages of using a computational model that takes the acoustic signal as its input. First, instead of using some kind of idealized input representation, the input used for the computational model can be identical to the stimuli used in the behavioral studies. Second, an end-to-end model necessarily has to deal with the whole range of issues related to the recognition of spoken words, instead of focusing on only parts of the speech recognition process like most other existing computational models of spoken-word recognition. Nevertheless, as the simulations have shown, there remain challenges for the future to improve Fine-Tracker's performance.

The simulation results showed that the target words are not always first best in the 50-best lists. There are multiple reasons why Fine-Tracker occasionally fails. First of all, as a result of the creation process of the target words, which were spliced from two different utterances, there is far more variability between the stimuli in the two conditions than just duration, as was also shown by Shatzman and McQueen (2006a). These acoustic differences between the stimuli will have an effect on the word scores of the target words due to differences in the averaged squared distance between the prelexical and lexical feature vectors for the two conditions. Second, as is shown in Tables II and III, some of the durational differences between the two conditions went in the opposite direction from the general trend which will result in problems for Fine-Tracker. Finally, the ANNs used at the prelexical level are not perfect. The frame classification error rates on the Salverda stimuli ranged from 75.9% correct for *manner* and 89.3% correct for *voice*. If the ANNs make initial errors then all following processes will be affected. Analysis of the failures of Fine-Tracker will inform us about areas where the model needs improvement.

In the current implementation, durational information is stored in the lexicon in the form of a multiplication of a feature vector. This setup allows for making lexical distinctions between, for instance, embedded words and their matrix words, while using an identical phoneme set for both words. In Fine-Tracker's lexicon, the [æ] in "ham" is identical to the representation of the [æ] in "hamster," the only difference being the number of feature vectors representing the [æ] in the lexical representation of the word, and thus its minimum duration in the signal. Segmental distinctions can be made in a similar fashion. The only difference between the feature vectors of a word final [s] and a word initial [s] is the number of feature vectors used to represent the phoneme. Fine-Tracker is therefore "able to use durational information […] both for segmental distinctions and for lexical distinctions that do not depend on differences between phonemes" (Shatzman and McQueen, 2006a). Currently, the number of feature vectors for each word in the lexicon is set by hand. However, as is shown by Tables II and III, there are large differences in the durational difference for the different stimuli between the two conditions. The positive significant correlation between Fine-Tracker's simulation results and the durational differences raises the question of what would happen if the number of feature vectors was determined specifically for each stimulus. For instance, it might be expected that for stimuli where the durational difference is (much) larger than the average difference, increasing the number of feature vectors might be beneficial. This was investigated for the target word "koekepan" and its embedded word "koe," as the durational difference between the two conditions for this

target word is much larger than average and Fine-Tracker was not able to produce a correct simulation for this target word. Increasing the number of feature vectors to four for each phoneme in the word indeed resulted in a correct simulation. This seems to suggest that making the number of feature vectors stimulus-dependent might improve Fine-Tracker's simulation power. Further research should shed light on issues such as finding the optimal number of feature vectors for each stimulus and whether this number should be equal for each phoneme in the lexical representation or not.

Within the psycholinguistic literature, the flow of information in spoken-word recognition is a controversial issue. There are computational models, such as TRACE (McClelland and Elman, 1986), that allow information to flow from the lexical to the prelexical levels, which are able to simulate well-known phenomena related to the involvement of lexical information in phonemic decision making. Simulations with Merge (Norris *et al.*, 2000), on the other hand, showed that it is possible to simulate these phenomena without information flowing back from the lexical to the prelexical level. As the issue of the flow of information is still unresolved, Fine-Tracker is based on the simplest model, i.e., without top-down information.

There are three aspects that are crucial to the model's performance: (1) the differentiation in the lexical representations between monosyllabic words and phonemically identical syllables which are part of polysyllabic words—which does not need to be encoded in the lexicon; (2) the ability to use the durational information at the prelexical level; (3) the use of this durational information at the lexical level to distinguish between the monosyllabic and the polysyllabic word. In Fine-Tracker, durational information is hard-coded in the lexicon. However, a perhaps more elegant implementation would be to incorporate durational information by allowing the lexical search to loop over a lexical feature vector, and assigning a probability to the self-transition loop, in order to allow for difference in length for monosyllabic and polysyllabic words within the same lexical representation. This would provide a way to use durational information at the prelexical level, for instance, in a prosodic analyzer such as proposed by Salverda *et al.* (2003) and Cho *et al.* (2007). It is to be expected that such an implementation, as long as it incorporates the three aspects that make the current implementation of Fine-Tracker work, will also be able to take benefit from durational cues to resolve temporarily ambiguous speech signals during word recognition, like Fine-Tracker.

In this study, we have investigated Fine-Tracker's simulation abilities with respect to a limited set of words. If one is interested in investigating more or other words than those investigated in this study, this can easily be done by including these words in Fine-Tracker's lexicon. Note, however, that, as for any computational model and ASR system, increasing the number of words in the lexicon will result in an increase in the difficulty of the task as more word hypotheses have to be investigated during the word search, which in turn often leads to a decrease in performance. If one is interested in simulating psycholinguistic findings that use pseudo- or nonwords, another issue arises. Computational models can in principle only recognize words that are in its lexicon. This means that if one is interested in simulating behavioral findings related to pseudo- or nonwords, these words have to be included in the lexicon, which means that in the strict sense they have become actual words. It is the experimenter's responsibility to remember which entries in the lexicon are actually pseudo- or nonwords and treat these words differently—the computational model in principle cannot do this.

The current study used the stimuli of the original eye-tracking studies, which consisted of a limited number of stimuli spoken by a single speaker in a single speaking style, and with little difference in overall speech rate. In future work, we will extend this research by investigating the performance of Fine-Tracker using speech from more speakers, using different types of lexical embedding, spoken in different speaking styles at different speech rates. To this end, data from the Spoken Dutch Corpus (Oostdijk *et al.*, 2002) will be extracted and analyzed. This will provide new knowledge about the nature and structure of subtle phonetic detail, and durational information specifically, in different types of speech. This knowledge can be used to further improve Fine-Tracker. Durational information extracted from real speech from the Spoken Dutch Corpus could be used to improve the lexical representations in terms of setting the number of feature vectors. Furthermore, this type of data can be used to investigate the effect of different types of acoustic cues, for instance those due to assimilation and coarticulation, on word recognition in a computational model. This can easily be investigated by using values in between 0 and 1 for the AF types in the lexical representation. Finally, Fine-Tracker does not have an explicit mechanism to deal with differences in speech rate, as so far this was implicitly controlled in the stimuli used. When using real speech, dealing with differences in speech rate will become an important issue. It is possible that a mechanism is needed that will provide a normalization of the speech rate.

To conclude, the implementation of Fine-Tracker and its successful simulations show that it is possible to develop a computational model of human spoken-word recognition that is sensitive to subtle phonetic detail, takes the acoustic signal as input, and in which subtle phonetic variation can be represented as continuous features. Fine-Tracker provides the first computational model of human spoken-word recognition that benefits from durational cues to resolve temporarily ambiguous speech signals during word recognition, as is found for humans. Fine-Tracker thus provides a good platform for further investigating the effect of durational cues in nonlaboratory speech and the role of other types of subtle acoustic cues on spoken-word recognition.

ACKNOWLEDGMENTS

This research was supported by a Veni-grant from the Netherlands Organization for Scientific Research to the author. The author is grateful to L. ten Bosch, L. Boves, A. Cutler, and J. M. McQueen for invaluable discussions related to the implementation of Fine-Tracker and the design of the experiments and comments on an earlier version of this ar-

ticle. Furthermore, the author would like to thank F. Kusters for the implementation of Fine-Tracker, T. Rietveld and F. van der Slik for help with the statistics, S. Creer for comments on an earlier version of this manuscript, and again A. Cutler for providing Fine-Tracker its name. Finally, the author would like to thank A. P. Salverda and K. Shatzman for kindly providing their data and answering questions related to them, and J. Sussman and two anonymous reviewers for their useful comments on an earlier version of this manuscript.

[1]Generally speaking, the more AF types there are to model with a single AF ANN, the more hidden nodes are needed. On the other hand, the more separable the AF types are within a single AF ANN, the fewer hidden nodes are needed.

[2]The word search module software of Fine-Tracker is implemented in JAVA and is distributed under the GNU General Public License (GPL) via http://www.finetracker.org (last viewed 4/12/2010).

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (**1998**). "Tracking the time course of spoken-word recognition using eye movements: Evidence for continuous mapping models," J. Mem. Lang. **38**, 419–439.

Andruski, J. E., Blumstein, S. E., and Burton, M. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**, 163–187.

Cho, T., McQueen, J. M., and Cox, E. A. (**2007**). "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," J. Phonetics **35**, 210–243.

Davis, M. H., Marslen-Wilson, W. D., and Gaskell, M. G. (**2002**). "Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition," J. Exp. Psychol. Hum. Percept. Perform. **28**, 218–244.

Denes, P. (**1955**). "Effect of duration on the perception of voicing," J. Acoust. Soc. Am. **27**, 761–764.

Gaskell, M. G. (**2003**). "Modelling regressive and progressive effects of assimilation in speech perception," J. Phonetics **31**, 447–463.

Gaskell, M. G., and Marslen-Wilson, W. D. (**1996**). "Phonological variation and inference in lexical access," J. Exp. Psychol. Hum. Percept. Perform. **22**, 144–158.

Gaskell, M. G., and Marslen-Wilson, W. D. (**1997**). "Integrating form and meaning: A distributed model of speech perception," Lang. Cognit. Processes **12**, 613–656.

Gow, D. W. (**2002**). "Does English coronal place assimilation create lexical ambiguity?," J. Exp. Psychol. Hum. Percept. Perform. **28**, 163–179.

Gow, D. W., and Gordon, P. C. (**1995**). "Lexical and prelexical influences on word segmentation: Evidence from priming," J. Exp. Psychol. Hum. Percept. Perform. **21**, 344–359.

Gow, D. W., and McMurray, B. (**2004**). "From sound to sense and back again: The integration of lexical and speech processes," in Proceedings of From Sound to Sense: 50+ Years of Discoveries in Speech Communication, Boston, MA, pp. 118–132.

Jurafsky, D., and Martin, J. H. (**2000**). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 1st ed. (Prentice-Hall, Upper Saddle River, NJ).

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. H. (**2005**). "Prosodic cues for morphological complexity: The case of Dutch plural nouns," Mem. Cognit. **33**, 430–446.

Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge, MA).

Luce, P. A., Goldinger, S. D., Auer, E. T., and Vitevitch, M. S. (**2000**). "Phonetic priming, neighborhood activation, and PARSYN," Percept. Psychophys. **62**, 615–625.

McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cognit. Psychol. **18**, 1–86.

McQueen, J. M. (**2005**). "Speech perception," in *The Handbook of Cognition*, edited by K. Lamberts and R. Goldstone, (Sage, London), pp. 255–275.

Norris, D. (**1994**). "Shortlist: A connectionist model of continuous speech recognition," Cognition **52**, 189–234.

Norris, D., and McQueen, J. M. (**2008**). "Shortlist B: A Bayesian model of continuous speech recognition," Psychol. Rev. **115**, 357–395.

Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**, 299–325.

Oostdijk, N. H. J., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., and Baayen, H. (**2002**). "Experiences from the Spoken Dutch Corpus project," in Proceedings of LREC—Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, pp. 340–347.

Salverda, A. P., Dahan, D., and McQueen, J. M. (**2003**). "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**, 51–89.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., and McDonough, J. (**2007**). "Effects of prosodically modulated subphonetic variation on lexical competition," Cognition **105**, 466–476.

Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**, 336–347.

Scharenborg, O., Norris, D., ten Bosch, L., and McQueen, J. M. (**2005**). "How should a speech recognizer work?," Cogn. Sci. **29**, 867–918.

Scharenborg, O., ten Bosch, L., Boves, L., and Norris, D. (**2003**). "Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition," J. Acoust. Soc. Am. **114**, 3032–3035.

Shatzman, K. B., and McQueen, J. M. (**2006a**). "Segment duration as a cue to word boundaries in spoken-word recognition," Percept. Psychophys. **68**, 1–16.

Shatzman, K. B., and McQueen, J. M. (**2006b**). "The modulation of lexical competition by segment duration," Psychon. Bull. Rev. **13**, 966–971.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**, 695–705.

Ström, N. (**1997**). "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," The Free Speech Journal **5**, 1–41.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., and Chambers, C. (**2000**). "Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing," J. Psycholinguist. Res. **29**, 557–580.

Turk, A. E., and Shattuck-Hufnagel, S. (**2000**). "Word-boundary-related duration patterns in English," J. Phonetics **28**, 397–440.