DIANA: towards computational modeling reaction times in lexical decision in North American English

L. ten Bosch\textsuperscript{1}, L. Boves\textsuperscript{1}, B. Tucker\textsuperscript{2}, M. Ernestus\textsuperscript{1,3}

\textsuperscript{1}Radboud University Nijmegen
\textsuperscript{2}Univ. of Alberta, Edmonton, Canada
\textsuperscript{3}Max Planck Institute for Psycholinguistics

\{l.tenbosch, l.boves\}@let.ru.nl, bvtucker@ualberta.ca, m.ernestus@let.ru.nl

Abstract

DIANA is an end-to-end computational model of speech processing, which takes as input the speech signal, and provides as output the orthographic transcription of the stimulus, a word/non-word judgment and the associated estimated reaction time. So far, the model has only been tested for Dutch. In this paper, we extend DIANA such that it can also process North American English. The model is tested by having it simulate human participants in a large scale North American English lexical decision experiment. The simulations show that DIANA can adequately approximate the reaction times of an average participant ($r = 0.45$). In addition, they indicate that DIANA does not yet adequately model the cognitive processes that take place after stimulus offset.

Index Terms: reaction times, local speed, participant-model comparison, computational modeling, spoken word recognition

1. Introduction

Lexical Decision is a versatile paradigm for investigating the cognitive processes involved in comprehending spoken and written words. The advent of (large-scale) experimental reference data is essential for comparing results of different experiments and the explanatory power of computational models and their underlying theories. For the recognition of written words, reference data are being created by e.g. the multi-language lexicon projects at the Center for Reading Research of Ghent University\cite{1, 2, 3}. From spoken word recognition, BALDEY\cite{4} is an example of a large-scale reference database for Dutch, while at the Department of Linguistics of the University of Alberta, the Massive Auditory Lexical Decision (MALD) project aims at providing a reference database for North American English\cite{5, 6}.

From MALD, first data (denoted MALD\textsubscript{PI}) are now available from a pilot experiment created to test stimuli and design for the full MALD database. BALDEY and MALD aim to provide an open source platform for developing and testing models of spoken word comprehension.

For BALDEY, we have shown that an end-to-end computational model of spoken word recognition, DIANA\cite{5, 6}, can accurately simulate reaction times (RT) averaged over 20 participants. DIANA can also simulate RTs of individual participants. While the theory and the computational model underlying DIANA is language-independent, the model has as yet only been tested for Dutch. In this paper we investigate whether DIANA can be extended for North American English, by using test data from MALD\textsubscript{PI}. Testing DIANA with a very different set of data might disclose issues in the model that are open for improvement. At the same time, the test might uncover issues related to the different design of MALD. While BALDEY selected a medium-size lexicon (about 2700 existing words), small enough to have all participants score all words, MALD aims for a very large lexicon, of which each participant scores only a very small (1\%) subset; as a result, most words in MALD\textsubscript{PI} have been scored by a very small number of listeners. All participants in BALDEY were native speakers of Dutch, whereas a substantial proportion of the listeners in MALD\textsubscript{PI} do not have English as their first language. Another, perhaps less important, difference between the two projects is that BALDEY used a female speaker, whereas the MALD speaker is male.

2. DIANA

As many influential models of decision making (e.g.\cite{7}), DIANA consists of three components: an activation, a decision, and an execution component (Fig. 1). As a computational model of spoken word recognition, DIANA differs from previous models, such as\cite{8, 9, 10}, in that it takes real speech as input and provides reaction times as output. Conceptually, the activation component in DIANA shares principles with Shortlist-B\cite{11}, while the decision component is reminiscent of theories proposed by\cite{12, 13, 14, 15, 16}.

Figure 1: Overview of DIANA. The model consists of three interrelated components: (1) an Activation component that takes speech as input; its output is a weighted lattice of hypotheses, evolving over time (2) a Decision component, which outputs the recognized word/non-word item and an estimated reaction time (3) an Execution component which models the time it takes from the mental decision until the eventual overt action (e.g. pressing a button).
Our approach is based on recent mathematical-psychological research on lateral inhibition between neural populations (e.g., [24, 25]), the Decision component in models of decision making that emphasize the role of lateral inhibition between neurons. Contrary to a family of neurally-based models of decision making, our approach automatically handles lexicons of approximately 40,000 entries. Entries in the lexicon are phonetically specified as (possibly several) sequences of phone symbols. Each entry is accompanied by a prior probability, which is derived from the relative frequency of the word in a text corpus. The activation of a word is determined by a weighted combination of the bottom-up acoustic match and the top-down prior probability. For the English implementation of DIANA we have used the unigram frequency counts in the Google 2012 N-gram corpus [17] and the CMU Pronouncing Dictionary [18]. Transcriptions of the words in MALDPI that were not present in the CMU dictionary were partly generated automatically, and partly constructed manually.

The activation component shares with e.g. Shortlist [8] and TRACE [9] the concepts of activation and competition between words in the mental lexicon as a function of phonetic information in the input. Similar to Shortlist B [11], words (and word sequences) are represented as competing paths in a lattice without lateral inhibition. The activation scores of word candidates are computed by a decoding algorithm borrowed from Automatic Speech Recognition (ASR). Its implementation is based on HTK [19]. For processing the utterances in MALDPI, speech is transformed into a sequence of vectors with spectral information based on Perceptual Linear Prediction (cf. [20]) augmented with delta and delta-delta parameters at a rate of 100 frames/s. This frame rate determines the phonetic evidence to be updated each 10 ms.

The acoustic models used in DIANA were taken from the American P2FA-Vislab project [21, 22] and adapted to the Canadian speaker who produced the stimuli used in MALDPI, by HERest in HTK [19], on a held out set of words from that speaker. The resulting acoustic models were assessed by performing a word recognition task with a uniform LM for this speaker: on an independent test set of 500 words and with a word accuracy of 82%, which was judged sufficient for our goal. Word confusions were found to be phonetically highly plausible.

Because the model must be able to distinguish between real words and non-words, DIANA uses two decoders in parallel, as in keyword detection [23]. The first decoder uses the lexicon to score lexical candidates. The parallel decoder computes the activation of phone sequences that are constrained by a probabilistic bigram.

Fig. 2 shows an example of the evolution of activations of the leading and runner-up word candidates for the acoustic input 'display'. The figure only shows the evolution of the top-30 runner-ups, which amounts to the top 0.1% of all potential word candidates; the majority of the remaining nearly 36,000 word candidates show activations staying far below the leading candidate activation.

2.1. Activation component

The input of DIANA is a real acoustic signal. The activation component uses the signal to compute time-varying activations for all words in the lexicon. The current implementation easily handles lexicons of approximately 40,000 entries. Entries in the lexicon are phonetically specified as (possibly several) sequences of phone symbols. Each entry is accompanied by a prior probability, which is derived from the relative frequency of the word in a text corpus. The activation of a word is determined by a weighted combination of the bottom-up acoustic match and the top-down prior probability. For the English implementation of DIANA we have used the unigram frequency counts in the Google 2012 N-gram corpus [17] and the CMU Pronouncing Dictionary [18]. Transcriptions of the words in MALDPI that were not present in the CMU dictionary were partly generated automatically, and partly constructed manually.

The activation component shares with e.g. Shortlist [8] and TRACE [9] the concepts of activation and competition between words in the mental lexicon as a function of phonetic information in the input. Similar to Shortlist B [11], words (and word sequences) are represented as competing paths in a lattice without lateral inhibition. The activation scores of word candidates are computed by a decoding algorithm borrowed from Automatic Speech Recognition (ASR). Its implementation is based on HTK [19]. For processing the utterances in MALDPI, speech is transformed into a sequence of vectors with spectral information based on Perceptual Linear Prediction (cf. [20]) augmented with delta and delta-delta parameters at a rate of 100 frames/s. This frame rate determines the phonetic evidence to be updated each 10 ms.

The acoustic models used in DIANA were taken from the American P2FA-Vislab project [21, 22] and adapted to the Canadian speaker who produced the stimuli used in MALDPI, by HERest in HTK [19], on a held out set of words from that speaker. The resulting acoustic models were assessed by performing a word recognition task with a uniform LM for this speaker: on an independent test set of 500 words and with a word accuracy of 82%, which was judged sufficient for our goal. Word confusions were found to be phonetically highly plausible.

Because the model must be able to distinguish between real words and non-words, DIANA uses two decoders in parallel, as in keyword detection [23]. The first decoder uses the lexicon to score lexical candidates. The parallel decoder computes the activation of phone sequences that are constrained by a probabilistic bigram.

Fig. 2 shows an example of the evolution of activations of the leading and runner-up word candidates for the acoustic input 'display'. The figure only shows the evolution of the top-30 runner-ups, which amounts to the top 0.1% of all potential word candidates; the majority of the remaining nearly 36,000 word candidates show activations staying far below the leading candidate activation.

2.2. Decision component

Contrary to a family of neurally-based models of decision making that emphasize the role of lateral inhibition between neuronal populations (e.g., [24, 25]), the Decision component in DIANA does not assume active inhibition between competitors. Our approach is based on recent mathematical-psychological models of decision and reaction times [12, 13, 14, 15, 16, 26, 27]. In DIANA, a decision about the winning word candidate is made at time $t$ when the activation of the leading candidate exceeds the activations of all other candidates. In this example DIANA will take a decision close to (or at) word offset, due to the very small differences between the leading candidate and the other runner-ups.

The execution component models the process from mental decision to overt behavior (e.g., pressing a button). In the present version of DIANA, the execution component adds a fixed delay to the time between the decision and the execution of the selected word.

2.3. Execution component

The execution component models the process from mental decision to overt behavior (e.g., pressing a button). In the present version of DIANA, the execution component adds a fixed delay to the time between the decision and the execution of the selected word.

3. Processing RT sequences

Reaction times are supposed to reflect the processes underlying speech comprehension [28]; therefore, they are a measure for the complexity of the cognitive processes [29]. In lexical decision experiments, RTs always come in sequences, and raw sequences tend to show fairly low correlations (0.1 – 0.3) between participants [5]. This is because individual RTs are superpositions of several different processes, each with their own time scale. At the shortest time scale, the features of individual stimuli (and the cognitive processes of interest) are at stake. In lexical decision experiments, these features include the lexical
status of the stimulus, its morphological complexity, the density of its lexical neighborhood, the word or lemma frequency, etc. (e.g. [29]). At the longest time scale, personal characteristics of the participants (e.g., their physical and mental condition, age, gender, and general cognitive abilities) affect the average RTs in a session [7]. At an intermediate time scale, slowly fluctuations of attention, change in strategy, learning effects, and fatigue affect the local speed in RT sequences, e.g. [30, 31]. Arguably, local speed effects are the most important processes that reduce the correlation between RT sequences. This is especially true in experiments in which all participants process stimuli in a different order.

In [5] we proposed a Moving Average (MA) filter to remove these local speed effects from RT sequences (detrending). The operation of the MA filter is controlled by a single parameter $0 \leq \alpha \leq 1$, which determines the number of preceding stimuli that affect the RT on the present stimulus. With data from BALDEY, it has been shown that values $\alpha \approx 0.15$ substantially increased the between-participant correlations. This range of $\alpha$ means that approximately six preceding stimuli affect the RT to the present stimulus, which is in agreement with findings reported in other reaction time studies where typical ranges of five to ten stimuli were found to have an effect [32, p. 409].

4. The Spoken English Lexicon data MALDPI

The MALDPI corpus contains data from a large scale auditory lexical decision experiment with over 26,000 words and nearly 10,000 non-words. The words were selected from a combination of three corpora: 8,000 words were extracted from the Buckeye Corpus [33], 21,000 most frequent words from COCA frequency list [34], 10,000 words from ELP [35], and 1,000 compound words, which when combined and after removal of duplicate entries resulted in a list of 28,510 words. About 9% of words were not yet processed in MALDPI. The list was designed to cover most of the high frequency words in the English lexicon; all morphological complexity was retained, as were function words. Proper nouns were removed from the list. Non-words were created using Wuggy [36] with the CMU dictionary set as the language and set to 2/3 overlap with the input item. Wuggy allowed us to quickly create nonwords written in IPA symbols based on the words in the corpus. The shortest words are monosyllabic, the longest comprise seven syllables.

Participants were undergraduate students in introductory Linguistics courses from a diverse linguistic background, with ages ranging from 17-44 years (mean 20). In an experiment session, a participant scored a set of 800 items, 400 words and 400 non-words. Due to the very large number of words and non-words, most items have only been scored by a small number of participants (min. 1, max. 12, mean 4.15). With participants showing implausible output removed, MALDPI contains RTs from 250 sessions, from native and non-native listeners.

RT sequences from an experiment as complex as MALDPI, in which many stimuli were only scored once, pose special challenges for detecting data points that are implausible, for example an RT that is improbably short or long. MA filter detrending for removing local speed effects from the individual RT sequences is part of the cleaning operation.

5. Results and discussion

For the assessment of DIANA, we selected the RT data of the 1,200 word types that were scored by 10 to 12 participants in MALDPI. This data set included RTs from both natives and non-natives. The RT scores were log transformed. Only trials with RT > 400 ms and logRT < $\mu + 2\sigma$ ($\mu$ and $\sigma$ of the logRT distribution), and for which the word/non-word decision was correct, were taken into account, leaving 9,880 trials.

Compared to BALDEY, MALDPI is more diverse in terms of the participants’ language background and seems to be more diverse in terms of participant behavior. On BALDEY, the increase on the between-participant correlation by detrending was on average 25%; on MALDPI this gain was less than 15%. In MALDPI several participant pairs show correlations $< 0.05$ between their logRT sequences (with or without detrending). There is, however, also a subgroup of participants with mutual correlations $\geq 0.15$, prior to detrending. On this subgroup, the improvement of average correlation as a function of $\alpha$ in the MA filter is small but consistent (Fig. 3). The maximum average correlation between coherent participants after optimal detrending is approximately 0.216, attained for $\alpha \approx 0.09$.

We are still in the process of investigating why part of the participants in MALDPI behave in a manner that reduces the correlation with other participants to almost 0. We suspected most of the deviant participants to be nonnatives, but this appears not to be the case. We also did not observe overall longer RTs for the nonnatives. We are currently in the process of conducting an in-depth analysis of the correlations of the ‘erratic’ participants with the RTs predicted by DIANA, which may uncover effects of specific (non)words.

Similar to the assessment of DIANA on BALDEY, we analyzed DIANA’s performance on MALDPI by comparing (in the logRT domain) its predicted RT sequence with the sequence of RTs averaged over all participants who scored the corresponding words. Before averaging of the RTs of individual participants, detrending with the optimal value of $\alpha$ was performed. The RTs simulated by DIANA depend on the setting of the parameters shown in Fig. 1. The parameter $\theta$ in the Decision component has by far the largest effect. Its value determines the proportion of words for which a decision is made before word offset. This proportion appears to affect both the average logRT predicted by DIANA and the correlation with the average RTs of the participants.

The effect of $\theta$ on the average logRT and the correlation with the average participants’ logRT is summarized in Fig. 4. The horizontal axis of the plots corresponds to the proportion of the words for which a decision is taken before word offset. For brevity, we indicate this proportion as $\tau$. The plot has two y-axes. The left y-axis relates to the correlation (circles); the right y-axis relates to the correlation (circles); the right y-axis relates to the correlation (circles).
decreases rapidly. This can be understood from the way D
all these frequent words. In this plot, the linear RT prediction
of linear RT in ms, averaged over all 1,200 words, as function
larger values of $\tau$

However, comparing absolute values of the parameters between
participants for the same value of $\tau$
found optimal on B
the correlation between D
and the average participant on B

The fast decrease of the correlation and the average RT of DIANA for higher values of $\tau$ suggests that the way in which the choiceRT is computed must be improved. Hick’s Law, on which the choiceRT contribution in DIANA’s RT prediction is based, only provides an approximate account of the decision processes that take place after stimulus offset. Quite likely, participant-dependent delays in terms of differential psychology [7] play a role, which may explain part of the seemingly “erratic” behavior discussed above.

6. Conclusions and future work

DIANA, originally implemented for Dutch, has successfully been extended to North American English, using only resources that are freely available. This strongly suggests that the theory underlying DIANA will also hold for other languages. Future research must show if similar performance can be obtained for non-IndoEuropean languages. The processing of raw RT sequences in the form of removing local speed effects to increase the correlation between participants, is useful for spotting and quantifying idiosyncrasies in RT data.

The test of DIANA on the MALDPI data has shown that the present version of the model is a powerful predictor of the average RTs of part of the participants. Another part of the participants shows almost zero correlations with the “homogeneous” group, and with DIANA. This raises questions with respect to the design of MALDPI, and the theory implemented in DIANA, both relating to the possibility that different participants use very different strategies and decision criteria in a speeded auditory lexical decision task. Decisions strategies will surely depend on the way in which pseudo-words are constructed. Early violations will induce a different strategy than late violations. DIANA does not have such a “strategic reasoning” component. Also, the threshold $\theta$ in the decision component is only based on the best runner-up. Quite possibly, the threshold in human participants also depends on the number of runners-up that are close to the leader, similar to the use of Hick’s Law [16] for computing the choice RT if no decision can be made before word offset. We have already pointed out that Hick’s law does not account for differences between participants. We will investigate whether implementation of the results of recent research on decision making such as [24, 25, 14, 15, 16, 26] will improve the correlation between DIANA and individual participants. Additionally, we will extend DIANA by taking into account findings about lexical access (e.g. [37]), phonetic details (e.g. [38]), and word representations (e.g. [39], [40]).

7. Acknowledgement

This work was funded by an ERC starting grant (284108) and an NWO VICI grant awarded to Mirjam Ernestus. Benjamin Tucker was awarded a Killam Cornerstone Grant and a SSHRC Insight Grant (435-2014-0678).
8. References


