# DIANA, AN END-TO-END COMPUTATIONAL MODEL OF HUMAN WORD COMPREHENSION

Louis ten Bosch[1], Lou Boves[1], Mirjam Ernestus[1,2]

[1]CLST, Radboud University, [2]Max Planck Institute for Psycholinguistics, Nijmegen, NL
{l.tenbosch,l.boves,m.ernestus}@let.ru.nl

## ABSTRACT

This paper presents DIANA, a new computational model of human speech processing. It is the first model that simulates the complete processing chain from the on-line processing of an acoustic signal to the execution of a response, including reaction times. Moreover it assumes minimal modularity.

DIANA consists of three components. The activation component computes a probabilistic match between the input acoustic signal and representations in DIANA's lexicon, resulting in a list of word hypotheses changing over time as the input unfolds. The decision component operates on this list and selects a word as soon as sufficient evidence is available. Finally, the execution component accounts for the time to execute a behavioral action.

We show that DIANA well simulates the average participant in a word recognition experiment.

**Keywords:** speech comprehension, computational model, end-to-end model, weak modularity

## 1. INTRODUCTION

Psycholinguistic studies show that the cognitive processes underlying speech comprehension are complex (for an overview, see e.g. [1]). Computational models that implement theories of these processes greatly facilitate our understanding of how humans recognize speech [10]. This paper presents DIANA, a new computational model of speech comprehension. DIANA builds on ideas that underlie previous computational models, most notably Shortlist B [15], and SpeM [19].

Importantly, DIANA differs from all models in two important respects. First, it simulates the complete processing chain from the on-line processing of an acoustic signal to the execution of a response, including reaction times (RTs). To our knowledge, only SpeM [19] and FineTracker [18] take the acoustic signal itself as its input, instead of some handcrafted symbolic representation of this signal, and no model simulates the execution of a response. Because DIANA simulates the complete processing chain, it can receive exactly the same input as human listeners in an auditory experiment, and its behavior can be directly compared with humans participants' behavior. Secondly, DIANA differs from all models in that it assumes minimal modularity, which implies that it does not make intermediate decisions but postpones all hard decisions towards the end of the decision making (see e.g. [12, 13, 15]).

In this paper we present DIANA and test its plausibility by comparing its simulation with participant's behavior in a word recogntion experiment.

## 2. DIANA

### 2.1. General description

DIANA simulates three interrelated processes: a word activation process, a decision process, and an execution ('effector') process. The activation process computes a probabilistic match between the input acoustic signal and the word representations in DIANA's lexicon, which, in combination with the word frequencies, determine the word activation scores. These scores are updated as the acoustic signal unfolds, and result in a ranked list of word candidates ('word hypotheses'), which changes over time. This time-varying list of word candidates forms the input to the decision process, which operates in parallel with the activation process. As soon as sufficient acoustic evidence is available, the decision process settles for a specific word. Once such a decision has been made, the execution process converts the decision into an observable action.

### 2.2. The Activation component

The activation component in DIANA takes the acoustic signal as input. Its output is a time-varying ranked list of word candidates. It makes use of a lexicon that contains orthographic and phonemic representatons of words, as well as the prior probabilities of these words. The design and implementation of DIANA supports the use of other units than phonemes, or even episodic representa-

tions [5, 6, 17].

In order to process acoustic input, DIANA converts the input signal into sequences of vectors (one vector per 10ms, 100 vectors/s), each vector consisting of 13 Mel-Frequency Cepstral Coefficients (MFCC, [2]), augmented with their first and second-order time derivatives [9].

To match the input signal with candidate words, each phone symbol in DIANA's lexicon is represented as a three-state hidden Markov model (HMM). The acoustic characteristics of each of these three states are represented by Gaussian mixture models (GMMs) that specify the distributions of the MFCC vectors associated to that state.

The Activation component of DIANA is implemented with the hidden Markov toolkit HTK [25]. As in Shortlist B [15], DIANA's matching function is based on a Bayesian framework, which allows for an elegant and mathematically principled way to compute the activation of a word candidate on the basis of acoustic evidence and the word's prior probability. During the course of the input, the activation of each word candidate is computed incrementally after each 10ms input.

In DIANA, the activation of a word candidate is computed via a weighting between the bottom-up acoustic match and the top-down prior probability:

(1)
$$P(word \mid acoustics) \sim P(acoustics \mid word) \cdot (P(word))^{\gamma}$$

The parameter $\gamma$ determines the relative weight of $P(word)$ and $P(acoustics \mid word)$. A value of $\gamma$ equal to 0 corresponds to total ignorance of word frequency; when $\gamma$ is high, the activation of a word is primarily determined by its frequency.

### 2.3. The Decision component

DIANA can make a decision as soon as the difference between the activation of the leading word candidate and alternative runner-up candidates exceed a threshold. (In this competition there is no lateral inhibition between the competing candidates, following Shortlist B[15], SPeM [19] and FineTracker [18].) That is, a decision is made if

(2)
$$\log(P(leader \mid acoustics)) - \log(P(runner\_up \mid acoustics)) > \theta$$

in which $\theta$ is a model parameter that depends on the task as well as on the strategy that the listener uses for completing the task.

**Reaction time** If the criterion (Eq. 2) is met before stimulus offset, DIANA's reaction time is the detection time (the time at which this criterion is met) plus the execution time.

It often occurs, however, that this criterion is not met at stimulus offset, in line with the finding that in many word recognition and lexical decision experiments a large proportion of the RTs (measured from stimulus onset) are (much) longer than the sum of the duration of the acoustic stimulus and the time it takes to externalize a decision by means of some action [22, 7, 3]. In that case, additional *choice reaction time*, required to differentiate between the highest ranked candidates [14, 4], is added to DIANA's reaction time. In DIANA, this additional reaction time is modelled by $\beta$ times the logarithm of the number of remaining candidates $word_i$ at stimulus offset for which $\log(P(leader \mid acoustics)) - \log(P(word_i \mid acoustics)) \leq \theta$.

### 2.4. The Execution component

The execution process in DIANA accounts for the time it takes to effectuate the decision in the form of overt behavior. In the current implementation, this delay is fixed at 200 ms for every participant and assumed to be independent of the stimulus.

## 3. WORD RECOGNITION EXPERIMENT

### 3.1. Experimental Method

**Participants** Twenty native Dutch listeners (10 male, 10 female), between 18 and 23 years of age (average 19.4), and all undergraduate students at Radboud University Nijmegen, were paid to participate in the experiment. None of them reported hearing loss or cognitive problems.

**Materials** The stimuli consisted of 613 real Dutch words: 314 nouns (125 singulars and 189 plurals), 80 adjectives, and 219 verb forms (80 infinitives, 52 present tense forms, 40 weak past tense forms, 4 strong past tense forms, and 43 participles). Each word was bisyllabic, and comprised only one stem. The average number of phonemes per word is 5.2, with a standard deviation of 1.0.

The same 613 words are also incorporated in BALDEY, a large auditory lexical decision experiment in Dutch [3], for which a female speaker has read aloud the stimuli carefully, one by one, in a sound attenuated booth. We re-used these audio recordings for our word recognition experiment. The durations of the 613 word realizations vary from 273 to 947 ms (mean: 552 ms; sd: 132 ms).

For the experiment, we created 20 random orderings of the 613 stimuli, one for each participant.

**Procedure** The participants were seated in a sound-

attenuated booth and listened to the stimuli over headphones at a comfortable loudness level. They had to press a button as soon as they had recognized the stimulus and to subsequently repeat the word.

The button box was connected to a dedicated PC running E-prime [20] as the main process. The auditory stimulus immediately stopped at the moment the button was pressed. The time interval between the onsets of subsequent stimuli was 3000 ms.

Per participant, the list of 613 stimuli was split into four sublists and participants were offered the opportunity to take short rests between sublists. One experiment session took approximately 50 minutes per participant.

**Analysis of the participant data** We analyzed the RT data in order to see whether participants show the expected patterns of word frequency and word duration. If they do, this will confirm that these RTs reflect word recognition times (instead of some task strategy). Moreover, the size of the word frequency effect will indicate the psychological plausibility of the gamma that we will need to simulate these data.

From this analysis we excluded the 1253 responses (10%) that were incorrect or for which the RT values were implausibly short (not shorter than 200 ms) or implausibly long (more than two standard deviations longer than the participant's mean).

Linear mixed effects modeling (in R v3.1.2, [23]) with as dependent variable $\log(RT)$, with as fixed effects previous $\log(RT)$, stimulus duration, trial index and word frequency (extracted from CGN, [16]), without interactions, and with subject and word as random effects (without random slopes), shows that word frequency is a significant ($\hat{\beta} = -3.22 \times 10^{-03}$, df = 1, $t = 2.42$) predictor, also after removal of outliers. In all models, (log) word duration is also a robust predictor ($\hat{\beta} = 6.52 \times 10^{-04}$, df = 1, $t = 27.9$). These and similar results show that this word recognition experiment provided valid data consistent with what could be expected on the basis of previous psycholinguistic experiments (e.g. lexical decision, [3]).

### 3.2. DIANA simulation

**DIANA's settings** For DIANA's Activation component we used speaker-independent HMM-based acoustic phone models for Dutch [8] (32 Gaussians/state), and adapted these models towards the speaker by applying HERest in HTK [25] on an independent set of 500 words.

DIANA's lexicon contains 24,878 words, including the 613 words in the experiment. The word frequencies were based on the Spoken Dutch Corpus (CGN) [16]. Words in DIANA's lexicon that did not occur in [16] were given a word frequency of 1.

**Assessment** We assessed DIANA by comparing its error rate and RTs for the correct responses with those produced by the human participants. Since DIANA does not simulate local trends in the RTs ([11, 24, 21]), local trends were removed from the observed human RTs by subtracting the folowing moving average filter:

(3)
$$\text{maRT}[i] = \alpha \cdot \log(\text{RT}[i-1]) + (1-\alpha) \cdot \text{maRT}[i-1]$$

Here, $\alpha$ is a parameter that determines the impact of previous RTs. The value of $\alpha$ is between 0 and 1. We filtered the RTs using an $\alpha$ of 0.17 since this value resulted in the highest correlation averaged across all pairs of participants (0.242). This value amounts to a history of approximately 6 previous stimuli having an effect on the RT on the current stimulus, in full agreement with [11].

DIANA RTs were assessed by computing the Pearson correlation between DIANA's simulated RT sequence and the RT sequence of the 'average participant', that is, the RT sequence that results from averaging the filtered RTs per stimulus across all participants. This average sequence can be considered the best estimation of the RTs as generated by the cognitive processes that are common to all participants.

**Modeling results** As discussed above, DIANA's performance is determined by the values of the parameters $\gamma$, $\theta$, and $\beta$. Here we only discuss the role of $\theta$. The parameters $\gamma$ and $\beta$ were set to their optimal values (0.8 and 0.1, respectively).

The parameter $\theta$ determines how much evidence the model needs to select a word candidate. A high $\theta$ implies that the model needs substantial evidence, and as a consequence produces only fewer errors but also long reaction times.

Figure 1 shows the role of $\theta$ on DIANA's behavior. In this figure, the squares show the correlation ($r$) between the log(RT) sequence produced by DIANA and the log(RT) sequence averaged over the 20 human participants, as a function of $\theta$. The circles show the percentage of stimuli for which DIANA did not make any decision before stimulus offset, as a function of $\theta$. The correlation between the simulated and the observed RT values reaches a maximum of 0.767 at $\theta = 100$. This value of $\theta$ corresponds to a rather conservative regime in DIANA's decision process: only about one third of all decisions in the Activation module are made before stimulus offset. DIANA's word error rate is then 5.2%.

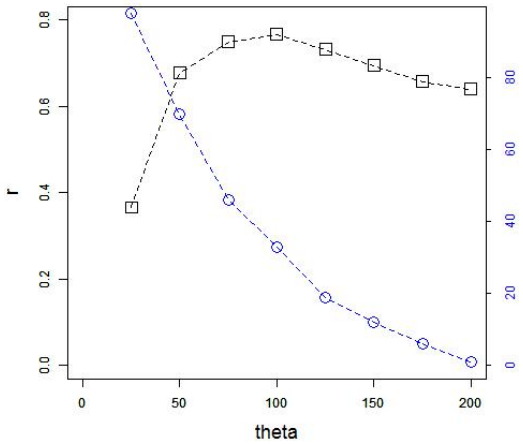


**Figure 1:** DIANA as function of $\theta$. The squares show the correlation ($r$, left axis) between the log(RT) sequences produced by DIANA and the average log(RT). The circles show the percentage (right axis) of stimuli for which DIANA did not made any decision before stimulus offset.

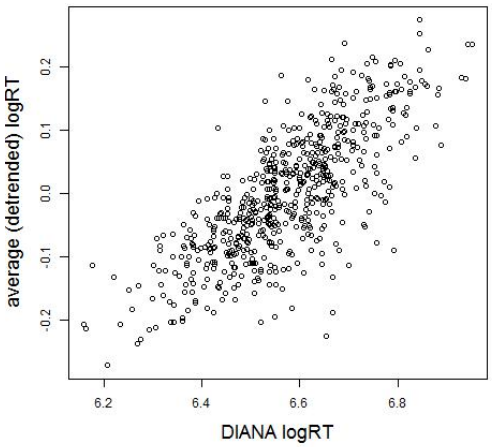


**Figure 2:** DIANA's RT predictions versus the averaged RTs from the 20 participants.

For the optimal parameter setting, Figure 2 shows the scatter plot of the log(RT) sequence generated by DIANA and the human average log(RT) sequence. The Pearson correlation is $r = 0.76$, so that DIANA accounts for 58% of the variance in the average scores of the participants.

## 4. DISCUSSION

In this paper we presented DIANA, a new computational model of word comprehension, which, in contrast to all previous models, assumes minimal modularity and simulates the complete chain from the online processing of an acoustic signal to the execution of a response. DIANA is based on previous computational models of word recognition and incorporates the knowledge that the field of psycholinguistics has gathered on human spoken word recognition in the last decennia.

In addition, we tested DIANA by conducting a word recognition experiment with human participants and by comparing the average RTs with those produced by DIANA. We found that DIANA can produce RTs that show a correlation of 0.77 with the average human RTs, so that DIANA accounts for 58% of the variance in the participants' average RTs. It obtains this correlation while producing 5.2% of word recognition errors, while the participants produced error rates in the range from 3% to 16% (average: 9.5%).

This correlation with the human participant data was obtained with values for the model parameters that are cognitively plausible. The optimal value of $\gamma$ (0.8) indicates that word frequency played a clear role, as also indicated by the statistical analyses of the human participant data. The optimal value of $\beta$ (0.1) indicates that the model does not simply select the word with the highest activation score at stimulus offset, but takes the serious word competitors into account leading to additional choice reaction time. Finally, the optimal value of $\theta$ (100) shows that the best match with participants is obtained if the model waits before selecting a word until the winning candidate substantially differed in activation score from the runner up.

These results strongly suggest that the assumptions on which DIANA is based approximate well the mechanisms underlying the word comprehension process in human listeners. Moreover, it implies that DIANA can now be taken as a baseline model that can be enriched to test hypotheses about other aspects of human speech comprehension, including the recognition of word pronunciation variants and the recognition of words (cognates and non-cognates) in a second language.

In conclusion, DIANA is a computational model of speech comprehension in that it can be tested by providing it with exactly the same input as human participants in psycholinguistic experiments and by comparing its output with these participants' output. Its performance on a word recognition task shows that it simulates well the mechanisms underlying human speech perception.

**Acknowledgement** This work was funded by an ERC starting grant (284108) and an NWO VICI grant awarded to Mirjam Ernestus.

## 5. REFERENCES


[1] Cutler, A. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. MIT Press.

[2] Davis, S., Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 357 – 366.

[3] Ernestus, M., Cutler, A. 2014. BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology* 14(1), 1–45.

[4] Friederici, A. 1995. The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language* 50(3), 259 – 281.

[5] Goldinger, S. 1996. Auditory lexical decision. *Language and Cognitive Processes* 11(0), 559–567.

[6] Goldinger, S. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological review* 105(2), 251 – 279.

[7] Goodman, J. C., Huttenlocher, J. 1988. Do we know how people identify spoken words? *Journal of Memory and Language* 27, 684–698.

[8] Hämäläinen, A., Gubian, M., ten Bosch, L., Boves, L. 2009. Analysis of acoustic reduction using spectral similarity measures. *Journal Acoustical Society of America* 126, 3227–3235.

[9] Holmes, J., Holmes, W. 2002. *Speech Synthesis and Recognition*. London and New York: Taylor and Francis 2nd edition.

[10] Jones, G., Gobet, F., Freudenthal, D., Watson, S., Pine, J. 2013. Why computational models are better than verbal theories: the case of non-word repetition. *Developmental Science* 1–13.

[11] Kelly, A., Heathcote, A., Heath, R., Longstaff, M. 2001. Response time dynamics: evidence for linear and low-dimensional non-linear structure in human choice sequences. *The quarterly journal of Experimental Psychology Section A: Human Experimental Psychology* 54(3), 805–840.

[12] McClelland, J. L. 1979. On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* 86, 287 – 330.

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

[14] Meyer, D. E., Kieras, D. E. 1997. A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic Mechanisms. *Psychological Review* 104(1), 3–65.

[15] Norris, D., McQueen, J. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115, 357 – 395.

[16] Oostdijk, N. 2000. The design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith, A., (eds), *Language and Computers, New Frontiers of Corpus Research*. Rodopi 105–112.

[17] Pierrehumbert, J. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee, J., Hopper, P., (eds), *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins 137 – 157.

[18] Scharenborg, O. 2010. Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America* 127, 3758 – 3770.

[19] Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J. 2005. How should a speech recognizer work? *Cognitive Science* 29, 867 – 918.

[20] Schneider, W., Eschman, A., Zuccolotto, A. 2002. E-prime reference guide. http://step.psy.cmu.edu/-materials/manuals/reference.pdf. Psychology Software Tools, Inc.

[21] Servant, M., Montagnini, A., Burle, B. 2013. Conflict tasks and the diffusion framework: Insight in model constraints based on psychological laws. *Proceedings of the ESCOP conference*.

[22] Taft, M., Hambly, G. 1986. Exploring the cohort model of spoken word recognition. *Cognition* 22(0), 259–282.

[23] Core Development Team, T. R. 2014. The R-project for Statistical Computing. http://www.r-project.org/. The R Foundation for Statistical Computing.

[24] Wagenmakers, E., Farrell, S., Ratcliff, R. 2004. Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin and Review* 11, 579 – 615.

[25] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. 2009. The HTK book (for HTK version 3.4). Technical report Cambridge University Engineering Department Cambridge, UK.