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

L. ten Bosch[1], L. Boves[1], B. Tucker[2], M. Ernestus[1,3]

[1]Radboud University Nijmegen
[2]Univ. of Alberta, Edmonton, Canada
[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 [1, 2, 3]. For spoken word recognition, BALDEY [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[1]. From MALD, first data (denoted MALDPI) 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 [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 MALDPI. 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 MALDPI 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 MALDPI 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.

[1]http://aphl.artsrn.ualberta.ca/?p=517

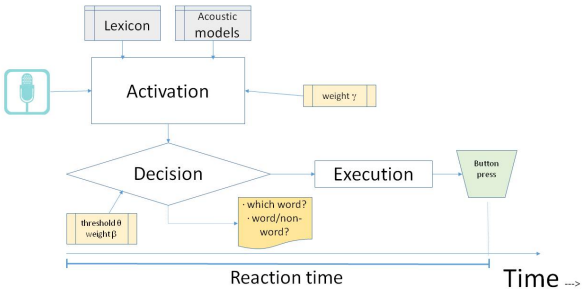

Figure 1: Overview of DIANA. The model consists of three interrelated components: (•1) an Activation component that takes speech as input; its output is a weighted lattice of hypotheses, evolving over time (•2) a Decision component, which outputs the recognized word/non-word item and an estimated reaction 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).

2. DIANA

As many influential models of decision making (e.g. [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 [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 [11], while the decision component is reminiscent of theories proposed by [12, 13, 14, 15, 16].

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 very high perplexity of 36,000, we obtained 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 phone 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 competitors with a specified threshold (a model parameter) θ. The distance between the leader and the runner-up varies in a non-linear way over time, due to the non-linear increase of acoustic evidence over time (see Fig. 2) and, evidently, the word-dependent density of the lexical neighborhood. Depending on the value of θ, the threshold may or may not be exceeded before stimulus offset. In the latter case, DIANA uses an extension of Hick's Law [16] to estimate the additional *choice reaction time*, i.e., the additional time it takes to choose between a number of alternatives left at stimulus offset:

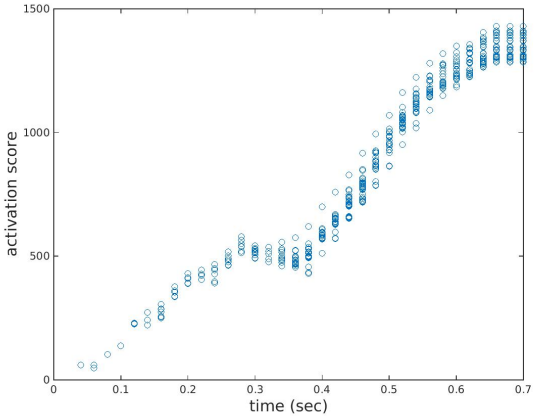

Figure 2: Activation scores for the leading word candidate and the top-30 runners-up for the word *display*. In this example DIANA will take a decision close to (or at) word offset, due to the very small differences between the leading candidate and competitors such as *displays, displace, displaced, displacing, displaying, displayed, disclaim, desolate*.

$$\text{choiceRT} = \beta \cdot H \tag{1}$$

where H denotes the *entropy* $-\sum p_i \log_2(p_i)$ of the set of probabilities p_i of the active word candidates at stimulus offset. If there is only one candidate left, $H = 0$ and so the additional choiceRT equals zero. In a typical lexical decision experiment, the entropy at word offset tends to be rather large, due to the presence of non-word phone sequences that are very similar to real words.

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 time it takes the decision component to identify the word. This delay models the time it takes to execute a planned movement.

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 α 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. Nonwords 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 nonwords, 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

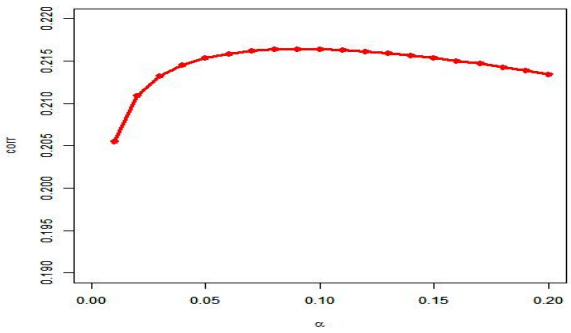

Figure 3: Average between-participant correlation for coherent participants as a function of detrending parameter α (along horizontal axis).

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 $< \mu + 2\sigma$ (μ and σ 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 ≥ 0.15, prior to detrending. On this subgroup, the improvement of average correlation as a function of α 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 α was performed. The RTs simulated by DIANA depend on the setting of the parameters shown in Fig. 1. The parameter θ 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 θ 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 τ. The plot has two y-axes. The left y-axis relates to the correlation (circles); the right

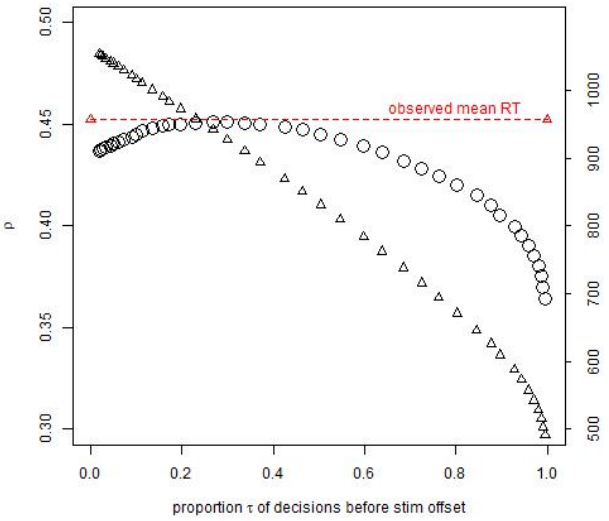

Figure 4: This figure shows the correlation between DIANA and the average participant (circles, left y-axis), as well as the predicted linear RT (triangles, right axis), both as function of the proportion τ of stimuli for which the decision is taken before stimulus offset (x-axis).

y-axis relates to the predicted *linear* RT (triangles). The plot with circles shows the correlation between DIANA's predicted logRTs and the logRTs for the virtual average participant for the set of 1,200 frequent words, as function of τ. The correlation first increases to about 0.451 (for $\tau = 0.29$) and decreases for larger values of τ. The plot with triangles shows the predicted linear RT in ms, averaged over all 1,200 words, as a function of τ; the horizontal red line is the observed average RT across all these frequent words. In this plot, the linear RT prediction decreases rapidly. This can be understood from the way DIANA simulates the RT: if a decision is made before offset, the RT will be equal to the decision moment. The number of competing candidates at word offset is zero, so that the value of H in choiceRT is 0. If no decision is taken before word offset, the RT predicted by DIANA is the duration of the word plus the choiceRT, which is now positive because $H > 0$. The right-hand side of the plot shows the unrealistic limit case in which DIANA is aggressively taking decisions before stimulus offset, yielding a very poor match with the linear RT prediction (right y-axis) and a poor correlation with the average participant in the logRT domain (left y-axis). The best regime for DIANA is a mild decision regime, in which decisions before offset are only made for a minority of the stimuli. The same mild regime was found optimal on BALDEY. However, the found optimal correlation (0.451) between DIANA and the 'average participant' on MALDPI is low compared to the correlation between DIANA and the average participant on BALDEY (> 0.61).

The fact that the curve of the correlation and the curve of the predicted average RT approximate the average RT of the participants for the same value of τ might suggest that the value of θ that corresponds to this value of τ is some kind of absolute optimum. This is the more so because for the Dutch BALDEY set the maximum correlation between DIANA and the average RTs of the participants was obtained for the value of θ for which a decision was taken before word offset for 25% of the words. However, comparing absolute values of the parameters between BALDEY and MALDPI can be misleading, if only because the nonwords in BALDEY were constructed in a different manner than in MALDPI. Also, it must be realized that the eventual RT predicted by DIANA depends on the delay in de Execution component. The average predicted RT curve in Fig. 4 was obtained with an execution delay of 100 ms. Different values would shift the triangle curve upward or downward, thereby changing the average predicted RT and shifting the value of τ for which this curve crosses the horizontal red line to the left of right.

The fast decrease of the correlation and the average RT of DIANA for higher values of τ 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 θ 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

[1] E. Keuleers, P. Lacey, K. Rastle, and M. Brysbaert, "The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words," *Behavior Research Methods*, vol. 44, no. 1, pp. 287–304, 2012.

[2] E. Keuleers, K. Diependaele, and M. Brysbaert, "Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords," *Frontiers in Psychology*, vol. 1, no. 174, 2010. [Online]. Available: http://www.frontiersin.org/language_sciences/10.3389

[3] L. Ferrand, B. New, M. Brysbaert, E. Keuleers, P. Bonin, A. Méot, M. Augustinova, and C. Pallier, "The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords," *Behavior Research Methods*, vol. 42, no. 2, pp. 488–496, 2010.

[4] M. Ernestus and A. Cutler, "BALDEY: A database of auditory lexical decisions," *Quarterly Journal of Experimental Psychology*, vol. Advance online publication, 2015.

[5] L. ten Bosch, L. Boves, and M. Ernestus, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in *Proceedings of Interspeech*, Lyon, France, 2013.

[6] ——, "Comparing reaction time sequences from human participants and computational models," in *Proceedings of Interspeech*, Singapore, 2014.

[7] J. J. Lee and C. F. Chabris, "General cognitive ability and the psychological refractory period: Individual differences in the mind's bottleneck," *Psychological Science*, vol. 24, no. 7, pp. 1226 – 1233, 2013.

[8] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, vol. 52, pp. 189–234, 1994.

[9] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, pp. 1 – 86, 1986.

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

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

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

[13] R. Ratcliff, "Continuous versus discrete information processing: Modeling the accumulation of partial information," *Psychological Review*, vol. 95, pp. 238 – 255, 1988.

[14] M. Usher and J. L. McClelland, "On the time course of perceptual choice: The leaky competing accumulator model," *Psychological Review*, vol. 108, pp. 550 – 592, 2001.

[15] R. Bogacz, E. Brown, Moehlis, P. E., Holmes, and J. D. Cohen, "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks," *Psychological Review*, vol. 113, pp. 700–765, 2006.

[16] M. Usher, Z. Olami, and J. L. McClelland, "Hick's law in a stochastic race model with speed-accuracy trade-off," *Journal of Mathematical Psychology*, vol. 46, pp. 704 – 715, 2002.

[17] "Google Books Ngram Viewer," http://storage.googleapis.com-/books/ngrams/books/datasetsv2.html, accessed: 2014-12-30.

[18] "The CMU Pronouncing Dictionary," http://www.speech.cs.-cmu.edu/cgi-bin/cmudict, accessed: 2014-12-20.

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

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

[21] "Penn phonetics lab forced aligner toolkit (p2fa)," http://www.ling.upenn.edu/phonetics/p2fa/, University of Pennsylvania Department of Linguistics, accessed: 2015-01-10.

[22] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," in *Proceedings of Acoustics '08*, 2008.

[23] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[24] J. Pearson and M. Platt, "Dynamic decision making in the brain," *Nature Neuroscience*, vol. 15, no. 3, pp. 341–342, 2012.

[25] L. T. Hunt, N. Kolling, A. Soltani, M. W. Woolrich, M. F. Rushworth, and T. E. Behrens, "Mechanisms underlying cortical activity during value-guided choice," *Nat Neurosci*, vol. 15, no. 3, p. 470, 2012.

[26] S. Brown and A. Heathcote, "The simplest complete model of choice response time: Linear Ballistic Accumulation," *Cognitive Psychology*, pp. 153–178, 2008.

[27] C. Donkin, L. Averell, S. Brown, and A. Heathcote, "Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator model," *Behavior Research Methods*, vol. 41, pp. 1095–1110, 2009.

[28] R. Whelan, "Effective analysis of reaction time data," *The Psychological Record*, vol. 58, no. 3, pp. 475 – 483, 2008.

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

[30] M. Ernestus and R. H. Baayen, "The comprehension of acoustically reduced morphologicaly complex words: the roles of deletion, duration, and frequency of occurrence," in *Proceedings of ICPhS*, Saarbrücken, 2013, pp. 773–776.

[31] I. Hanique, E. Aalders, and E. Ernestus, "The robustness of exemplar effects in word comprehension," *The Mental Lexicon*, vol. 8, pp. 269–294, 2013.

[32] T. L. Thornton and D. L. Gilden, "Provenance of correlations in psychological data," *Psychonomic Bulletin & Review*, vol. 12, pp. 409 – 441, 2005.

[33] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," www.buckeyecorpus.osu.edu, 2007, columbus, OH: Department of Psychology, Ohio State University (Distributor).

[34] M. Davies, "The corpus of contemporary American English as the first reliable monitor corpus of English," *Literary and Linguistic Computing*, vol. 25, no. 4, pp. 447–65, 2010.

[35] D. Balota, M. Yap, M. Cortese, K. Hutchison, B. Kessler, B. Loftis, J. Neely, D. Nelson, G. Simpson, and R. Treiman, "The English Lexicon Project," *Behavior Research Methods*, vol. 39, pp. 445–459, 2007.

[36] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior Research Methods*, vol. 42, no. 3, pp. 627–633, 2010.

[37] M. Gaskell and N. Dumay, "Phonological variation and inference in lexical access," *Journal of Experimental Psychology*, vol. HPP 22, pp. 144–158, 1996.

[38] S. Hawkins, "Roles and representations of systematic fine phonetic detail in speech understanding," *Journal of Phonetics*, vol. 31, pp. 373–405, 2003.

[39] C. McLennan, P. Luce, and J. Charles-Luce, "Representation of lexical form," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 29, pp. 539 – 553, 2003.

[40] L. Ranbom and C. Connine, "Lexical representation of phonological variation in spoken word recognition," *Journal of Memory and Language*, vol. 57, pp. 273–298, 2007.