Towards an end-to-end computational model of speech comprehension: simulating a lexical decision task

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

[1]CLS/CLST, Radboud University, [2]Max Planck Institute for Psycholinguistics, Nijmegen, NL

l.tenbosch@let.ru.nl, l.boves@let.ru.nl, m.ernestus@let.ru.nl

Abstract

This paper describes a computational model of speech comprehension that takes the acoustic signal as input and predicts reaction times as observed in an auditory lexical decision task. By doing so, we explore a new generation of end-to-end computational models that are able to simulate the behaviour of human subjects participating in a psycholinguistic experiment. So far, nearly all computational models of speech comprehension do not start from the speech signal itself, but from abstract representations of the speech signal, while the few existing models that do start from the acoustic signal cannot directly model reaction times as obtained in comprehension experiments. The main functional components in our model are the perception stage, which is compatible with the psycholinguistic model Shortlist B and is implemented with techniques from automatic speech recognition, and the decision stage, which is based on the linear ballistic accumulation decision model. We successfully tested our model against data from 20 participants performing a large-scale auditory lexical decision experiment. Analyses show that the model is a good predictor for the average judgment and reaction time for each word.

Index Terms: speech comprehension, computational model

1. Introduction

One of the central questions in psycholinguistics is how humans process speech. Researchers approach this question by conducting behavioural and (more recently) brain imaging experiments, and by formulating theories that account for the experimental data. Most theories are "verbal", that is, they are expressed as narratives or scenarios. Some of these verbal theories have also been implemented in computer programs, which has the advantage that the theories have to be formulated explicitly and in detail. More importantly, computer models make it easier to determine whether the theories indeed account for all experimental data and to derive predictions about the outcomes of experiments which have not yet been conducted. These experiments can subsequently be conducted in order to test the theory. So far, nearly all computational models of speech comprehension start from abstract representations of the speech signal, often in the form of phonemes, which are hand-crafted by the researchers themselves, rather than from the actual acoustic signals [13, 12]. The few models that do start from the acoustic signal (e.g. FineTracker, [17]) do not directly model reaction times obtained in comprehension experiments, such as lexical decision experiments. This paper introduces a model of speech comprehension that takes the acoustic signal as its input and predicts reaction times in a lexical decision task. The design of this model is part of a long-term project aiming at understanding human speech comprehension with a focus on word representations and processing of reduction in speech. In this paper we present the skeleton implementation of the model and show that it can already simulate the behaviour of subjects in a lexical decision task. In addition, we discuss several fundamental issues that play a role in designing and testing computational models of human speech comprehension.

The structure of this paper is as follows. Section 2 describes the data from a large-scale lexical decision experiment. The computational model is presented in section 3. Sections 4 present some results of the lexical decision experiment, and the approximation of these results by the model. In section 5 we discuss several issues that emerged from the modelling exercise. Section 6 presents the main conclusions of our work.

2. The lexical decision experiment

In an auditory lexical decision experiment, participants must decide whether a spoken item is, or is not, an existing word in a predefined language (e.g. English). Behavioural measures include reaction times (RTs) and proportions of erroneous decisions. The RTs have been shown to correlate, among others, with the frequency of occurrence of the word, the number of similar sounding words (density of the lexical neighborhood), and the position of the uniqueness point.

The lexical decision experiment that we aim to model uses Dutch words and non-words as auditory stimuli, carefully spoken in isolation by a Dutch female speaker in a neutral voice. The entire experiment contained 5541 different items (words and non-words). Subjects were 10 female and 10 male native speakers of Dutch, who heard all these stimuli. In our simulation experiment we focused on modelling the RTs in a subset consisting of 613 bisyllabic monomorphemic existing words. Furthermore, since RT distributions typically comprise some extremely fast and a long tail of very slow reactions that arguably are not representative of the 'normal' cognitive processes, we ignored all RTs shorter than 550 ms and all RTs $> \mu + 2\sigma$. This removed approximately 1% and 7% of the RTs, respectively. As a result we had 9986 RTs from 20 subjects to be modeled.

3. Description of the model

3.1. A three stage architecture

Our model hinges on the concept of activation and competition between words in the mental lexicon as a function of phonetic information at the input, which was introduced in previous models such as Shortlist [13] and TRACE [12]. Similar to Shortlist B [14], words (and word sequences) are represented as competing paths in a phone lattice; therefore, there is no lateral inhibition. Contrary to Shortlist and Shortlist B, the input of our

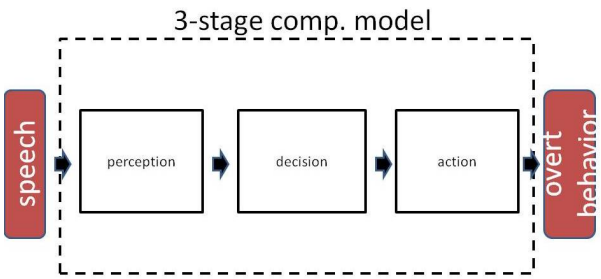

Figure 1: Overview of the computational model simulating a lexical decision experiment as performed by human subjects. The model consists of three interrelated stages (1) a perception stage that takes speech as input; its output is a weighted lattice of hypotheses, evolving over time (2) a decision stage, which outputs the recognised word/non-word item and an estimated reaction time (3) a stage modelling the time it takes from the mental decision until the eventual action (e.g. pressing a button) is actually performed.

model is real speech, so that it can be tested with the speech heard by the subjects in the lexical decision experiment. While previous models only accounted for the recognition process per se, our model incorporates two additional stages, a decision stage and an action stage. The resulting three-stage architecture is depicted in Fig. 1. In each stage, a number of model parameters play a role that govern the behaviour of that stage and thereby the behaviour of the entire model. Depending on the exact parameter settings in the different stages, the model can be entirely deterministic or rather probabilistic.

3.2. Perceptual stage

The perceptual stage is implemented in the form of an Automatic Speech Recognition (ASR) algorithm. Speech is transformed into a sequence of 13 Mel Frequency Cepstral Coefficients vectors ('frames') augmented with delta and delta-delta parameters (cf. [7]) at a rate of 100 frames/s. Therefore, the phonetic evidence in the model is updated at 10 ms intervals. ASR systems consist of acoustic models of the phone(me)s of a language, a lexicon in which words are represented as sequences or networks of phones and a language model that specifies the prior probability of sequences of discrete units such as phones, morphemes and words.

The acoustic models used in the model were taken from an existing speaker-independent ASR system [5] (Gaussian Mixture Model, 32 gaussians/state, 3-state HMM per phone) and adapted to the speaker that produced the stimuli used in the lexical decision experiment. We adapted the means and the weights of the Gaussians in the mixtures, by using HERest in HTK [8], on a randomly chosen subset of 500 stimuli from the target speaker that did not include the stimulus set used for the reaction time modelling. The resulting acoustic models were assessed by performing a word recognition task for this speaker: on an independent test set of 2780 words we obtained a word accuracy of 95.2%.

Because the model must be able to distinguish between existing words and non-words, each stimulus is processed in two ways, as is usual in keyword detection in ASR [9]. The first way is a conventional ASR decoding using a lexicon that contains the canonical pronunciation of all *words* in the task. In this decoding all words had the same prior probability. In parallel, the input speech is decoded using a lexicon that only contains the phones, in combination with a language model that reflects the canonical phone sequences of all 2780 existing words that were used in the lexical decision experiment.

The parallel decoders each produce list of the 20 most likely outputs (words in the first decoder, phone sequences in the second), together with the likelihood of each output given the input speech. These scores are computed by accumulating the scores of all 10 ms frames on the path in the lattice, and therefore the final scores depend on the duration of the input speech. To allow comparison of the scores for inputs of different durations, the final scores were divided by the number of frames in the path. The resulting "normalised" scores are the input for the next model stage. In the current implementation of the model these scores are only available after the processing of the input speech is completed.

A stimulus is likely to be a 'word' if the first decoder wins, i.e. if the score of the first entry in the word-based output exceeds the score of the first entry in the phone-based output. Since words also are phone sequences, it may happen that the corresponding phone sequence receives a higher score than the full word. To avoid erroneous non-word decisions, the model has a balance parameter λ to give preference to deciding 'word'; the value of λ is negative, stimulus-independent, and added to the phone sequence score (see section 4.2).

3.3. Decision stage

The decision stage in our model is based on a recent mathematical-psychological model of decision and reaction times [2], [3]. Decision models, and in particular models aiming at modelling reaction times, can be complex, even for binary choices. RT distributions of correct and incorrect responses tend to be complex, and attempts to simulate these distributions have resulted in ever more complex models (e.g. [11, 16, 18]; see [1]). Recently, [2] showed that it is possible to generate all distributions obtained in behavioural experiments by means of a relatively simple model, the linear ballistic accumulator (LBA) model. The LBA model is based on the idea that items on which decisions must be made accumulate evidence over time. Competing hypotheses are represented by individual accumulators, which may grow at different rates. Also, different accumulators may start from different levels. The decision is determined by the accumulator that first races across a common evidence threshold θ, while the time point at which the threshold is crossed determines the reaction time. In the LBA, this rate of growth is constant for each accumulator: evidence thus increases in a linear and deterministic manner. RT *distributions* can be obtained by randomly choosing initial evidence values from a uniform distribution and growth rates from a Gaussian distribution. The LBA model successfully simulates empirical phenomena that have proven difficult to simulate by concurrent models, in both binary and multi-class tasks (see [19]).

In our model we use a simplified version of LBA. All alternatives (the 20 hypotheses from the word decoder and the 20 hypotheses from the phone sequence decoder) are given the same *starting value* (so we do not model yet a possible prior effect of the relative frequency of existing words), while the *growth rate* of the 40 accumulators is determined by the normalized score of the 40 hypotheses, computed in the perception stage.

3.4. Action stage

The action stage models the process from mental decision to overt behaviour (e.g., pressing a button). In the present version

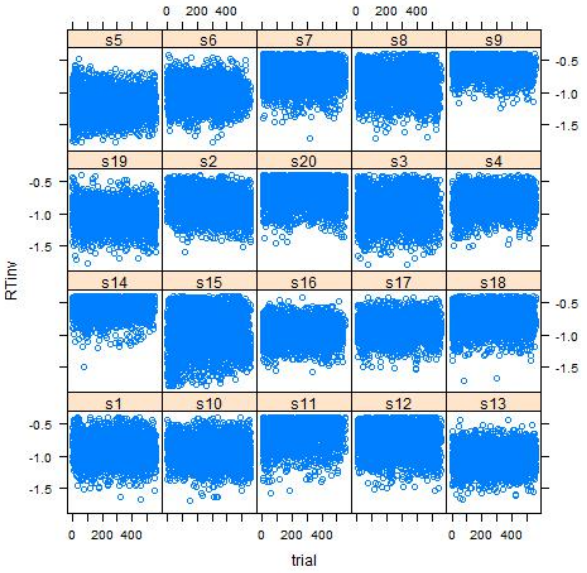

Figure 2: Reaction times on stimuli, grouped per subject.

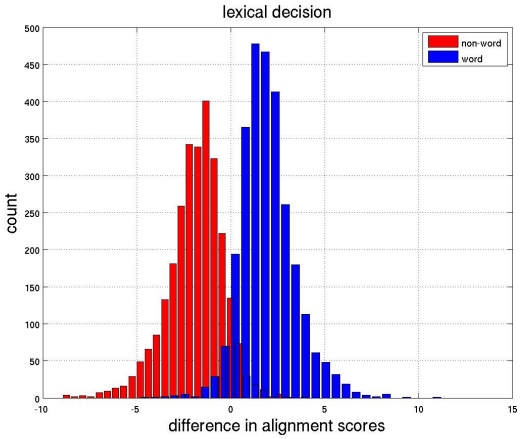

Figure 3: This plot shows the separability of items linguistically tagged as words (blue, right distribution, N=2780) and non-word items (red, left distribution, N=2761)). The horizontal axis represents the difference between two model scores: the model score based on the decoding lexicon with words, and the model score based on the phonotactically constrained phone loop. Positive and negative values indicate 'words' and non-words as classified by the model, respectively.

of our model this stage only adds a fixed delay (100 ms) between the output of the decision stage and observable action. It is easy to make this delay subject-dependent, and to add random variation to each stimulus.

4. Results

4.1. Behavioural data

Figure 2 shows the RT data of the 20 subjects in the lexical decision experiment. The inverses of the RTs (in ms) are displayed along the vertical axis; the stimuli are displayed along the horizontal axis. It is obvious that some subjects are relatively slow and others are relatively fast. Also, the variance differs between the subjects. Pearson correlation values (paired by stimuli) between pairs of subjects ranged between 0.1 and 0.3 (with mean 0.18) when measured using $1/\text{RT}$ instead of RT, a transformation that makes the per-subject distributions more Gaussian.

The apparent differences between the 20 subjects raise the question what a computational model should simulate. Is there a unique set of cognitive processes shared by all subjects, so that a model with a unique set of parameter values should be used, or are the processes different between individual subjects or groups of subjects, in which case it would be necessary to have the model operate with different set of parameter values?

4.2. Word/non-word modelling

The word/non-word decisions made by our model are based on the score s_w for the first hypothesis in the word list output and the score s_p of the best-scoring phone sequence in the perception stage. If $s_w - (s_p + \lambda) > 0$ the stimulus is labelled as an existing word. The value of λ was optimised on a held-out set consisting of 200 words and non-words to minimise the equal-error-rate (EER) between the word and non-word sets. The minimal EER obtained was 6.0% for a value of $\lambda = -2.8$. In Fig. 3, the separation of the word and the non-word stimuli is visualised. Along the horizontal axis, the parameter $s_w - (s_p + \lambda)$ is plotted. The left and right histograms are associated to non-word and word items, respectively. The estimation is not error-free: of the words (right histogram) about 4% is classified as non-word; of the non-words, about 7% is classified as words. For the human subjects, these percentages are 6% and 5%, respectively.

4.3. Modelling reaction times

We assess the accuracy with which our model simulated RTs by means of Pearson correlations between the (noise free) RTs predicted by the model and the (noisy) RTs obtained for the 20 individual subjects in the lexical decision experiment. Again, RTs were transformed into 1/RT to enhance the Gaussianity of the distributions. Figure 4 shows the correlations of the $1/\text{RT}$ in the form of a heat plot, in which the model serves as the 21^{th} subject. The average correlation between the model and each of the subjects is 0.47, significantly larger than the average subject-subject correlations (0.1–0.3). This suggests that the model captures cognitive processes that are shared by all subjects, and that the seemingly low correlations between pairs of subjects are due to the presence of substantial noise in the RTs of individual subjects to individual stimuli. At the moment it is not possible to say where that noise is generated, in the perception, decision or action stage.

5. Discussion

The three-stage model for predicting RTs in a lexical decision task is able to simulate reaction times and the associated word/non-word decisions on the basis of real speech input. A comparison of the reaction times obtained in an auditory lexical decision experiment to the model simulations shows a correlation with each of the human subjects (Fig. 4) that is significantly higher than the subject-subject correlations in this experiment, which range between 0.1 and 0.3. These between-human-subject values are low, but it is still reasonable to assume that the subjects participating in a lexical decision task

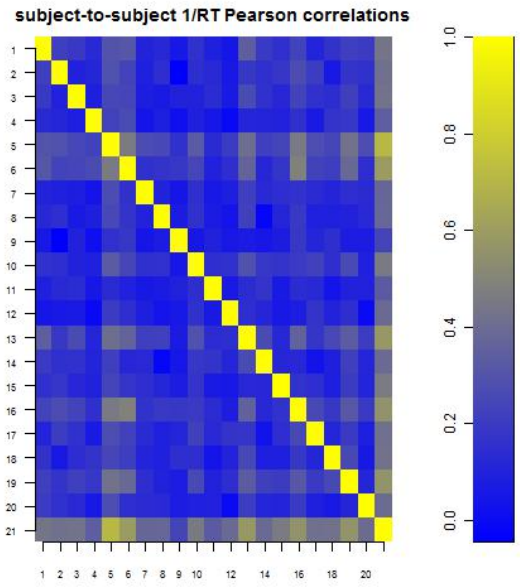

Figure 4: Matrix plot of Pearson correlations of $1/\text{RT}$ between all 20 subjects and the model, the model serving as 21^{st} subject. The correlation between the model and each of the subjects is significantly larger than the between-subject correlations.

share a common underlying cognitive processes, which are inherently quite noisy. The fact that the average correlation between the model output and individual subjects is much higher (0.47) suggests that if we could remove the noise from the subjects' responses, the between-subject correlation will improve by at least 0.1-0.15. This finding raises the question what the behaviour of a virtual 'average' subject S_{virt} would be. To investigate this, a virtual subject was designed by assigning an RT to each word based on the RTs from the 20 human subjects $(S_i, i = 1, \ldots, 20)$ as follows:

$$\frac{1}{\text{RT}_{S_{virt}}} = \text{mean}\left(\frac{1}{\text{RT}_{S_i}}\right)_{i=1,\ldots,20} \quad (1)$$

Not surprisingly, this virtual subject has a high average Pearson correlation with the 20 human subjects (0.52). (The averaging in eq. 1 takes place in the $1/\text{RT}$-domain, as all correlations are based on $1/\text{RT}$ values.) It is remarkable, however, that the computational model (with its average correlation with the subjects of 0.47) comes close to this virtual 'average' subject.

From the scatter plots in Fig. 2 it can be seen that subjects differ in two aspects: the location of their RT distribution (slow versus fast subjects), and the variance of the RTs. In our model the location of the RT distribution can be simulated in at least two ways: by making the growth rate in the decision stage subject-dependent and by making the delay in the action stage subject-dependent. Future simulation experiments must show if the two options yield different results (within reasonable limits of the parameter values). In addition, behavioural experiments must be designed that can tear apart the contributions of these stages in the human subject.

The model presented in this paper was deliberately simplified. For one thing, it does not model the effect of the relative frequency of the words. Even with reliable estimates of these relative frequencies, their application must be conceptually justified: in the perception stage (as a language model in de decoding stage, in which the scores of paths are a weighted sum of acoustic and language model scores, in accordance with the conventional ASR-based approach) and/or in the decision stage (as prior evidence, where accumulators start with this prior evidence before stimulus onset). We also refrained from adding random noise in any stage of the model. In future versions of this model these additions will be made, for example by adding frequency information, by sampling the delay in the action stage from a Gaussian distribution, and by adding noise to the starting levels of the accumulators and the growth rates in the decision stage. For refinements of the updated model, especially of the first stage, we will attempt to take into account findings about lexical access (e.g. [4]), phonetic details (e.g. [6]), and word representations (e.g. [10], [15]).

In the current implementation of our model (as in the LBA model) the growth rate is assumed to be constant over time. This is equivalent to assuming that the phonetic evidence for a word increases linearly over time, an assumption that may not be valid. The constant rate assumption allowed us to make the operation of the perception and decision stages fully sequential. In future versions of the model we intend to have the two stages operate in parallel, such that the growth rate of the accumulators can be adapted at 10 ms intervals based on the phonetic evidence generated in the perception stage.

The *comparison* between a computational model of human behaviour on the one hand and experimental behavioural data on the other hand is arguably the most difficult issue in computational modelling. Such a comparison can be performed in various ways. In this paper, we decided to assess the model on the basis of its correlation with RTs obtained from 20 subjects, as if the model were the $(N + 1)^{th}$ subject. However, correlations disregard differences between the means and variances of the RTs of individual subjects. As long as we do not know whether these differences are due to different values of parameters in essentially the same cognitive processes in all subjects, so that it is safe to assume that all subjects use the same processes, it is necessary to also try to simulate the absolute values of the RTs of individual subjects. The extent to which such simulations are successful allows us to estimate the probability that the assumption that all subjects use essentially the same cognitive processes can be maintained.

6. Conclusions

A computational model of word comprehension has been presented. It is the first computational model of speech comprehension that is able to take speech as input and simulate reaction times. It architecture consists of three interrelated stages (perception, decision, and action). The model has been used to simulate reaction times and word/non-word judgements obtained from a lexical decision experiment based on nearly 10000 measurements. The model estimates reaction times with a correlation of about 0.47 with human subjects, which is substantially larger than the between-subject correlations which lie in the range 0.1-0.3. Using correlation as a measure for similarity, the model is very similar to a virtual 'average' subject, which reaches an average correlation with the human subjects of 0.52.

7. Acknowledgements

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

8. References

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

[2] Brown, S.D., and Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. Cognitive Psychology, pp. 153–178.

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

[4] Gaskell, M.G. and Dumay, N., (1996). Phonological variation and inference in lexical access, J. Exp. Psychol. HPP 22, pp. 144–158.

[5] Hämäläinen, A., Gubian, M., ten Bosch, L., and Boves, L. (2009). Analysis of acoustic reduction using spectral similarity measures. J. Acoust. Soc. Am. Volume 126, Issue 6, pp. 3227–3235.

[6] Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. Journal of Phonetics 31, pp. 373–405.

[7] Holmes, J. and Holmes, W. (2002). Speech Synthesis and Recognition, Taylor and Francis Group.

[8] Young, S., et al. (2009). The HTK Book (for HTK version 3.4). Cambridge University Engineering Department.

[9] Keshet, J., Grangier, D., and Bengio, S. (2009). Discriminative Keyword Spotting. Speech Communication 51 (4), pp. 317–329.

[10] McLennan, C.T. Luce, P.A., and Charles-Luce, J. (2003). Representation of Lexical Form, Journal of Experimental Psychology: Learning, Memory, and Cognition 29: pp. 539–553.

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

[12] McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18(1), pp. 1–86.

[13] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52(3), pp. 189–234.

[14] Norris, D. & McQueen, J. (2008). Shortlist B: A Bayesian Model of Continuous Speech Recognition. Psychological Review 115, No. 2, pp. 357–395.

[15] Ranbom, L.J., and Connine, C.M. (2007). Lexical representation of phonological variation in spoken word recognition, JML 57, pp. 273–298.

[16] Ratcliff, R. (1988). Continuous versus discrete information processing: Modelling the accumulation of partial information. Psychological Review, 95, pp. 238–255.

[17] Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. Journal of the Acoustical Society of America, 127 (6), pp. 3758–3770.

[18] Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. Psychological Review, 108, pp. 550–592.

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