Thresholding word activations for response scoring – Modelling psycholinguistic data

Christina Bergmann[1,2], Louis ten Bosch[1], Lou Boves[1]

[1] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2] International Max Planck Research School for Language Sciences,
Radboud University Nijmegen, The Netherlands
{c.bergmann, l.tenbosch, l.boves}@let.ru.nl

Abstract

In the present paper we replicate simulations of infant word learning and the effect of variation in the input. We then investigate to what extent the results are influenced by the way in which the continuous response functions are treated and what effects the use of thresholds can have on the data. Our results show that the underlying response pattern, as uncovered by different thresholds, varies greatly. Nonetheless, the overall output of the model is often correct and able to generalise to unseen data. Thus, we show that the model can give correct responses even in uncertain circumstances. Links of this finding to language acquisition research are discussed.

Index Terms: First language acquisition, Machine learning, Modelling infant data

1. Introduction

Language acquisition research is becoming an increasingly interdisciplinary field, ranging from psychology and linguistics to artificial intelligence and machine learning. These disciplines make different theoretical assumptions and employ various research methods in investigating the same phenomenon, namely how an infant can acquire the native language in only a few years and from what seems essentially incomplete and partly inconsistent input. As a consequence of the theoretical and methodological differences, there are substantial gaps between the disciplines [1], which have to be bridged to allow for mutually beneficial collaborations and possible cross-fertilisation.

One way in which bridges can be designed (if not built) is by means of computational simulations of what otherwise would remain abstract theories. Obviously, this approach requires data from behavioural experiments on which theories can be constructed, and that can subsequently be used to test computational implementations of the theories. One example of such an effort is a number of studies that investigate the internal representation of 'words' depending on whether infants learn from a single or from multiple speakers, as reviewed in [2]. In summary, infants seem to encode speaker identity when they hear a new word spoken by a single speaker and have difficulty recognising the same word spoken by a different speaker. When infants are familiarised with the new words with speech of multiple speakers, their performance on speech of 'strangers' is significantly better. One way to interpret these results is as an indication that children use variability during language acquisition to learn which part of the information/variation in the input is critical and which parts of the acoustic signal can be ignored. Simulation data from a computational model essentially replicate the behavioural data [3, 4].

The results of experiments with infants can only be reported in terms of observable behaviours, even if one is interested in fundamentally unobservable phenomena such as representations in the mental lexicon [1]. However, 'observable behaviours' is a concept that must be carefully defined. In looking-while-listening paradigms behaviours are usually measured in terms of discrete categories (e.g., right or left, correct or wrong). Actually, much richer data is available (e.g., for eye tracking studies: dynamic information of fixations, saccades, scanning behaviour, duration of fixations, etc.), which is discretised to make it possible to apply conventional statistical analysis techniques such as ANOVA or *t*-tests.

A common procedure for discretisation of continuous (fixation) data is the use of thresholds, which determine what is counted as a fixation, whether a look was on the correct part of the screen, and so forth. Usually, these thresholds are fixed and motivated by previous literature that might concern different tasks, purposes or even age groups. Thus, the choice of thresholds is likely to affects outcomes.

The aim of this paper is to investigate the effect of such decision thresholds on the assessment of infant behaviour in more detail. Rather than trying to re-analyse data from infant experiments, we investigate the issue by means of computational simulations. The output of computational models of language acquisition, such as the models developed in the ACORNS (ACquisition Of Recognition and communicatioN Skills) project (http://www.acorns-project.org) (e.g. [4]), can be (semi-)continuous functions, much like the proportion of the time an infant fixates one of the pictures in a looking-while-listening experiment. Previous experiments with the ACORNS models almost invariably discretised the output of the model (a graded activation vector) into two response categories: correct or incorrect. In the present paper we replicate model simulations [3] and investigate to what extent the results and the interpretation of the model simulations are influenced by the use and choice of thresholds.

2. The ACORNS Model

To investigate cross-situational word discovery in young infants, the ACORNS project used computational models. An important feature of all approaches within the ACORNS project is the use of continuous real speech as input, without preprocessing steps to discretise the input via segmentation or forced alignment of phone labels (for a number of recent models that discretise the input, see [5]). As a consequence, no (unwarranted) lexical, phonetic or phonological knowledge is assumed in the learner.

The present paper focuses on a particular implementation of the ACORNS model (ACORNS-NMF), which is based on Non-negative Matrix Factorisation (NMF) [6]. Until now, studies using ACORNS-NMF, such as [4], have reported on the model's performance in terms of a specific definition of response accuracy that is best described as 'winner-takes-all': while all words being learned may be activated to some extent, the hypothesis with the highest activation is considered as the only response. Although this can be justified, there are obviously alternative ways for representing the responses.

2.1. Non-Negative Matrix Factorisation

Non-negative Matrix Factorisation (NMF) [7] simulates learning by finding a decomposition of an $n \times m$ dimensional input matrix $\mathbf{V}$, consisting of m utterances, each encoded as a vector $\vec{v}$ of total dimension n (representing the acoustic features $\vec{v}_a$ and conceptual keyword labels $\vec{v}_k$ of an utterance) into the product of two smaller matrices $\mathbf{W} \cdot \mathbf{H} \approx \mathbf{V}$ by minimising the Kullback-Leibler divergence between the input $\mathbf{V}$ and the product of $\mathbf{W}$ and $\mathbf{H}$. The dimension of $\mathbf{W}$ is $n \times r$, and the dimension of $\mathbf{H}$ is $r \times m$. The constant r (a model parameter) is chosen such that $(m+n)r \ll m \times n$, i.e., information is compressed. In the present experiment, r equals 70.

The internal matrix $\mathbf{W}$ has the same structure as the input in $\vec{v}$, namely an acoustic part and a conceptual keyword encoding part. Hence, each column vector in $\mathbf{W}$ can be considered to represent an association between acoustic and semantic information of keywords. $\mathbf{H}$ contains information about activation of columns in $\mathbf{W}$ during training. In the present paper, an incremental version of NMF is used [6]. This version can claim substantial cognitive plausibility, because it needs only to memorise a small number of most recent utterances, in addition to the internal representations in the matrix $\mathbf{W}$ of the words that are being learned.

During testing, a new utterance is given to the model in the same acoustic encoding $\vec{v}_a$ as in the training [8], but without providing the corresponding keyword part $\vec{v}_k$. The meaning of the utterance (the identity of the keyword) is then inferred by first searching $\hat{h}$ such that $\vec{v}_a \approx \mathbf{W}_a \cdot \hat{h}$ (via minimising the Kullback-Leibler divergence). Here, $\hat{h}$ is estimated using only the learned acoustic representations within $\mathbf{W}_a$. Next, the missing keyword information is reconstructed by $\hat{\vec{v}}_k = \mathbf{W}_k \cdot \hat{h}$, which uses the conceptual parts of the columns in $\mathbf{W}$ weighted by $\hat{h}$. Note that both the activation values in $\hat{h}$ and the keyword values in $\hat{\vec{v}}_k$ take real values, unlike the binary keyword labels in $\vec{v}_k$ presented to the learner during training.

2.2. Training and Testing

The model was trained with nine different keywords, embedded in varying carrier sentences. Recordings of four native speakers of English, two male and two female, were used. To investigate the possible effects of the way in which the real-valued keyword labels in the response vectors are converted to decisions whether an utterance was correctly or incorrectly recognised, we replicated parts of a previous study [3]. That study investigated whether variation during learning aids generalisation, in particular how learning from one or multiple speakers affects recognition of unknown talkers (as in infants [2]). The authors found that when the model learned keyword – speech associations from one speaker only, recognition accuracies for the same speaker were very high, but for unknown speakers they were degraded. However, when learning took place with multiple speakers, recognition was vastly improved for unknown speakers.

We ran a variant of the study, modelling the difference that variation makes by training the model with one speaker at a time. Thus, the model first experienced no variation and only learns from utterances spoken by a single speaker. Then, a second speaker was used for training, which increased variation in the model's input. This held then, too, for the onset of the third and fourth speaker. The model was first trained with 60 occurrences of all keywords spoken by one speaker (for a total of 540 utterances), then by the second speaker, and so forth until the model was trained with all 2160 utterances. Testing was conducted independently for each speaker, so that the difference in performance for yet unknown speakers could be assessed. Thus, we took full advantage of this blocked presentation of speakers during training and could assess the model's generalisation abilities to unknown speakers when it has observed one to three speakers.

To test the model's recognition performance, a held-out test set was used containing 20 utterances per keyword and speaker, amounting to a total of $20 \times 9 \times 4 = 720$ test samples. During testing, the model was frozen in its current state and thus could not learn from being exposed to the test utterances or new speakers. Therefore, the sentences and even speakers in the test set remained unknown to the model and could be used at every time point. To get insight into the model's behaviour as training proceeds, we tested the model at every tenth utterance for 100 utterances from the point of a speaker change onwards. For the remainder of the training, we assessed the model's responses after 90 training utterances had been observed.

2.2.1. *Relative and absolute thresholds*

To make full use of the model's response, that is, to take advantage of the fact that each test utterance yields a vector of activations instead of a single answer, we assessed the model's recognition performance using thresholds for the activation of the correct concept. To this end, the model's response for all nine keywords is normalised to sum up to 1. Then, two thresholds are defined, a relative and an absolute threshold (θ_{rel} and θ_{abs}, respectively). Responses of the model that fall below the thresholds were counted as *undecided*, a separate response category next to correct and incorrect. The thresholds can be interpreted as determining the 'certainty' of the model regarding its decoding of a test utterance.

The relative threshold θ_{rel} takes into account the difference in normalised activation of the maximum and the runner-up. By setting this threshold, a minimal difference between the two highest activated keyword labels is enforced. The absolute threshold θ_{abs} imposes a minimum on the normalised activation, irrespective of the competition that the winner faces. Due to the normalisation, both thresholds depend on each other to a certain extent. Nonetheless, several scenarios are still possible. The model can have a high activation and an almost as highly activated runner-up or have the remaining activation distributed relatively evenly over all other competitors.

'Accuracy' is defined as the ratio of the number of correct responses (that meet a given threshold criterion) and the number of test utterances that are presented to the model. Thus, 'undecided' responses are counted as incorrect when computing accuracy. To obtain a complete picture of the performance, accuracy must be combined with the proportion of utterances for which no decision is made. Accuracy values in previous reports correspond to $\theta = 0$ for both types of thresholds. Due to the nor-

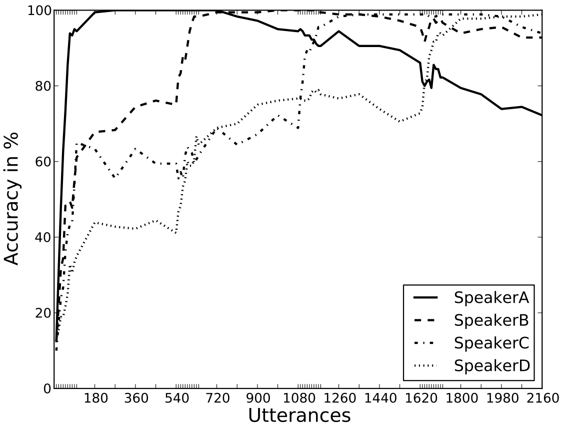

Figure 1: Recognition accuracies in a winner-takes-all coding scheme. Speaker changes occur every 540 utterances, indicated by the increased density of tests after onset of a new speaker.

malisation and the winner-takes-all approach to discretising the model's response, it can be expected that an absolute threshold θ_{abs} does not have an impact when it is lower than chance level. Thus, $\theta_{abs} < \frac{1}{9}$ (for nine keywords) should not have an effect on the number of correctly recognised items.

3. Results

The 'traditional' accuracy results, shown in Fig. 1, show that we replicated previous findings on the model's behaviour. Steep improvements for the speakers currently being trained are observable, where the accuracy levels quickly approach the ceiling. For speakers not yet trained, a beneficiary effect of variation during training can be observed, since all speakers perform above chance level. Performance remains at a high level even after the model adapts to a new speaker.

However, a decay can be seen for the first speaker (speaker A), from 100% at utterance 540 to 72% at the end of the training session. Most probably, previous reports obscured this effect by only reporting accuracy averaged over all speakers. Hence, this finding is independent of the introduction of thresholds.

3.1. Activations and Thresholds

In the previous reports only the maximally activated concept was considered, and if the winning label was equal to the 'true' label, the response was counted as 'correct'. The recognition accuracies at different relative and absolute thresholds for the first and the last speaker trained are depicted in Fig. 2.

The most important effect of setting $\frac{1}{9} < \theta_{abs} < .2$ is that previously misrecognised stimuli are now being classified as 'undecided'. Other than the classification of 'wrong' responses as 'undecided', small values of θ_{abs} have no effect. This can be seen in Fig. 2, where the accuracy plots for $\theta_{abs} = \{.0, .1, .2\}$ overlap. When further increasing the threshold, recognition performance in terms of items recognised correctly and with a 'certainty' exceeding the threshold begins to degrade.

The amount of items classified as 'undecided' is not uniform across the whole time course of training. For the period in which a speaker is trained, $\theta_{abs} = 0$ shows recognition accuracies above 95% already after a small number of training utterances for each keyword. This high level of accuracy is maintained throughout training with a particular speaker. However, when θ_{abs} is increased, accuracy improves more gradually. In the left hand panel of Fig. 2 it can be seen that with $\theta_{abs} = .4$ an accuracy of 100% is only reached after all training utterances of speaker A have been processed. With $\theta_{abs} = .5$ the 100% ceiling is not reached anymore. Contrary to the conclusion in [4] learning does not saturate after a small number of utterances, at least not in the sense that the confidence of the model keeps increasing until the end of training.

When a speaker is not yet trained or the model learns from new speakers, the recognition accuracies that remained fairly high with $\theta_{abs} = .0$ deteriorate much more rapidly with increasing θ_{abs}. For $\theta_{abs} = .5$ a final accuracy of only 18% is obtained. From the dotted lines in Fig. 2 it can also be seen that the beneficial effect for speaker D when the model is trained with speakers A, B, and C rapidly disappears for values of $\theta_{abs} > .2$. With $\theta_{abs} = .5$ the performance of speaker D only rises above chance level after the first training sentences of this speaker have been processed. So here too we see that the interpretation of the performance of the model strongly depends on the criteria used for deciding that a response is correct. Obviously, guessing in low confidence situations is rewarded.

In the right hand panel of Fig. 2 it can be seen that for the relative threshold θ_{rel}, a .1 minimal difference between highest and second highest activated keyword label already affects recognition performance. The curves for $\theta_{rel} = 0$ and $\theta_{rel} = .1$ do not overlap. Furthermore, an increase of θ_{rel} leads to all errors now being marked as 'undecided' in all conditions and for all speakers. Further increasing the relative threshold leads to degraded performance, similarly to the absolute threshold $\theta_{abs} > .2$. The effects described in the previous paragraphs are thus present for the relative threshold, albeit to a greater extent due to its consideration of competitors.

4. Discussion

The present paper addressed the possible difference between a mere winner-takes-all scoring and the use of thresholds when assessing responses in a (modelled) psycholinguistic study. Furthermore, the choice of threshold was examined by investigating the effect that two different types of thresholds at various values had on the measured responses.

To this end, we trained a word-learning model in a multi-speaker condition with a blocked presentation of four speakers to replicate effects of variation of recognition that were found in infant studies reviewed in [2]. Our results in terms of a pure winner-takes-all measurement of performance replicate both previous infant and modelling studies. In addition, we could report a decrease in accuracy for speakers that are not trained any longer.

Using a relative and an absolute threshold for interpreting the model's response, a more differentiated picture emerged. First, the introduction of thresholds led to the discovery that learning is not complete after only a few utterances. $\theta = 0$, the pure winner-takes-all assessment, suggested that only a few training utterances suffice to achieve almost perfect recognition accuracy of all keywords spoken by a particular speaker. However, the use of thresholds (corresponding to the confidence of the model in its decision) revealed that the learner-internal word representations still profit from additional training, as both relative and absolute activations steadily increase during training.

Second, our results challenge previous conclusions on the extent to which the model replicates the finding that infants

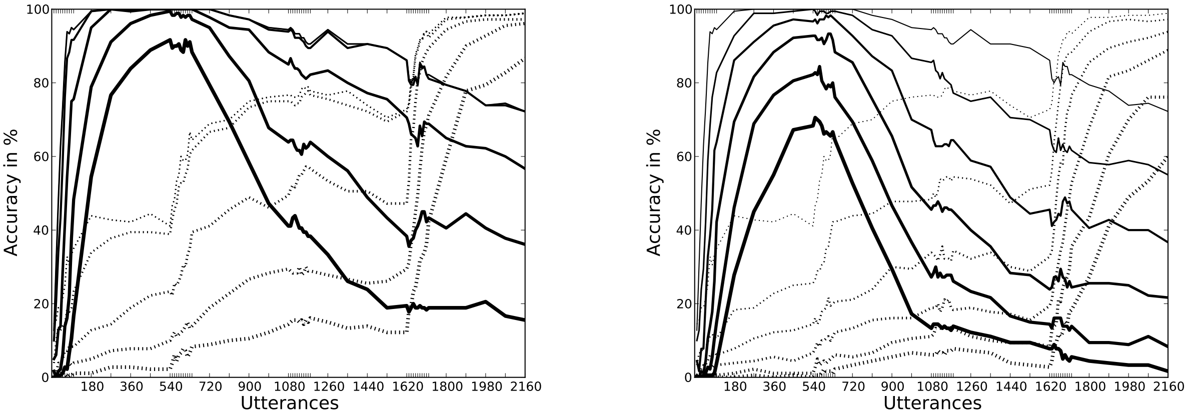

Figure 2: Effect of thresholds on recognition accuracies for the first and the last speaker (speaker A, solid line, and speaker D, dotted line). The left panel shows the effect of θ_{abs}, the right panel depicts increased θ_{rel}. Both θ increase in steps of 10% from 0% to 50%. Since the application of thresholds has a detrimental effect on accuracy, lower lines indicate higher thresholds with the lowest lines in both panels corresponding to $\theta = 50\%$.

profit from variation during learning. Whether or not that effect is replicated depends on the setting of the thresholds. Of course, this finding also raises questions about the criteria and thresholds used in the infant experiments for deciding that a response was right or wrong.

These results show a very different pattern from what has been observed and described in previous reports [4]. In addition, our findings point to the model's ability to make a correct overall decision without being too 'certain' of its response. This is most impressive for the yet unknown speakers and for the early moments of training, where the learner has only observed a few instances of each keyword. Yet, it was possible for the learner to guess the correct keyword label, albeit with little certainty. That his guessing behaviour did not lead to random chance performance shows that the model can generalise, cautiously, from only little experience.

The fact that the two thresholds led to a similar behaviour in the model's performance is related to the way in which the ACORN-NMF model computes activations. As absolute activation of the winner decreases, the activations of the competing keywords become stronger. While we believe that this behaviour is cognitively plausible, other models can be conceived that would concentrate activations in only a few competitors. In those models the effect of absolute and relative thresholds will be different.

When trying to bridge the gap between model output and 'observable behaviours', the present study indicates that the use of fixed thresholds can indeed obscure behaviour. Like in our study, criteria must be set for deciding whether the proportion of the response behaviours that corresponds to the expected behaviour is large enough to consider the response as 'correct'. Here too, using low or high thresholds clearly has a large impact on the proportion of 'correct' responses, and can therefore lead to different conclusions regarding the plausibility of some theory about underlying processes. Moreover, as in our study, using low thresholds may obscure the fact that children are learning over time, perhaps even during the course of an experiment.

To conclude, we show that considering the structure of the responses and exploring different possibilities of grading responses can lead to insights that are not immediately visible in the mere accuracy data, mean values, or the likes. Thus, our results are in line with the studies on data treatment summarised in [9].

5. Acknowledgements

The research of Christina Bergmann is supported by grant number 360-70-350 from the Dutch Science Organisation NWO.

6. References

[1] Scharenborg, O. and Boves, L. "Computational modelling of spoken-word recognition processes: Design choices and evaluation", Pragmatics and Cognition, 18(1): 136 – 164, 2010.

[2] Newman, R.S. "The level of detail in infants' word learning". Current directions in Psychological Science, 17(3): 229 – 232, 2008.

[3] ten Bosch, L., Van hamme, H., Boves, L. "Unsupervised detection of words – questioning the relevance of segmentation", ISCA ITRW, Speech Analysis and Processing for Knowledge Discovery, paper 046, 2008.

[4] ten Bosch, L., Räsänen, O., Driesen, J., Aimetti, G., Altosaar, T., Boves, L., and Corns, A. "Do Multiple Caregivers Speed up Language Acquisition?", Proc. Interspeech 2009, 704 – 707, 2009.

[5] MacWhinney, B. "Computational models of child language learning: an introduction", Journal of Child Language, 37(3): 477 – 485, 2010.

[6] Driesen, J., ten Bosch, L., and Van hamme, H. "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition", Proc. Interspeech 2009, 1731 – 1734, 2009.

[7] Lee, D. and Seung, S. "Learning the parts of object by non-negative matrix factorization", Nature, 40: 788 – 791, 1999.

[8] Van hamme, H. "HAC-models: a novel approach to continuous speech recognition", Proc. Interspeech 2008, 2554 – 2557, 2008.

[9] Forster, K. and Masson, M. "Introduction: Emerging data analysis", Journal of Memory and Language (Special Issue: Emerging Data Analysis), 59(4): 387 – 388, 2008.