Modelling the effect of speaker familiarity and noise on infant word recognition

Christina Bergmann[1,2], *Michele Gubian*[1], *Lou Boves*[1]

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

Abstract

In the present paper we show that a general-purpose word learning model can simulate several important findings from recent experiments in language acquisition. Both the addition of background noise and varying the speaker have been found to influence infants' performance during word recognition experiments. We were able to replicate this behaviour in our artificial word learning agent. We use the results to discuss both advantages and limitations of computational models of language acquisition.

Index Terms: language acquisition, statistical learning, background noise

1. Introduction

Language acquisition, an arguably extremely complex task, is approached by infants with at least some skills and capacities that seem to emerge at a very young age. Some examples are a neural processing system dedicated to language-like acoustic input [1] and an ability to attend specifically to speech (rather than non-speech sounds) [2]. Using these facilities, infants have to detect meaningful patterns in the stream of speech that is often perceived under non-perfect conditions. Most every-day linguistic input can be assumed to occur with at least some background noise, be it non-speech, such as the engine while driving in a car, or speech, such as a television or a parent on the phone.

A few experiments have examined infants' ability to focus on a stream of speech and detect known structure under noisy conditions. A word-segmentation experiment [3] showed that at a signal-to-noise ratio (SNR) of 5 dB children at 7.5 months of age succeed at recognising familiar frequent words such as *dog* within passages against a competing voice, but fail to do so at an SNR of 0 dB, that is when both the target voice and the competing voice are equally loud. Turning to a even more frequent and well-known word, Newman [4] investigated the recognition of the child's own name in multi-talker babble across three age groups. At the age of 5 and 9 months, infants successfully recognised their name at an SNR of 10 dB, but not at a 5 dB SNR. At the lower SNR, a stress-matched foil was confused with the child's name. Only at the age of 13 months, children succeeded in this task at both SNRs and could discriminate between their name and a stress-matched foil.

All these experiments used a well-known word (the child's name or frequent words such as *dog*) with either an unfamiliar speaker or a speaker to which the child was familiarised as part of the experiment. The identity of a speaker, however, has been found to be detected and used by children at the age of 7.5 months [5] to an extent that seems to affect even the encoding of lexical items. The effect of speaker familiarity and a possible tuning in on characteristics of the speaker's voice was subsequently used by Barker and Newman [6] to assess a child's ability to recognise words in multi-talker environments. 7.5 month old infants were able to detect familiarised well-known target words in passages spoken by their mother with a female voice talking in the background at a 10 dB lower intensity. When those passages were uttered by a stranger in similar conditions infants failed to detect the familiarised words.

The results laid out above indicate some ability of infants to segment and comprehend speech in noisy environments. At the same time it is also evident that infants have not yet acquired the specific skills that enable adults to understand speech - even of unfamiliar speakers - in conditions where the SNR is as low as – 5 dB [7]. Comparable research on human *speaker* recognition in noise is sparse, but it has been shown that performance degrades somewhat when the reference speech is recorded over the fixed telephone network, while unknown samples are recorded in a mobile network [8].

In the present paper we employ a general-purpose word-learning model in an attempt to simulate the effects of background noise and speaker identity on word recognition. Previous experiments computationally simulated cross-situational word discovery using statistical information using this general-purpose word-learning model (e.g. [9], [10]). One important result is the successfully replication of infants' ability to associate meaning with words that appear across different situations and within different utterances. Furthermore, the finding that speaker-dependent information seems to be encoded by infants during word learning [5] has also been replicated by this model. More precisely, a training with four speakers in a block-wise fashion led to a moderate improvement in accuracy for yet untrained speakers, whereas training with four intermixed speakers led to a significantly higher improvement of learning across the board [10].

In the research reported in this paper we investigate whether the model can also simulate the results of the speech-in-noise experiments alluded to above. By testing the model under noisy conditions, we intend to gain further insight into the capabilities and limitations of the model. One limitation is immediately evident (and was also an issue in previous experiments): With few exceptions the performance of the model is expressed in terms of the proportion of correctly recognised keywords within test utterances. In order to compare these accuracy measures with behavioural measurements obtained in experiments with infants we need to make the assumption that word recognition accuracy is related to looking times or listening preferences. However, despite these limitations computational model simulations can

provide insights in cognitive processes that cannot be directly observed in infants [11].

Bearing these limitations in mind, we trained our model with clean speech by one female speaker (the *Mother*), which contained keywords in various carrier sentences. Subsequently, we tested the recognition accuracy of those keywords with test sentences from the Mother and another female speaker, the *Stranger*. To draw upon one of the obvious advantages of computational modelling, namely the possibility to investigate numerous variables and their interaction with relative ease, we aim not merely at reproducing the limited amount of known child data. Rather, we vary the SNR in steps of 5 dB between 30 dB (which is effectively clean speech) and 0 dB. We also use two different types of noise, pink noise and background babble stemming from recordings in a cafeteria. All combinations of those three factors, speaker identity, noise type and SNR, are explored to yield a thorough assessment of the model's performance.

2. The model

The ACORNS (ACquisition Of Recognition and communicatioN Skills) project (http://www.acorns-project.org) [10] aimed at investigating language acquisition using computational models. More precisely, to simulate cross-situational word discovery within real acoustic speech utterances paired with keyword-labels, a number of machine learning approaches were used. Importantly, the computational models developed in ACORNS learn from real speech input. No previous lexical, phonetic or phonological information is provided to the learner, nor is information on the number of various items to be learned from the input given beforehand. Therefore, these models offer an excellent starting point for investigating the impact of background noise on the performance of a learner.

In the current study, we use the Non-negative Matrix Factorisation (NMF) [12] implementation of the ACORNS models. This model can replicate the advantage of learning from multiple speakers over learning from a single speaker [10].

Input is presented to the model by pairing an acoustic part with a corresponding keyword label. In NMF, this input is coded as a vector $\mathbf{v} = [\mathbf{v_a v_k}]$. NMF simulates learning by decomposing a high-dimensional input consisting of m vectors (utterances) $\mathbf{v}$ of a total length n (representing the acoustic $\mathbf{v_a}$ and keyword encoding $\mathbf{v_k}$ features of an utterance) into the product of two more compact internal matrices $W \cdot H \approx V$ by minimising the Kullback-Leibler divergence between the input and the dot product of the decomposed matrices. The size of the internal matrices for W is $n \times r$ and for H it is $r \times m$. The constant r is chosen such that $(m + n)r < mn$, i.e. information is compressed. W has the same internal structure as V, namely an acoustic and a 'visual' keyword-encoding part. Hence, it can be assumed to store acoustic information associated to keywords. H contains information about episodic activation of columns in W during training. The particular version of NMF used here, which updates the content of W after each input utterance, has previously been described in [9]. 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 W of the words that are being learned.

To assess the performance of the model during and after training, a new utterance containing a previously learned keyword is given in the form of $\mathbf{v_a}$, without providing the corresponding keyword part $\mathbf{v_k}$. The missing keyword information has to be reconstructed by approximating $\mathbf{v_k} \approx W_k \cdot \hat{h}$ (again by minimising the Kullback-Leibler divergence), where $\hat{h}$ is estimated using the learned representations within W. The reconstructed keyword is compared against the original information given in the test item in order to establish whether the correct keyword was recognised.

3. The effect of noise and speaker identity

3.1. Training and testing

During training, the learner was presented with 500 utterances containing one out of nine keywords within a carrier sentence spoken by a female speaker, the Mother. Each sentence was accompanied by the corresponding keyword in form of a boolean vector.

In the test phase, we aimed at assessing the model's word recognition accuracy for utterances spoken by the Mother or by a new speaker, the Stranger. To this end, we generated two test sets with a similar structure, one for each speaker. Each of those test sets only contained utterances that were not part of the training. 20 test items per keyword were used, resulting in 180 utterances per test set. Accuracy tests were conducted after 20 training steps, that is after 20 new utterances haven been used successively to update the internal representations of the model, up to the point when 200 utterances have been observed. For the remainder of the experiment testing occurred after 50 training steps. These intervals allow for a sufficiently fine-grained assessment during early parts of the learning, where the most drastic changes of recognition accuracy have been found to occur [10]. During testing the learning is disabled. Therefore, the same utterances can be used at each test step, and the test utterances spoken by the Stranger remain equally unfamiliar during the complete experiment. While this is certainly not ecologically plausible, we consider this as an important advantage of computational modelling, because it enables us to make strict comparisons that are impossible in infant experiments.

To assess the recognition ability of the learner in noisy environments, two types of noise, namely pink noise and background babble, with SNRs degrading from 30 dB to 0 dB in steps of 5 dB, were added to the test items. The resulting acoustic signal was then transformed into the fixed-length vector required by NMF. In our implementation, each vector $\mathbf{v_a}$ has length $n_a = 110,002$ and it is based on a Vector Quantization coding of the MFCC vectors derived from an input utterance. This high dimensionality is a consequence of the coding scheme that captures co-occurrence counts of acoustic events at specified time lags [13]. Note that it is not possible to resynthesise the original speech signal from a vector $\mathbf{v_a}$.

The addition of noise to the test items aims at paralleling the infant experiments described in Sec. 1. During the infant experiments, a change in listening behaviour was interpreted as the expression of a preference based on word-recognition. In our model, word recognition is assessed by accuracy scores. Hence, an increase in accuracy implies a higher rate of recognition and should therefore model the cause for the behaviour observed in infant experiments.

3.2. Results

Generally, for both types of noise, the model performed in a comparable manner; hence we only present the results of testing with babble noise. All statements also apply to the pink noise condition, unless noted otherwise. To control for possible idiosyncratic effects of both the training and the test set, we

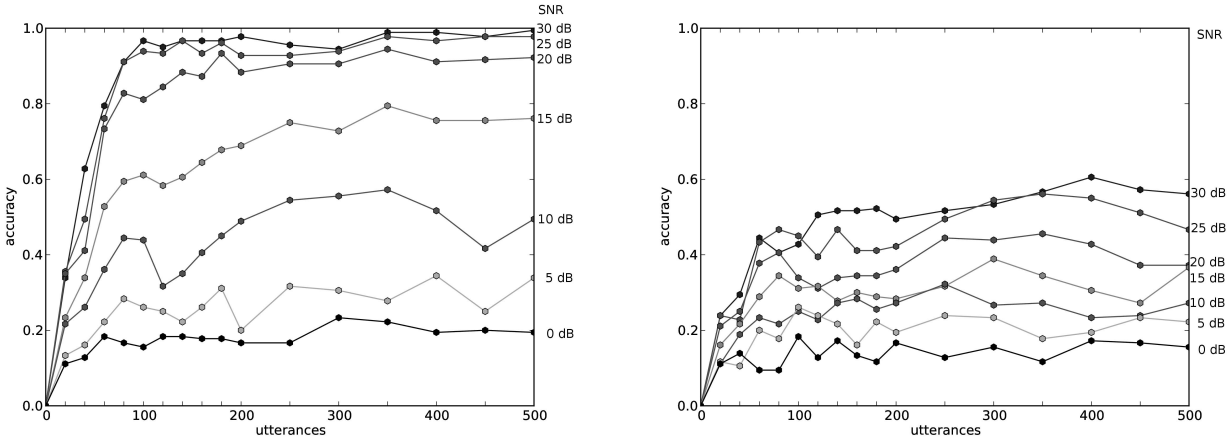

Figure 1: Recognition accuracies in babble noise; left panel shows the familiar speaker, the Mother; right panel the unfamiliar speaker, the Stranger. SNRs are annotated at the final utterance.

ran a simulation with the roles of Mother and Stranger reversed. This procedure yielded comparable results. To assess apparent differences and similarities in the accuracy data we employed the McNemar test [14]. We used the Bonferroni correction to account for multiple comparisons and consequently multiplied each *p*-value by the number of comparisons undertaken. A significance level of $\alpha = 0.01$ was used.

Fig. 1 depicts the accuracy scores using both the Mother (left) and the Stranger (right) test sets for the babble noise across SNRs. The accuracy refers to the percentage of correctly recognised keywords in the 180 test items (20 for each of the nine keywords) in the held-out test set. Each line in the graph represents one SNR condition with the respective SNR value annotated to the right of the panels.

Inspecting the two panels, it stands out that most learning takes place within roughly the first 100 utterances, which corresponds to about eleven training sentences per keyword. The highest overall accuracy within the first 100 training steps is 96%. This occurs in the case where the test items are spoken by the Mother with virtually no added noise. Only comparatively small improvements take place after the first 100 training items with an average increase in accuracy of 4% (SE is 0.75%).

A comparison of the accuracy in word recognition for the Mother's versus the Stranger's test items shows that utterances spoken by the same speaker during both training and testing are much easier to recognise than utterances spoken by an unfamiliar speaker. The highest accuracy for the Stranger is at 60%, whereas the Mother's test sentences lead to a recognition rate of up to 99%. In both conditions, maximal performance occurs when the SNR is at 30 dB, which corresponds to virtually clean speech. For the Mother's test sentences training and testing occur under matched conditions, but using different carrier sentences and different realisations of the keywords.

Adding noise to the test utterances of the Mother or the Stranger leads to a graceful degradation of recognition accuracy. There is a significant difference in accuracy for the Stranger between an SNR of 30 dB and of 25 dB. Decreasing the intensity of the signal by 5 dB leads to a loss of accuracy at this point already. This is not the case when assessing the Mother, where recognition rates at 30 dB and at 25 dB SNR are indistinguishable and at ceiling with up to 99% correctly recognised test items. We thus consider the performance of the model at ceiling when tested with the Mother's speech at an SNR of 25 dB.

According to a McNemar test the accuracy of the Stranger with clean speech (30 dB SNR) is indistinguishable from the recognition performance of the Mother at SNR 15 dB. Thus, the Mother has an advantage of at least 10 dB over the Stranger (relative to the 25 dB SNR for the Mother's speech, which yields accuracy scores equivalent to 30 dB SNR). For lower SNR values the advantage of the Mother over the Stranger decreases, but remains statistically significant. Only at 0 dB, where the performance of both Mother and Stranger is around chance level, the advantage of the Mother disappears.

4. Discussion

In our study, we set out to model the effect of noise on infant word recognition. The results presented above show that the model is sensitive to noise in the test items and that it seems to have tuned in on specific properties of one speaker. This is evident in the overall advantage of the familiar speaker, the Mother, across SNRs (excluding 0 dB). Furthermore, the addition of noise led to a gradual decrease in accuracy for both speakers with chance performance being reached when the signal and the noise are of equal intensity.

When comparing the model's performance to the behaviour of infants in experimental settings, a number of findings laid out in Sec. 1 have been replicated. First, we could show that a known speaker has a general advantage over an unknown speaker. This result is in line with the finding that words spoken by a child's own mother are recognised in a 10 dB SNR condition, whereas a stranger's voice does not elicit a behavioural response under the same noise condition [6]. We could additionally quantify the difference between talkers with respect to our model and found that the Mother has a 10 dB advantage over the Stranger. Second, decreasing the SNR led to a graceful degradation of the model's performance, as opposed to a sudden breakdown of overall performance at a positive SNR. This is in line with infant behaviour, who show a decreasing listening preference with increased noise intensity [3].

Our model failed to improve strongly with a moderate amount of additional training, be it for the known speaker in noisy conditions or for the unfamiliar speaker. Contrastingly,

children's performance improve with older age [4]. Hence, our training data are not suitable for modelling a developmental trajectory comparable to children's increased linguistic skills, even under noisy conditions. At the same time it is fair to say our model learned more in the experiment than do infants in the typical preferential looking experiment.

Two different but possibly interrelated explanations can be used to account for the behaviour of the present model: On one hand, as visible in Fig. 1, performance seems to reach a stable level after about 100 training utterances. This apparent saturation of learning is underlined by the continuously high and stable accuracy scores under matched conditions after the first 100 training steps. Thus, there is no need to drastically change the internal representations of acoustic input during training. It has to be noted that we cannot directly assess the actual form of the internal representations due to the encoding of the acoustic signal in the form of co-occurrence counts of VQ-labels explained in Sec. 3.1. Moreover, without any form of noise compensation this encoding is not likely to be robust against additive noise. Based on the performance of the model, however, we can still assume that the representations are generalised enough to extend to new tokens of a given keyword.

On the other hand, the current model employs a single-level representation to store and recognise acoustic information. There is no hierarchical organisation or multi-level information flow. Hence, neither fully episodic information characteristic for the speech of the Mother nor abstract knowledge about speech and the native language can be used to aid word recognition, especially in adverse conditions.

In children, knowledge about general properties and the structure the native language has been found to surface around the first birthday [4]. At the same time, multi-level representations of linguistic input seem to emerge. This is illustrated by a difference in behaviour when confronted with identical stimuli in two different tasks (e.g. [15]). When children have to merely discriminate a native phonetic contrast in syllable-initial position (e.g. *bin* versus *din*), they show the perceptual abilities to do so. When one of those syllables is taught as a new word, in contrast, children do not notice a switch. This seemingly contradictory behaviour is assumed to originate in several layers of internal representation encoding different levels of acoustic detail. Each task consequently taps into a different level of generalisation.

Our model, unlike children, does not yet develop such a multi-level analysis of acoustic input after sufficient training and maturation. It consequently provides a snapshot rather than a developmental account of child language acquisition. Other current models use multi-level representations, recent examples being PHOCUS and PUDDLE (as reviewed by [16]). However, the multi-level organisation is usually hand-crafted and predefined, instead of being established from the input. Consequently, the models also provide a snapshot, albeit of a later developmental stage. Furthermore, most computational models rely on symbolic input, often in the form of transcribed or otherwise heavily pre-processed speech. Our learning system, in contrast, has to discover meaningful information in the signal as a blank slate and without additional information.

We are exploring several directions for allowing our model to develop hierarchical representations. By doing so, we hope to gain further insight into the very early stages of language acquisition. However, it is not evident how that can be done without wiring at least some aspects of a linguistic theory into the architecture, even if that is something we would want to avoid.

In summary, we have presented how a recent word-learning model can reproduce major aspects of the findings from experiments on infant language acquisition. The behaviour of the model shows analogies to an early phase of word-learning in infants, which is usually not covered by simulations of child language acquisition.

5. Acknowledgements

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

6. References

[1] Dehaene-Lambertz, G., Dehaene, S., and Hertz-Pannier L., "Functional Neuroimaging of Speech Perception in Infants", Science, 298: 2013 – 2015, 2002.

[2] Vouloumanos, A. and Werker, J.F., "Tuned to the signal: the privileged status of speech for young infants", Developmental Science, 7(3): 270 – 276, 2004.

[3] Newman, R.S. and Jusczyk, P.W., "The cocktail party effect in infants", Perception and Psychophysics, 58(8): 1145 – 1156, 1996.

[4] Newman, R.S., "The cocktail party effect in infants revisited: Listening to one's name in noise", Developmental Psychology, 41(2): 352 – 362, 2005.

[5] Houston, D.M. and Jusczyk, P.W., "The role of talker-specic information in word segmentation by infants", Journal of Experimental Psychology, 26(5): 1570 – 1582.

[6] Barker, B.A. and Newman, R.S., "Listen to your mother! The role of talker familiarity in infant streaming", Cognition, 94: B45 – B53, 2004.

[7] Lippmann, R., "Speech recognition by machines and humans", Speech Communication, 22(1): 1 – 15, 1997.

[8] Alexander A., Dessimoz D., Botti F., and Drygajlo A., "Aural and Automatic Forensic Speaker Recognition in Mismatched Conditions", International Journal of Speech, Language and the Law, 12(2): 214 – 234, 2005.

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

[10] 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.

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

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

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

[14] McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages", Psychometrika, 12(2): 153 – 157, 1947.

[15] Altvater-Mackensen, N. and Fikkert, P., "The acquisition of the stop-fricative contrast in perception and production", Lingua, In Press, 2010.

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