The Effects of Background Noise on Native and Non-native Spoken-word Recognition: A Computational Modelling Approach

Themis Karaminis (themis.karaminis@edgehill.ac.uk)
Department of Psychology, Edge Hill University, Ormskirk, UK

Odette Scharenborg (o.scharenborg@let.ru.nl)
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

Abstract

How does the presence of background noise affect the cognitive processes underlying spoken-word recognition? And how do these effects differ in native and non-native language listeners? We addressed these questions using artificial neural-network modelling. We trained a deep auto-encoder architecture on binary phonological and semantic representations of 121 English and Dutch translation equivalents. We also varied exposure to the two languages to generate 'native English' and 'non-native English' trained networks. These networks captured key effects in the performance (accuracy rates and the number of erroneous responses per word stimulus) of English and Dutch listeners in an offline English spoken-word identification experiment (Scharenborg et al., 2017), which considered clean and noisy listening conditions and three intensities of speech-shaped noise, applied word-initially or word-finally. Our simulations suggested that the effects of noise on native and non-native listening are comparable and can be accounted for within the same cognitive architecture for spoken-word recognition.

Keywords: spoken-word recognition; non-native listening; noise; computational modelling; deep neural networks

Introduction

The presence of background noise deteriorates speech perception and this effect is particularly pronounced in non-native listeners (e.g., Cooke, García Lecumberri, & Barker, 2008; Kilman et al., 2014, Scharenborg, Coumans, & van Hout, 2017). Theories of spoken-word recognition often assume that native and non-native listeners are affected by background noise differentially (Cooke et al., 2008, Meador, Flege, & Mackay, 2000). The pronounced difficulties of non-native listeners in noisy listening conditions are thought to reflect the dual challenge of understanding imperfect speech signals with imperfect language knowledge (García Lecumberri, Cooke, & Cutler, 2010). There is a fairly good understanding of the differential effects of noise on native and non-native sound perception (see for a review: Garcia Lecumberri et al., 2010). Far less is known about the effects of noise on spoken-word recognition upstream of sound processing, and how these differ in native and non-native listeners.

A recent study from our lab (Scharenborg et al., 2017) aimed to address this gap. The study employed an offline spoken-word identification experiment, in which English and Dutch students listened to English words in clean listening conditions and with background noise. Noise was applied at the onset or the offset of stimuli words and at different intensities. Scharenborg et al. (2017) found that even though non-native listeners performed overall worse than native listeners, the patterns resulting from the systematic manipulations of the position and intensity of noise were strikingly similar in the two groups. Based on these results, Scharenborg et al. (2017) hypothesised that, in contrast to standard theories of spoken-word recognition (Cooke et al., 2008; Meador et al., 2000), noise has similar effects on native and non-native spoken-word recognition and that the lower overall performance of non-native compared to native listeners is primarily due to differences in exposure, rather than fundamental differences between the native and non-native spoken-word recognition systems.

In this study, we aim to further investigate this hypothesis from a computational modelling perspective. We developed a novel computational model of spoken-word recognition which addresses the effects of noise on native and non-native listenening. Our model, referred to as ListenIN, is based on an autoencoder deep neural network (DNN) architecture for word learning trained on composite representations of words, consisting of simplified representations of phonological forms of words (in line with, e.g., the TRACE model, McClelland & Elman, 1986; see also, Gaskell & Marslen-Wilson, 1997; Smith, Monaghan, & Huettig, 2017) and their meanings (cf. Plunkett et al., 1992). The DNN architecture is cross-linguistically general (cf. Karaminis & Thomas, 2010) and can be exposed to different linguistic environments (e.g., English words only vs. English and Dutch translation equivalents, with greater exposure to the latter) to simulate different types of lexical knowledge ('native English' vs. 'non-native English/native Dutch', correspondingly).

We developed two versions of the model, one corresponding to a native English listener and one corresponding to a non-native English/native Dutch listener. We tested these models on an English spoken-word-identification task, parallel to the human-listener experiment (Scharenborg et al., 2017). Crucially, the two models have emerged from the same initial neural network architecture being exposed to different linguistic environments. To foreshadow our results, using this procedure, we found that the models captured all key patterns in the human data. Our computational simulations demonstrate that, consistent with the hypothesis put forward in Scharenborg et al. (2017), the comparable effects of noise on native and non-native listening can be accounted for within the same cognitive architecture for spoken-word recognition.

The human data

The target empirical data for our model came from an offline English spoken-word identification task, administered to 61 native Dutch and 50 native English students (Scharenborg et al., 2017). The target data included two measurements: overall accuracy in word-identification and the number of different erroneous responses per incorrectly identified word. The first measure provides a general evaluation of offline spoken-word identification performance; the second measure addresses errors in further detail and taps on the notion of the (size of the) 'competitor space' during the activation process in spoken-word recognition (Scharenborg et al., 2017).

The stimuli in Scharenborg et al. (2017) consisted of 126 English words (45 disyllabic, 81 monosyllabic). Each word was presented without added noise (i.e., in the clear), and with stationary speech-shaped noise (SSN) added word-initially or word-finally and at three signal-to-noise ratios (SNRs): 0, −6, and −12 dB. The average number of phones masked by noise was 2.44 ± 0.54 for the word-initial and 2.70 ± 0.94 for the word-final noise condition.

Participants were instructed that they would be listening to English words partially obscured with noise, and were asked to type in the word they thought they heard. Each participant was tested on 168 stimuli, corresponding to 84 words presented with combinations of the three SNRs with the word-initial or the word-final condition, and the same 84 words presented in the clear. Obvious spelling mistakes and homophones were corrected prior to data analysis.

The upper left panel in Figure 1 shows overall accuracy in spoken-word identification in native (continuous lines) and non-native (dashed lines) listeners, in the clear listening condition and the three SNRs, and in conditions of word-initial (thick lines) and word-final (thin lines) noise. With regards to accuracy, the key findings consisted of three main effects and three two-way interactions, which taken together indicated that there were no large differences in the effect of background noise on the processes underlying native and non-native spoken-word recognition. More precisely, the statistical modelling of the accuracy data showed significant main effects of SNR, noise position, and (listener) group. These results suggested that overall accuracy was lower in higher SNRs, in the word-initial than the word-final masking condition, and in non-native compared to native listening. Two-way interactions between the position of noise and group, the position of noise and SNR, and SNR and group suggested that the detrimental effects of word-initial relative to word-final noise were more pronounced in native listeners and in higher SNRs, and that accuracy decreased with SNR more rapidly in non-native listeners. However, as these interactions were ordinal (i.e., the lines did not cross), the combined results implied that the effects of noise on accuracy were not drastically different in native and non-native listeners (Scharenborg et al., 2017).

For a complementary account of this result, the reader may inspect the upper-left plot of Figure 1. Two forceps-like patterns, consisting of a thick and a thin line, correspond to the effects of noise in each group. Crucially, the most

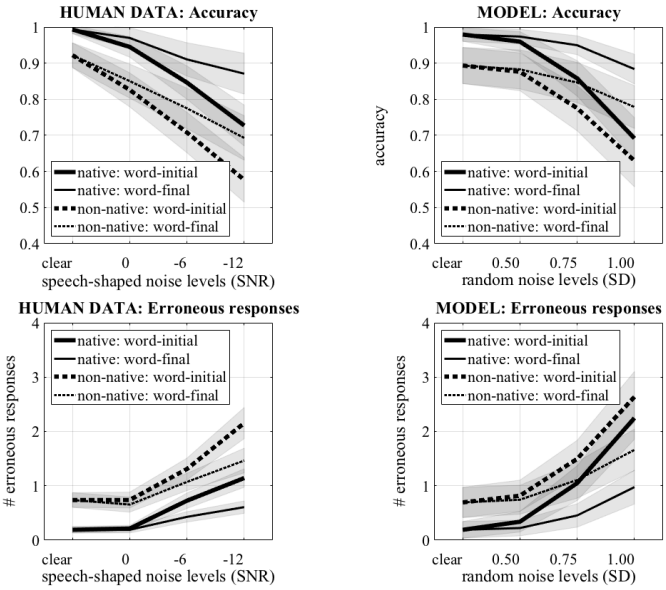

Figure 1: Empirical data (left; from Scharenborg et al. 2017) and simulation results (right) on accuracy rates (top) and the number of erroneous responses (bottom).

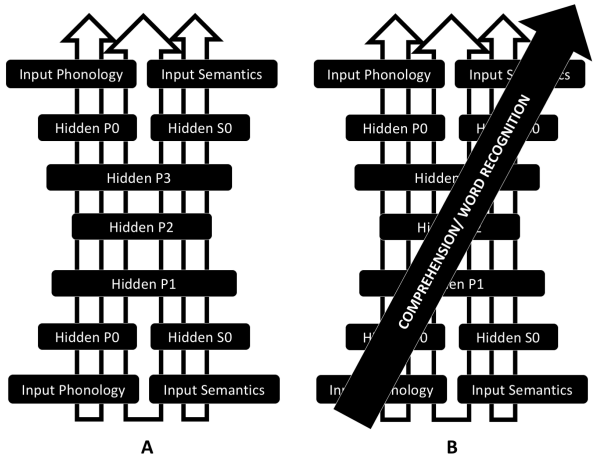

Figure 2: Architecture, training, and testing of the neural network model. A: training on three types of mapping (indicated by the three vertical arrows); B: testing on the novel task of comprehension/word identification (indicated by the diagonal arrow).

pronounced difference between native and non-native listening is the downward shift of the forceps-like pattern, rather than differences in shape, slopes, and direction.

Turning to the number of alternative responses (lower left panel in Figure 1), the key, statistically significant findings were the main effects of group, SNR, and noise position (Scharenborg et al., 2017), showing again no large differences in the effects of background noise on native and non-native spoken-word recognition. Thus, not only did non-native listeners make more errors, they also made more variable errors. The number of alternative responses increased with SNR and was higher in the word-initial than

the word-final condition. The statistical modelling also included the effects of word frequency (Scharenborg et al., 2017) and showed that the number of alternative responses was higher for lower-frequency words (Scharenborg et al., 2017). Figure 1 again shows that the effects of the systematic manipulations of noise manifest as forceps-like patterns, which are highly similar for native and non-native listeners (group differences accounted for by a vertical shift, rather than changes in shape/slope/direction).

Computational modelling

Experimental set-up

Our research design involved a training and a test phase. During training, we developed two versions of the deep autoencoder neural network shown in Figure 2: one exhibiting robust knowledge of a monolingual vocabulary of English words (i.e., the 'native' listener); and one exhibiting knowledge of a bilingual vocabulary of English and Dutch translation equivalents, however, with less robust knowledge of English compared to Dutch words due to greater exposure to the latter (the 'non-native' listener) (cf. Filippi, Karaminis, & Thomas, 2014). The models were trained on composite word representations, which consisted of phonological forms and their meanings. The models were exposed to three types of input-output mapping, illustrated with vertical arrows in Figure 2A: between phonological forms; between meanings; and between combinations of phonological forms and their meanings (see Plunkett et al., 1992).

At test, the models were assessed on their abilities for comprehension/spoken-word identification − that is, when presented with a phonological form only (no semantics input), they should generate the appropriate meaning in the output layer (diagonal arrow in Figure 2B). The neural networks received no explicit (supervised) training on these mappings. Comprehension/word-identification was thus a novel task for the model and correct word identification implied that it had learned to auto-associate phonological forms with their meanings.

We tested the two versions of the trained neural network on a simulated English word-identification task, parallel to Scharenborg et al. (2017). In this task, the phonological representations presented in the input layer included word-initial or word-final noise at three different intensities, as in Scharenborg et al. (2017). For each network, we obtained measures of accuracy in spoken-word identification and of the number of erroneous responses per incorrect response in the two conditions for the position of noise and in different noise intensities. We analysed the model-based measures with statistical modelling procedures similar to Scharenborg et al. (2017), and compared our results to the empirical study.

We note that we also performed control simulations in which we simulated non-native listening by testing 'native English/non-native Dutch' models, that is, models trained on English and Dutch interleaved with a 3 : 1 ratio. These simulations (not reported here) showed parallel results to those reported in this paper, and allowed us to establish that the differences in performance between native and non-native models do not depend on whether the 'native' version is exposed to monolingual or bilingual linguistic environments.

Architecture and representations

The deep autoencoder neural network (Figure 2) used in the simulations comprised a phonological and a semantics pathway and had a symmetric structure, horizontally and vertically. The input and the output layer consisted of 292 units representing phonology and 300 units representing semantics. The architecture had five hidden layers. The first hidden layer was bipartite and consisted of two banks of 150 units, one fully connected to input phonology and another fully connected to input semantics. The second hidden layer (200 units) was composite, that is, it was fully connected to all the units of the first layer. The third hidden layer had 150 units. As the autoencoder was symmetric, the fourth hidden layer was identical to the second hidden layer; and the fifth hidden layer was identical to the first hidden layer.

The phonological form of words was represented using a feature-based representational scheme. Our scheme encoded 51 distinct phones, 20 vowels and 31 consonants, using 22 articulatory/phonological features: consonant, vowel, obstruent, sonorant, aspirated, voiced, plosive, continuant, nasal, lateral, rhotic, strident, labial, coronal, dorsal, glottal, distributed, high, mid, low, retracted, and long (cf. Karaminis & Thomas, 2010).

Phonological forms of words were fitted to a 13-slot disyllabic template: CCCVVCCCVVCCC, where C denotes a consonant and V denotes a vowel. We used alignment to the left (similar to Shook & Marian, 2013) so as to incorporate in our model the incremental nature of speech processing. The phonological representations also included prosodic information, namely syllabic length and syllabic stress. Syllabic length was represented with thermometer encoding over 2 bits (monosyllabic = 01; disyllabic = 11), syllabic stress was represented with 2 one-hot bits. Finally, language information (2 one-hot units, English or Dutch) was also included in the phonological representations. This information is useful for the modelling of production (producing phonological forms given meanings) in future extensions of this model but was not used in the current simulations. In sum, the phonological forms of words are represented in a distributed manner over a 292-bit vector with an average of 25.94 ± 6.31 'ones' per word.

Word semantics was represented with a binary scheme based on a 300-dimensional word-embedding model trained on a corpus of 100 billion words from Google News (Mikolov et al. 2012; model retrieved from the gensim Python library, Řehůřek & Sojka, 2010). The real-numbered word-embeddings were transformed to binary representations by setting all values lower than -0.175 to 1, and all other values to 0. As a result of this transformation, word semantics were represented with an average of 52.73 ± 11.48 'ones' over a 300-bit vector.

Training

The training set of the model consisted of 121 English words, and their 121 Dutch translations, taken from Scharenborg et al. (2017) (excluding five word pairs, where the Dutch translation equivalent was trisyllabic). English and Dutch translation equivalents had different phonological forms, and exactly the same semantics.

Training the deep autoencoder combined three techniques, namely weight initialisation with pretraining, weight fine-tuning with three phases of training, and denoising. We included these techniques in the design of the model based on pilot simulations (not reported here), which suggested that this combination enabled networks to auto-associate phonology and semantics, and show abilities for word comprehension and word production.

Weight initialization with pretraining. The weights of the deep autoencoder were initialised using a pretraining method (Hinton & Salakhutdinov, 2006). Weights between the individual layers of the deep autoencoder (say between 'Input Phonology' and 'Hidden P0') were trained separately, within shallow (one hidden layer) autoencoders, considered specifically for the pretraining phase. For a demonstration of this method, see Hinton and Salakhutdinov (2006). Pretraining was implemented in MatLab with the neural network toolbox (Release 2016a, The Mathworks, 2016), using the Scaled Conjugate Gradient Algorithm (Moller, 1993), with sparsity regularisation (Olshausen & Field, 1997) and the following parameters: 2000 epochs, L2 WeightRegularization = 0.01, SparsityProportion = 0.10.

Weight fine-tuning with three phases of training. After the weights of the deep network were initialised with pretraining, they were fine-tuned (within the deep network). Weight fine-tuning lasted for 1000 epochs and used the back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986) with the cross-entropy error criterion (Hinton, 1992) (lr = 0.05, momentum = 0.0). Similar to Plunkett et al. (1992), the following three phases are considered: (A) The phonology-to-phonology phase, in which only input phonology was presented to the network, which was trained on producing the same phonological pattern in the output layer. No changes happened in the semantics pathway during this phase; (B) The semantics-to-semantics phase, which focused learning on the semantics side of the network; (C) The mixed phase, in which both phonology and semantics were presented in the input layer and learning happened across the whole network.

Each epoch included 242 (2*121) training sweeps. In each sweep, a word was chosen probabilistically and was presented to the network. The probabilistic training regime was designed to implement either a 'native English' condition, in which networks were presented with English words only, or a 'non-native English/native Dutch' condition, in which networks were presented to English and Dutch words with a 1:3 ratio. For a given word, one of the following two sequences of the learning phases (chosen randomly) was applied: (A)−(C)−(A) or (B)−(C)−(B). The weight changes estimated by the learning algorithm were also multiplied by the log-transformed CELEX word frequencies (cf. Baayen, Piepenbrock, & Gulikers, 1995).

Denoising. During the fine-tuning phase, random noise was also injected in the input layer of the network ('denoising autoencoder'; see Vincent, Larochelle, Bengio, & Manzagol, 2008; Zur, Jiang, Pesce, & Drukker, 2009). Noise was injected probabilistically, in half of the training sweeps and to the phonological part of the input or the semantics part or both. In all cases, the network should output denoised activations. The random noise added to a given input unit's activation had a zero mean and SD equal to 7 times the SD of the unit's activation across the representations of the training set. The injected noise was therefore very high, often distorting the input patterns.

Simulated spoken-word identification task

In the simulated word-identification task, we presented the native and non-native versions of the models with phonological representations of words (and no semantics) at the input and evaluated whether they produced appropriate semantics in the output layer. The output was evaluated with a nearest neighbor criterion (output semantics should be closer to the appropriate semantics than any other semantics pattern). Similar to the human-listener experiment (Scharenborg et al., 2017), we included a zero-noise condition and three increasing levels of added noise. The added noise was implemented as a real-numbered vector selected from a Gaussian distribution with mean = 0 ('clear') and SD = .50 ('low intensity'), .75 ('medium intensity'), or 1.00 ('high intensity'). We also considered conditions of word-initial and word-final noise, by adding noise only to either the leftmost or the rightmost slots (correspondingly) of the phonological template. For individual words, word-initial and word-final noise was applied to the same phones as in Scharenborg et al. (2017).

We performed two sets of simulations, with 10 replications (random seeds for weight initialisation) each. In the first set, we tested English word identification in models trained only on English words (i.e., 'native listeners'). In the second set, we tested English word identification in the model trained on English and Dutch translation equivalents interleaved with a 1:3 ratio (i.e., 'non-native listeners').

Results

The right panels of Figure 1 present the simulation results on the accuracy rates (top) and the number of erroneous responses for incorrect responses (bottom). Visual inspection of the combined results in Figure 1 suggests that the modelling results were highly similar to the human data (left panels) on both measures. The modelling output replicated the forceps-like patterns for the effects of the systematic manipulations of noise on the two measures of spoken-word identification in each group. And again, the effect of background noise was highly similar in the 'native' and 'non-native' models: group differences were accounted for by a

vertical shift of the forceps-like performance patterns. However, the model presented bigger differences in the listening performance in the word-initial and the word-final condition (forceps-like performance patterns are more open and overlap in the right panels).

The statistical analysis of the models' accuracies showed main effects of group (β = -.348, SE = .144, p = .016), noise intensity (β = 1.138, SE = .089, p < .001), and position (β = -1.734, SE = .136, p < .001), as well as two-way interactions between noise intensity and group (β = -.501, SE = .071, p < .001), position and noise intensity (β = .254, SE = .101, p < .001), and position and group (β = .393, SE = .143, p < .001). The models thus captured all six key findings of the human accuracy data (Scharenborg et al., 2017). Moreover, as interactions were ordinal: the combined results implied that the effects of noise on accuracy were not drastically different in the native and non-native models. This is also in line with the human data (Scharenborg et al., 2017).

Turning to the number of erroneous responses, our statistical analysis, which accounted for error rates and word frequency (cf. Scharenborg et al., 2017), showed main effects of group (β = .515, SE = .088, p < .001), noise intensity (β = -1.310, SE = .135, p < .001), and position (β = -.528, SE = .094, p < .001), and a non-significant trend for an effect of word frequency (β = -.0854, SE = .051, p = 0.092). Importantly, there were no significant interactions between these factors and group. The model therefore captured the key findings in Scharenborg et al. (2017), including the similar effects of noise on the number of erroneous responses in native and non-native listening.

Discussion and Conclusions

In this paper, we present a computational model which is novel in addressing the effects of noise on the cognitive processes underlying spoken-word recognition in native and non-native listening. A key assumption of the model, grounded on the hypothesis put forward in Scharenborg et al. (2017), is that a common cognitive architecture underlies native and non-native spoken-word recognition. Based on this common architecture, which is differentiated only through its exposure to two different training sets, our model emulates human native and non-native lexical knowledge and captures key effects in the human data from Scharenborg et al. (2017). Indeed, 'non-native' models presented overall lower accuracy rates and overall higher numbers of erroneous responses than 'native' models. Yet, and consistent with the human data (Scharenborg et al., 2017), the effects of the systematic manipulations of noise on the two measures of spoken word-recognition were similar in the 'native' and the 'non-native' models (vertical shifts in Figure 1).

Our results support the idea put forward in Scharenborg et al. (2017) that noise affects the cognitive processes underlying native and non-native spoken-word recognition in a similar fashion, and that performance differences between native and non-native listeners when listening in noise are mostly due to differences in exposure to the non-native language. An important question is whether this hypothesis also applies to online spoken-word comprehension, which involves subtler temporal dynamics than offline spoken-word recognition. Our current work focuses on extending our model, to address human data on online native and non-native spoken-word recognition in noise (by presenting phonological input incrementally, see Gaskell & Marslen-Wilson, 1992) (Hintz & Scharenborg, 2016).

Our model captures the empirical data by bringing together principles and assumptions of earlier models of the bilingual lexicon and speech recognition, for example, feature-based, spatially encoded, representations of phonology with alignment to the left (Shook & Marian, 2013; Smith et al., 2017), a simplified scheme for the representation of word meanings (Smith et al., 2017), bilingual linguistic environments consisting of translation equivalents (Filippi et al., 2013; Shook & Marian, 2013, though see Zhao & Li, 2010), and (semi-supervised) autoassociation of phonological and semantic representations (cf. Plunkett et al., 1992).

Our model is not without shortcomings. Firstly, the feature- and slot-based representations of phonology and the implementation of speech-shaped noise as a random vector added to the binary phonological patterns are simplifying assumptions, which overlook key characteristics of speech. Future versions should consider input representations that are closer to speech signals (cf. Norris & McQueen, 2008; Ten Bosch, Boves, & Ernestus, 2015; Scharenborg, 2010), and which include more realistic implementations of noise. Future versions of the model could also employ recurrent DNN architectures. These are naturally suited to address the incremental nature of speech input, while they have supported some important recent advances in automatic speech recognition (Yu & Deng, 2016). Finally, non-native language learning involves considerable individual variability, in the mode, the timing, the exposure to the two languages, and in language typology. This variability will also be investigated in future versions of the model.

In conclusion, in this paper, we present a computational model, which is novel for addressing the effects of noise on spoken-word recognition in native and non-native listening. The model successfully simulated human performance in a spoken-word identification task, which was administered to native and non-native listeners and included elaborate manipulations of listening conditions (Scharenborg et al., 2017). The model's success in capturing the human data supports a unified account of spoken-word recognition in noise in native and non-native listening using a neural-network-modelling framework.

Acknowledgements

This research was funded by a Vidi-grant from the Netherlands Organization for Scientific Research (NWO; grant number: 276--89--003) awarded to OS. We would like to thank Jiska Koemans for her help with the phonetic transcriptions. Thanks also to Aditi Lahiri, Caroline Floccia, Kim Plunkett and Jeremy Goslin for their contributions in an earlier version of this model.

References

Baayen, R.H., Piepenbrock R., & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Cooke, M., García Lecumberri, M. L. & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *Journal of the Acoustical Society of America, 123,* 414-427.

Filippi, R., Karaminis T., & Thomas, M. S. C. (2014). Language switching in bilingual production: empirical data and a computational model. *Bilingualism: Language and Cognition, 2,* 294–315.

García Lecumberri, M.L., Cooke, M. and Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication, 52,* 864–886.

Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes,* 12, 613—656.

Hinton, G.E. (1989). Connectionist learning procedures. *Artificial Intelligence,* 40, 185–234.

Hinton, G.E. & Salakhutdinov, R.P. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science,* 313, 504–507.

Hintz, F. & Scharenborg, O. (2016). The effect of background noise on the activation of phonological and semantic information during spoken-word recognition. In *Proceedings of Interspeech 2016,* San Francisco, CA, pp. 2816–20.

Karaminis, T.N. & Thomas, M.S.C. (2010). A cross-linguistic model of the acquisition of inflectional morphology in English and Modern Greek. In S. Ohlsson and R. Catrambone, (Eds.). *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Cognitive Science Society, Austin, TX, pp. 730–735.

Kilman, L., Zekveld, A.A., Hällgren, M. & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology, 5*, 651.

The MathWorks Inc. (2016). MATLAB and Neural Network Toolbox Release 2016a. Natick, Massachusetts: MA.

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

Meador, D., Flege, J. E., &, Mackay, I.R.A (2000). Factors affecting the recognition of words in a second language, *Bilingualism: Language & Cognition, 3,* 5–56.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2012). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems,* 3111–19.

Moller, M.F. (1993). A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning, *Neural Networks,* 6, 525–533.

Norris, D. & McQueen, J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition, *Psychological Review, 115,* 357–395.

Olshausen, B.A., & Field, D.J. (1997). Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1, *Vision Research,* 37, 3311–3325.

Plunkett, K., Sinha, C., Møller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4,* 293–312.

Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, Valletta: University of Malta, pp. 46-50.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1: Foundations),* Cambridge, MA: The MIT Press, pp. 318–362.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition, *Journal of the Acoustical Society of America,* vol. 127, no. 6, pp. 3758–70.

Scharenborg, O., Coumans, J., & van Hout, R. (2017). The effect of background noise on the word activation process in non-native spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2017. doi:10.1037/xlm0000441.

Shook, A., & Marian, V. (2013). The Bilingual Language Interaction Network for Comprehension of Speech. *Bilingualism: Language and Cognition, 16,* 304–324.

Smith, A., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93,* 276–303.

Ten Bosch, L., Boves, L., & Ernestus, M. (2015). DIANA, an end-to-end computational model of human word comprehension. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, & J. Scobbie (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015),* Glasgow: University of Glasgow.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML-08)*, pp. 1096–1103.

Yu, D. & Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach.* Springer.

Zhao X., & P. Li (2010). Bilingual lexical interactions in an unsupervised neural network model, *International Journal of Bilingual Education and Bilingualism, 13,* 505-24.

Zur, R.M., Jiang, Y., Pesce, L.L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics, 36*, 4810–18.