Interspeech 2018
2-6 September 2018, Hyderabad

Wide Learning for Auditory Comprehension

Elnaz Shafaei-Bajestan[1], *R. Harald Baayen*[1]

[1]Quantitative Linguistics, Eberhard Karls Universität Tübingen, Tübingen, Germany

elnaz.shafaei-bajestan@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

Abstract

Classical linguistic, cognitive, and engineering models for speech recognition and human auditory comprehension posit representations for sounds and words that mediate between the acoustic signal and interpretation. Recent advances in automatic speech recognition have shown, using deep learning, that state-of-the-art performance is obtained without such units. We present a cognitive model of auditory comprehension based on wide rather than deep learning that was trained on 20 to 80 hours of TV news broadcasts. Just as deep network models, our model is an end-to-end system that does not make use of phonemes and phonological word form representations. Nevertheless, it performs well on the difficult task of single word identification (model accuracy 11.37%, Mozilla DeepSpeech: 4.45%). The architecture of the model is a simple two-layered wide neural network with weighted connections between acoustic frequency band features as inputs and lexical outcomes (pointers to semantic vectors) as outputs. Model performance shows hardly any degradation when trained on speech in noise rather than on clean speech. Performance was further enhanced by adding a second network to a standard wide network. The present word recognition module is designed to become part of a larger system modeling the comprehension of running speech.

Index Terms: naive discriminative learning, auditory word recognition, wide neural networks, deep speech, frequency band summary features

1. Introduction

The question of how we understand speech is under investigation in many disciplines, ranging from linguistics, cognitive science and neuroscience, to natural language engineering [1]. Almost all current linguistic theories assume speech recognition is a two-stage process, with an initial stage at which the acoustic signal is mapped onto a sequence of phonemes, and a subsequent stage at which the stream of phonemes is segmented into a sequence of words. Accordingly, a substantial body of research has focused on linking properties of the acoustic signal to linguistic units such as phonemes and phonological word form representations [2, 3], and cognitive architectures have been put forward that specify how these representations are accessed [4]. Classical automatic speech recognition (ASR) systems build on hidden Markov models (HMMs) in which phonemes again play a pivotal role [5]. However, deep learning has enabled considerable progress, replacing hand-engineered processing with end-to-end approaches that directly learn from data. “Deep Speech” is an example of a state-of-the-art ASR system based on end-to-end deep learning that does not depend on the concept of a “phoneme” as theoretical construct or computational unit [6].

The present study is a progress report on a linguistic approach to auditory comprehension that, like deep learning, rejects the phoneme as a pivotal unit for language comprehension, reflecting the discomfort that also exists within the linguistics community about the validity and usefulness of the phoneme as a theoretical construct [7, 8]. Unlike deep learning, we make use of wide learning, in combination with substantial investment in the development of linguistically and cognitively well-motivated input features. The general framework within which this development takes place is that of naive discriminative learning (NDL) [9, 10]. NDL implements error-driven learning based on the learning rule proposed by Rescorla and Wagner [11], which has a strong history in the field of animal learning and more recently also human learning [12, 13].

The network architecture used for the standard NDL model is a simple two layer network where the weights on connections from input units (henceforth, cues) to output units (henceforth, outcomes) are gradually updated based on the Rescorla-Wagner learning rule (for more details, see section 2.3). The aim of NDL is to build end-to-end models, with in the case of auditory comprehension low-level form features as cues and semantic units as outcomes. Importantly, the standard implementation of NDL (available as an R package [14] and a python library [15]) does not make use of any hidden layers, and hence explores to what extent it is possible, given well-chosen acoustic features, to discriminate between lexical meanings using simply a linear network. However, Sering et al. [16] proposed an extension of the NDL architecture with a second two-layer network, that is trained independently of the first, that further enhances classification performance.

NDL has been successfully employed in modeling the data from a range of experimental studies, showing promising results in explaining human language processing [9], [17], [18] as well as lexical learning in animals [19]. For human auditory comprehension, Arnold et al. [20] developed an NDL-based model of single word recognition, and applied it successfully to spontaneous conversational German speech. A comparison of model performance on lexical discrimination with human performance on the same speech tokens revealed that model accuracy was within the human range.

The current study builds upon this model, and tests it on more and different kinds of speech data, and at the same time also explores whether the second network proposed by Sering et al. indeed improves classification accuracy. Results for single word recognition are compared to that of Mozilla DeepSpeech [21]. A further contribution of the present study is a method for distinguishing between relatively clean speech and speech in noise in the input corpus which comprises TV broadcast videos recorded in studio or outdoors, along with music and background noise.

2. Materials and Methods

2.1. Material

The data resource employed for this study is a subset of the big data from the Distributed Little Red Hen Lab, a vast repository of multi-modal TV news broadcasts. We used 500 audio files containing 266 hours of national and cable broadcasts

10.21437/Interspeech.2018-2420

from the United States in English, recorded in 2016, which were accompanied by high quality transcripts and had been aligned successfully for more than 97% of their words by the Gentle forced aligner [22]. The advantage of working with a huge archive such as the Red Hen data is that a substantial amount of speech is recorded in noisy conditions. The archive not only contains relatively clean speech recorded in a studio, but also speech recorded when reporters are outside, in which case considerable background noise can be present. Furthermore, even recordings made in the studio often carry not only speech, but music playing in the background as well.

We developed an algorithm to automatically distinguish between relatively clean parts where there is speech without background noise or music, from noisy data. To do so, a threshold of 350 in a CD quality (44,100 Hz sampling frequency and 16 bit resolution) pulse-code modulation (PCM) encoded speech stream, was defined to mark the level of amplitudes close to zero. This threshold ($\approx 3\%$ of the peak amplitude) was chosen to capture pauses during speech and the short periods of silence during the closure of plosives. Background noise typically results in such short periods of silence being absent. Sliding a non-overlapping time window of 30 s over the audio files, we traced the number of pause markers which are completely contained in the 30-second window. All speech chunks having more than 40 paus markers were considered *clean*. As a result, a total of 5924 audio files, each with a duration of 30 s, was selected to represent almost 50 h of clean speech. A subset of 970 randomly selected files were manually evaluated by an American English native speaker. The proportion of files for which no background noise could be detected anywhere for the full 30 seconds was 0.35. Thus, the *clean* dataset comprises both truly clean speech files, and speech files with mild background noise.

A *noisy* subset was also compiled using the "sound to textgrid analysis (silences)" from Praat [23] with a silence threshold of -26 dB and a minimum silence interval duration of 0.09 s. Audio chunks with speech, as opposed to those tagged as *silence*, with a duration of at least 5.6 s were included as representing noisy speech, the noise being either outside noise or music playing in the background. In this way, 19,602 audio files of varying durations were obtained, to a total of 80 h of speech. A random subset of 2000 noisy files was evaluated by the same native-speaker, who reported that 91.9% of the files are indeed noisy speech and music snippets.

From the clean and noisy data sets with 50 and 80 hours of speech respectively, henceforth clean-50 and noisy-80, we randomly sampled subsets of 20 and 50 hours of speech (clean-20, noisy-20, and noisy-50), in order to enable comparison with the original results of Arnold et al., which were based on 20 hours of speech, and to provide insight into how the classification algorithm performs as the amount of speech is increased.

2.2. Acoustic features

The acoustic features that served as input cues for the NDL network were the Frequency Band Summary (FBS) features developed by Arnold et al. [20]. FBS features are derived as follows. Given the audio signal for a word, minima of the Hilbert amplitude envelope of the speech wave are used to segment the speech into chunks of varying duration. When no clear minima are present, the signal contributes one chunk. Next, each chunk is evaluated on 21 MEL spectrum frequency bands, which are motivated by the different receptive areas of the cochlea that are known to be responsive to variation in different frequency ranges in acoustic signals [24]. For each chunk, and each of the 21 frequency bands of these chunks, an FBS feature brings together band number, chunk number, and a summary of the temporal variation in the band by means of the median, minimum, maximum, initial, and final intensities of the values in the band. We used the `AcousticNDLCodeR` R package [25] to extract the FBS features from our speech files.

2.3. The NDL classifier

Consider a set of cues C with m unique members c_i ($i = 1, \cdots, m$) and a set of lexical outcomes O with n unique members o_j ($j = 1, \cdots, n$). C and O elements occur with repetition in a pair called a learning event. A sequence of learning events E_{train} of length r compose the training data. The NDL network is defined by an $m \times n$ matrix $\boldsymbol{W}$ of connection weights w_{ij}, where w_{ij} is the association strength from c_i to o_j. The connections weights in W are initialized with zeros; $w_{ij}^{(t=0)} = 0$ ($i = 1, \cdots, m; j = 1, \cdots, n$). During learning, events are visited one at a time. At time t ($t = 1, \cdots, r$), the learning event e at t comprises a set of active cues C_t ($C_t \subseteq C$) and a set of observed outcomes O_t ($O_t \subseteq O$) that drive the updating of the weights of $\boldsymbol{W}$. Denoting the weight from c_i to o_j at time t by $w_{ij}^{(t)}$, the update in weights from c_i to o_j at time t is given by $\Delta w_{ij}^{(t)}$, i.e.,

$$w_{ij}^{(t)} = w_{ij}^{(t-1)} + \Delta w_{ij}^{(t)}. \tag{1}$$

The update $\Delta w_{ij}^{(t)}$ itself is given by the Rescorla-Wagner learning rule:

$$\Delta w_{ij}^{(t)} = \begin{cases} 0 & \text{if } c_i \notin C_t, \\ \alpha_i \beta_j (\lambda - \sum\limits_{c_k \in C_t} w_{kj}^{(t-1)}) & \text{if } c_i \in C_t \wedge o_j \in O_t, \\ \alpha_i \beta_j (0 - \sum\limits_{c_k \in C_t} w_{kj}^{(t-1)}) & \text{if } c_i \in C_t \wedge o_j \notin O_t. \end{cases} \tag{2}$$

The parameters of the Rescorla-Wagner learning rule were set to $\lambda = 1.0$, $\alpha_i = 1.0$ and $\beta_j = 0.001$ for all i, j, following earlier modeling studies with NDL, and never changed in the course of the present study.

Given $\boldsymbol{W}$ and active cues at learning event e_i, $c_k \in C_i$, the support a_{ij} (henceforth, activation) of these cues for a given outcome o_j is given by the sum of the weights from the active cues to this outcome:

$$a_{ij} = \sum_{c_k \in C_i} w_{kj}. \tag{3}$$

More generally, given an $r \times m$ matrix $\boldsymbol{C}$ of learning events by cues that is zero but one for those cues that are present at a given learning event, we have that

$$\mathbf{A} = \mathbf{CW}. \tag{4}$$

The activation matrix $\boldsymbol{A}$ provides, for each word presented to it, the network's support for all possible outcomes. To assess network performance, the lexical outcome with the highest activation is selected and compared with the targeted outcome. Alternatively, the number of targeted outcomes among the top n most highly activated outcomes can be considered.

Following Sering et al. [16], a second network was stacked on top of the first one. This second network is defined by an $n \times n$ decision matrix $\boldsymbol{D}$ which linearly seeks to rotate the activation matrix $\boldsymbol{A}$ of the first network onto an $r \times n$ target matrix $\boldsymbol{T}$ specifying for each learning event whether a given outcome is present (1) or absent (0), i.e.,

$$\mathbf{AD} = \mathbf{T}. \tag{5}$$

Therefore, $\boldsymbol{D}$ can be estimated by solving

$$\hat{\mathbf{D}} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}\mathbf{A}^t\mathbf{T}, \tag{6}$$

resulting in a matrix of predicted outcome strengths $\hat{\boldsymbol{T}}$,

$$\hat{\mathbf{T}} = \mathbf{A}\hat{\mathbf{D}}. \tag{7}$$

As for the standard NDL network, model performance is based on whether the most activated outcome in the relevant column of $\hat{\mathbf{T}}$ is identical to the targeted outcome. Here too, evaluation can be extended to include targeted outcomes among the top n best supported outcomes.

For each of the five datasets introduced above, 10-fold cross validation was applied, resulting in a total of 50 models for the standard NDL model (using $\boldsymbol{A}$) and a second set of 50 models for the extended NDL model, henceforth NDL+ (using $\hat{\mathbf{T}}$). All models are trained and tested on single word tokens (as given by the word boundaries provided by the aligner) with FBS Features of the audio file as cues and orthographic form of the word types as identifiers for lexical outcomes. Out-of-vocabulary word types were discounted when computing accuracy. From the clean-50 corpus, 72,711 FBS features and 15,698 lexomes were extracted from 401,015 word tokens. The noisy-80 corpus contained 66,106 FBS features, 13,523 lexomes, and 289,245 word tokens. The `ndl2` (version 0.1.0.9002) R package [14] was used to estimate $\boldsymbol{A}$. The matrices of NDL+ were obtained using python code developed in the context of [16].

3. Results

Figure 1 visualizes model accuracy across cross-validation runs for the 5 data sets when using standard NDL (left) and NDL+ (right). NDL reaches, on average, a recognition accuracy of 11.72 on the clean datasets (11.58 on clean-20 and 11.86 on clean-50) and 11.13 on the noisy datasets (10.75, 11.36, and 11.29 on noisy-20, noisy-50, and noisy-80, respectively). A Wilcoxon Rank Sum Test indicated that NDL accuracy on 50 hours of clean speech was higher than accuracy on 20 hours of clean speech $W = 8$, $p < .001$. There were also statistically significant difference in mean NDL accuracy between the three noisy datasets (Kruskal-Wallis rank sum test; $H(2) = 19.66$, $p < .001$), with post hoc pairwise multiple comparisons using the Nemeneyi test and p-value adjustment using the Bonferroni correction indicating that mean NDL accuracy for the 20 hours dataset is lower than that for the 50 hours ($p < .001$) and the 80 hours datasets ($p < .01$). As expected, accuracy decreases when replacing clean speech by noisy speech (Wilcoxon rank sum test; $W = 560$, $p < .001$), but the decrease is modest, around 1%.

As illustrated by Figure 1, the performance of NDL+ was superior to that of NDL ($W = 332$, $p < .001$, Wilcoxon test), increasing by 9.14 % of the NDL accuracy, going from 11.37 ± 0.41 to 12.41 ± 0.76. The Wilcoxon test showed that the observed improvement for model accuracies going from NDL to NDL+ are significant across all corpora (Noisy-20: $W = 13$, $p < .005$; other corpora: $W = 0$, $p < .001$). Analysis of variance revealed a significant effect of corpus ($F(4, 45) = 50.73$, $p < .001$) and corpus size ($F(2, 47) = 61.03$, $p < .001$) on the amount of increase from NDL to NDL+ in the model accuracy.

Mean NDL+ accuracy was 12.27 ± 0.21 for clean 20 hours, 13.34 ± 0.13 for clean 50 hours, 11.16 ± 0.38 for noisy 20 hours, 12.45 ± 0.19 for noisy 50 hours, and 12.81 ± 0.1 for noisy 80 hours of speech. Exposing NDL+ to more clean speech (20 hours to 50 hours) significantly increased model accuracy ($W = 0$, $p < .001$, Wilcoxon test). Mean accuracy also differed for the three datasets with noisy speech (Kruskal-Wallis test; $H(2) = 24.6$, $p < .001$). Furthermore, mean NDL+ accuracy differed significantly for both the noisy20–noisy50 and noisy20–noisy80 pairwise comparisons ($p < 0.01$, Nemeneyi post hoc with Benferroni correction). A 0.7 drop is observed in comparing NDL+ accuracy for noisy speech against that of the clean speech (Wilcoxon test; $W = 428$, $p < .05$).

A linear model for NDL accuracy, excluding the noisy-80 condition, showed significant interaction between the clean/noisy status and size of the corpus. However, there is no such interaction with NDL+ (Table 1). Furthermore, all 3 two-way interactions in a model of accuracy as a function of corpus size, corpus clean/noisy status, and method (NDL or NDL+) are well supported.

Table 1: *Linear regression model output for accuracy using NDL (top) and NDL+ (bottom) with the clean/noisy status and size of the corpus (excluding 80 hours) as two-level predictors.*

	Estimate	Std. Error	t-value
Intercept***	11.58	0.06	197.32
Noisy***	−0.84	0.08	−10.07
Size50**	0.28	0.08	3.35
Interaction**	0.33	0.12	2.81
Intercept***	12.27	0.08	156.05
Noisy***	−0.11	0.11	−9.98
Size50***	1.07	0.11	9.58
Interaction	0.22	0.16	1.41

Significance codes '***': $p < .001$; '**': $p < .01$

The recognition of isolated words sliced out of running speech is a hard task both for ASRs and humans. Human accuracy on the German data of [20] ranged from 20% to 40% (NDL performance with training on 20 hours of speech of 20 females was around 20–25%). Recognition accuracy of the present models is lower, ranging from 10.37 to 13.53, unsurprisingly as there is much greater speaker variability while at the same time we are working not with lab-recorded speech but with speech with a much lower signal to noise ratio. To put the present results in perspective, the performance of the open source Mozilla DeepSpeech [21] speech-to-text engine with a pre-trained English model was assessed on the isolated words from the clean-50 and noisy-80 corpora that the NDL models were trained on. Accuracy of single word recognition was 6.28% for the clean corpus and 2.62 for the noisy corpus.

4. Discussion and Conclusions

We presented a cognitively motivated model of speech recognition trained and evaluated on single word tokens taken from real speech data of the Red Hen Lab, using 10-fold cross validation for assessing model accuracy across five datasets that were automatically sampled from the data, including both relatively clear speech and speech with substantial background noise. We also extended previous work with NDL on auditory comprehension by increasing the volume of data in hours from 20 to 50 and 80. We also tested a recent extension of the model, NDL+, which adds a second network that takes the activation vectors of the first network as input, and is trained to map these onto one-hot encoded output vectors for the lexical outcomes.

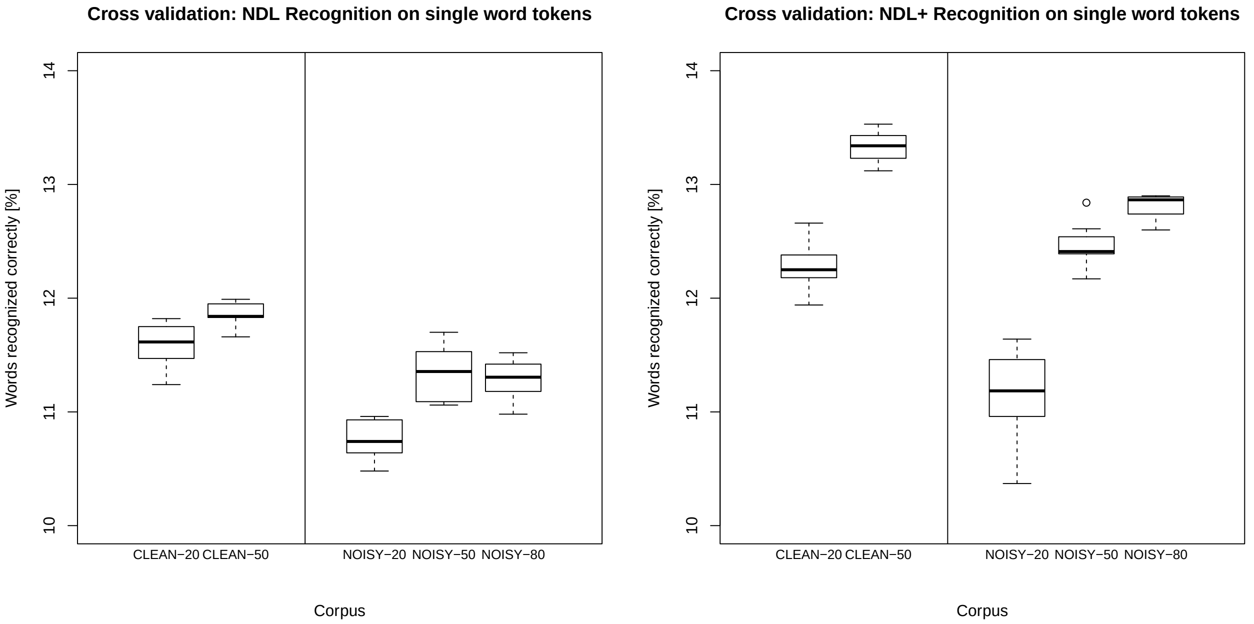

Figure 1: *Box-and-whiskers plots for the accuracy of word identification [%] across 10-fold cross-validation on five corpora for the NDL model in isolation (left panel) and for the NDL+ model paired with NDL (right panel).*

The results show that NDL and NDL+ accuracies improve when the model is exposed to more training data, across both clean and noisy speech. Increasing the amount of training data was more beneficial for the noisy compared to the clean data. For NDL+, but not for NDL, accuracy improved for 80 in comparison with 50 hours of noisy speech, suggesting that with greater quantities of training data, further improvement is possible. Also, training on more data was more advantageous for NDL+ than NDL. We therefore plan to test NDL+ on much larger volumes of speech, with hundreds and perhaps thousands of hours of speech, as available in the Red Hen repository.

As expected, NDL and NDL+ accuracies dropped when the models were exposed to speech in noise, compared to relatively clean studio-recorded speech, but the drop in accuracy was surprisingly modest. To our knowledge, other cognitive models of speech comprehension trained on real speech have been restricted to clear laboratory speech only [26, 27]. We also observed that the number of cues was lower in the noisy environment compared to the clean environment, which dovetails well with reduced sensitivity to speech and degraded comprehension performance. The number of outcomes was also lower in the noisy condition, suggesting that speakers when communicating in noise fall back on a more restricted and presumably better-transmittable vocabulary.

We have evaluated model performance by calculating the proportion of targets that had the highest activation of, on average, 12,030 lexical outcomes. When we consider the number of targeted lexical outcomes among the top 5 and top 10 best supported outcomes, accuracy reaches 30.80% and 40.46% for the clean data and 29.72% and 38.82% for the noisy data. In future work, we plan to compare NDL performance with human performance on words sampled from the Red Hen Lab datasets.

To place the performance of NDL and NDL+ in the context of ASR systems, we compared the performance of our wide learning networks with that of the Mozilla DeepSpeech. The NDL and NDL+ wide models outperformed the DeepSpeech system by roughly 6 to 9%. This comparison does not do justice to the deep speech model, as this model is optimized to recognize words in context rather than isolated words, and is in all likelihood trained on a broader range of registers of spoken English than our news broadcast data. However, we note that the present NDL models are developed as part of a wider project addressing word recognition in running speech. A blueprint of the envisioned general framework can be found in [28].

The results of the present study indicate that a simple error-driven wide network, or a pair of such networks but trained independently, without any back-propagation of errors, can go quite far in modeling auditory comprehension, given the challenges of the task: discriminating between thousands of different lexical outcomes with huge variability in respect of background noise and speakers' accent, dialect, sociolect, speech rate, age and gender. We hope the model will prove useful also for understanding, predicting, and modeling the sensitivity of human listeners to the many social features that characterize speakers and that are part and parcel of what they communicate when speaking.

5. Acknowledgements

This research was funded by the European Research Council (ERC grant 742545, Project WIDE). The authors thank the Distributed Little Red Hen Lab, co-directed by Francis Steen and Mark Turner, and Peter Uhrig for making the Red Hen data available to us. The authors are indebted to Denis Arnold, Konstantin Sering, and Fabian Tomaschek for feedback and technical assistance.

6. References

[1] B. C. Moore, L. K. Tyler, and W. Marslen-Wilson, "Introduction. the perception of speech: from sound to meaning," *Philosophical Transactions of The Royal Society B*, pp. 917–921, 2008.

[2] A. M. Liberman, *Speech: A special code*. MIT press, 1996.

[3] R. L. Diehl, A. J. Lotto, and L. L. Holt, "Speech perception," *Annual Review of Psychology*, vol. 55, pp. 149–179, 2004.

[4] D. Norris and J. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review*, vol. 115, no. 2, pp. 357–395, 2008.

[5] R. K. Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: conventional methods (part i)," *International Journal of Speech Technology*, vol. 14, no. 4, p. 297, 2011.

[6] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[7] G. Sampson, "Is there a universal phonetic alphabet?" *Language*, vol. 50, no. 2, pp. 236–259, 1974.

[8] R. F. Port and A. P. Leary, "Against formal phonology," *Language*, vol. 81, pp. 927–964, 2005.

[9] R. H. Baayen, P. Milin, D. F. Durdević, P. Hendrix, and M. Marelli, "An amorphous model for morphological processing in visual comprehension based on naive discriminative learning." *Psychological review*, vol. 118, no. 3, p. 438, 2011.

[10] P. Milin, L. B. Feldman, M. Ramscar, P. Hendrix, and R. H. Baayen, "Discrimination in lexical decision," *PLOS-ONE*, vol. 12, no. 2, p. e0171935, 2017.

[11] R. A. Rescorla, A. R. Wagner *et al.*, "A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," *Classical conditioning II: Current research and theory*, vol. 2, pp. 64–99, 1972.

[12] M. Ramscar, D. Yarlett, M. Dye, K. Denny, and K. Thorpe, "The effects of feature-label-order and their implications for symbolic learning," *Cognitive Science*, vol. 34, no. 6, pp. 909–957, 2010.

[13] M. Ramscar, C. C. Sun, P. Hendrix, and R. H. Baayen, "The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline," *Psychological Science*, 2017, https://doi.org/10.1177/0956797617706393.

[14] Cyrus Shaoul, Samuel Bitschau, Nathanael Schilling, Antti Arppe, Peter Hendrix, Petar Milin, and R. Harald Baayen, "ndl2: Naive discriminative learning: an implementation in r," R package, 2015.

[15] K. Sering, M. Weitz, D.-E. Künstle, and L. Schneider, "Pyndl: Naive discriminative learning in python," Jan. 2018. [Online]. Available: http://pyndl.readthedocs.io/en/latest/

[16] T. Sering, P. Milin, and R. H. Baayen, "Language comprehension as a multiple label classification problem," *Statistica Neerlandica*, pp. 1–15, 2018.

[17] H. Pham and H. Baayen, "Vietnamese compounds show an anti-frequency effect in visual lexical decision," *Language, Cognition and Neuroscience*, vol. 30, no. 9, pp. 1077–1095, 2015.

[18] P. Milin, L. B. Feldman, M. Ramscar, P. Hendrix, and R. H. Baayen, "Discrimination in lexical decision," *PLOS-ONE*, vol. 12, no. 2, p. e0171935, 2017.

[19] M. Linke, F. Broeker, M. Ramscar, and R. H. Baayen, "Are baboons learning "orthographic" representations? probably not," *PLOS-ONE*, vol. 12, no. 8, p. e0183876, 2017.

[20] D. Arnold, F. Tomaschek, K. Sering, F. Lopez, and R. H. Baayen, "Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit," *PLOS-ONE*, vol. 12, no. 4, p. e0174623, 2017.

[21] "Project deepspeech," https://github.com/mozilla/DeepSpeech, Mozilla Organization, 2013.

[22] R. Ochshorn and M. Hawkins, "Gentle: A forced aligner," 2016.

[23] D. W. Paul Boersma, "Praat: doing phonetics by computer," 2006.

[24] S. Zerlin, "Traveling-wave velocity in the human cochlea," *The Journal of the Acoustical Society of America*, vol. 46, no. 4B, pp. 1011–1015, 1969.

[25] D. Arnold, "Acousticndlcoder: Coding sound files for use with ndl," 2017, r package version 1.0.1. [Online]. Available: https://CRAN.R-project.org/package=AcousticNDLCodeR

[26] O. Scharenborg, D. Norris, L. Bosch, and J. M. McQueen, "How should a speech recognizer work?" *Cognitive Science*, vol. 29, no. 6, pp. 867–918, 2005.

[27] L. ten Bosch, M. Ernestus, and L. Boves, "Comparing reaction time sequences from human participants and computational models," in *Proceedings of Interspeech 2014: 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 462–466.

[28] R. H. Baayen, C. Shaoul, J. Willits, and M. Ramscar, "Comprehension without segmentation: A proof of concept with naive discriminative learning," *Language, Cognition and Neuroscience*, vol. 31, no. 1, pp. 106–128, 2015.