Learning from the acoustic signal: Error-driven learning of low-level acoustics discriminates vowel and consonant pairs

Jessie S. Nixon (jessie.nixon@uni-tuebingen.de)
Quantitative Linguistics, Wilhelmstr. 19, Eberhard Karls University of Tübingen, Tübingen, Germany

Fabian Tomaschek (fabian.tomaschek@uni-tuebingen.de)
Quantitative Linguistics, Wilhelmstr. 19, Eberhard Karls University of Tübingen, Tübingen, Germany

Abstract

Until the last couple of decades, research on speech acquisition generally assumed that infants were born with innate knowledge of a universal set of phonetic features that occurred across all the world's languages and that learning one's native language involved selecting the appropriate subset from this broader group. Over the last two decades, this account has given way to the idea that speech sounds are learned via general learning mechanisms. Statistical clustering models have become the most common way to explain how infants learn the sounds of their language. Over the first few months of life, infants go from being able to discriminate the sounds of all languages, to perceiving speech sounds in a way that is increasingly honed to their native language. However, recent empirical and computational evidence suggests that purely statistical clustering methods may not be sufficient to explain speech sound acquisition. The present study used discriminative, error-driven learning, an implementation of the Rescorla-Wagner model, to model early development of speech perception. Expectations about the upcoming acoustic speech signal were learned from the surrounding speech signal, with spectral slices from a spontaneous speech corpus as both inputs and outputs of the model. After training, the model was tested on vowel and fricative continua. The model was able to discriminate both the vowels and the consonants, with decreasing activation along the continuum for steps further away from the target sounds. Inspection of cue weights showed that both the vowels and the fricatives were discriminated based on cues in the expected spectral frequency ranges. Vowel discrimination occurred in the frequency bands corresponding to vowel formants; fricative discrimination occurred in the lowest frequencies corresponding to presence versus absence of voicing. These results suggest that a discriminative error-driven approach may provide a viable alternative to statistical clustering for modelling early infant speech acquisition.

Keywords: error-driven learning; discriminative learning; statistical learning; Rescorla-Wagner model; speech acquisition; first language acquisition

Introduction

One of the key questions in speech perception research is how human speech perception becomes specialised for the specific language(s) in the learner's environment. In the 1970s, and even up to the 1990s, many researchers assumed that infants were born with innate knowledge about the sound units of language. Language was thought to be too complex to be learned through general learning mechansims and so innate neural functions were proposed that were specialised for speech sounds (e.g. Eimas & Corbit, 1973).

Recent decades have seen a shift in thinking regarding both the ability of infants to learn from their environment through general learning mechanisms and, relatedly, the ability to learn from a variable acoustic signal, rather than a set of discrete units. Statistical (or distributional) learning models in particular are now arguably the dominant models of language acquisition. Yet recent research has raised the question of whether purely statistical models can adequately account for speech acquisition. We present an alternative account of how listeners can learn from the distribution of information in speech: through error-driven learning of the acoustic signal.

According to statistical learning models, listeners keep track of the frequency of occurrence of events in their environment (Saffran, Aslin, & Newport, 1996; Maye & Gerken, 2000). In the case of speech sounds, listeners are thought to keep track of acoustic cue values, which form clusters around the peaks of distributions (Maye & Gerken, 2000; Maye, Werker, & Gerken, 2002). Based on how many clusters occur for the various acoustic cues, listeners learn how many speech categories occur in their language. Discriminative cues form at least two clusters, such as short and long voice onset time in English (e.g. /b/ vs. /p/), while non-discriminative cues form only one cluster, such as pharyngealisation in English.[1]

Statistical learning models have been highly successful in suggesting a possible mechanism by which learners could learn their language from the input. This has likely been one of the major contributing factors in the transformation in thinking regarding the learnability of language. Yet, despite their successes, recent empirical and computational evidence suggests that purely statistical models may not be able to account for speech sound acquisition (Baayen, Shaoul, Willits, & Ramscar, 2016; Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009; McMurray & Hollich, 2009; Nixon, 2020; Soderstrom, Conwell, Feldman, & Morgan, 2009).

McMurray and Hollich (2009) argue that behavioural studies cannot answer the question of whether statistical learning is the mechanism underlying learning and that computational modelling is essential for addressing this question. Computational models have exposed a number of challenges for statistical models in explaining first language speech sound acquisition. For example, clustering algorithms without competition fail to converge on the right number of categories (McMurray et al., 2009). If phonemes are taken to be the

[1] Pharyngealisation discriminates consonants and vowels in a number of languages, including Arabic (e.g. Mohamed, 2001)

units that are to be learned, the distribution of tokens in categories such as vowels is not such that it can be learned through unsupervised clustering mechanisms, because there is often too much overlap between categories (Feldman et al., 2013). Analysis of the time course of effects of statistical variance on eye movements suggests that the statistical variance effects are more likely to stem from uncertainty or error in the system, rather than perceptual learning of statistical distributions (Nixon, van Rij, Mok, Baayen, & Chen, 2016). In addition, in the distributional learning paradigm, the greatest learning results from rare surprising events, as predicted by error-driven learning models, rather than an equal amount of learning from all events (i.e. a linear or veridical representation of the distributions), as predicted by statistical learning models (Olejarczuk, Kapatsinski, & Baayen, 2018). It has also recently been demonstrated that speech sound acquisition involves cue competition and is affected by the predictive structure of learning events, findings that cannot be explained by purely statistical models (Nixon, 2020). Together these findings suggest that it will be necessary to find alternatives to purely statistical clustering models in order to explain first language speech sound acquisition.

In the present study, using computational modelling, we present an alternative account of early infant speech sound acquisition. The model is based on the Rescorla-Wagner learning equations (Rescorla & Wagner, 1972). The Rescorla-Wagner model was developed based on decades of research, particularly in animal learning (Pavlovian conditioning). Since its publication, it has had a profound impact on a number of areas of psychology (see e.g. Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996, for reviews). Insights from the model and from error-driven learning theory have only very recently begun to be applied to language (e.g. Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Ramscar, Dye, & McCauley, 2013), yet a wide range of linguistic phenomena have been successfully modelled using error-driven learning (Arnold, Tomaschek, Sering, Lopez, & Baayen, 2017; Kapatsinski, 2018; Nixon, 2020; Tomaschek, Plag, Ernestus, & Baayen, 2019; Shafaei-Bajestan & Baayen, 2018).

The model is very simple and there are no hidden layers. Compared to deep neural networks, which generally have many hidden layers, this transparency can be considered an advantage in cognitive science, as the results are generally more interpretable in terms of cognition. The Rescorla-Wagner equations and their implementation in the R package `NDL` (Arppe et al., 2015) update connection strengths between present cues and all encountered outcomes. Connection strength increases between cues and outcomes that are present during a learning event (or trial). Connection strength decreases between cues that are present and outcomes that are not present. For cues that are not present, no adjustment is made. Outcomes have a maximum connection strength (here set to the default, 1). The adjustments to cue-outcome connection strength are a constant proportion of the error between the current connection strength and either the maximum connection strength for present outcomes or zero for absent outcomes (see Nixon, 2020, for a straightforward introduction to the Rescorla-Wagner equations and their application in second language speech sound acquisition).

We propose that infants learn by using all available perceptual cues to predict upcoming important outcomes. One aspect of this process is that infants should learn to use the acoustic signal to predict upcoming acoustic signal.[2] Therefore, we model the learning of speech as a process of prediction and discrimination of the upcoming signal from the surrounding acoustic signal itself. We use `NDL` for training the corpus. We test the model on discrimination of vowel and consonant continua.

Method

Training materials and procedure

The model was trained on the Karls Eberhard Corpus of native German spontaneous speech (KEC; Arnold & Tomaschek, 2016). The corpus consists of 79 speakers, with approximately one hour of spontaneous dialogue between two speakers known to each other (on average 30 minutes of signal per speaker). Each speaker was recorded in separate sound-attenuated booth at 44100 Hz.

We trained individual NDL networks for 38 of the speakers in the KEC. The Rescorla-Wagner learning algorithm takes discrete cues and outcomes as input and output. Therefore, the acoustic features that served as cues and outcomes of the NDL network consisted of discretised spectral slices of the acoustic signal. The continuous speech signal was divided into 25 ms windows with 15 ms overlap as is common in speech processing and machine learning (e.g. Chapaneri & Jayaswal, 2013). Most information conveyed in speech occurs below 10,000 Hz; therefore, to reduce processing cost, only frequencies up to 10,200 Hz (49 mel) were included. To reflect the non-linear perceptual sensitivities of the human cochlea to spectral change (Allen, 2008), spectral frequencies were divided into 104 equidistant (approx. 0.47) steps on the mel scale. For each *spectral component* (i.e. 25 ms by 0.47 mel cell), the log power spectrum was calculated and rounded to one decimal place as a measure of amplitude.

Thus, for each 25 ms time step, an amplitude value was obtained for each 0.47 mel spectral component from 0 to 49 mel. Note that the numerical value of neither the spectral frequency nor the amplitude was available to the model due to discretisation. Each spectral component was named according to its spectral frequency component and amplitude values, with each spectral component-by-amplitude combination simply forming a unique code (e.g. SC12A4 for the 12th spectral component at 344 Hz - 375 Hz / 5.16 mel - 5.63 mel with a log amplitude of 4).

In each learning event, the outcomes were the 104 spectral components for a single 25 ms window. Cues consisted of the

[2] This learning process presumably involves all sensory input, not just acoustic information. But for practical reasons, we restrict ourselves to the acoustic modality in the present study.

104 spectral components for each of the two previous windows and one following window (104 spectral components × three time windows = 312 cues per learning event). Repetitions of identical cues were permitted, if they occurred. Thus, the model was trained to use low-level acoustic cues to predict low-level acoustic outcomes.

Training was conducted with a moving window of these four time steps across the whole speech file for each individual speaker. At the end of training, the model consisted of a matrix of cue-outcome connection strengths. For each encountered amplitude value for each spectral component, the value of the connection strengths indicate the degree of expectation that the upcoming signal will contain each of the various possible amplitude values for the different spectral components.

Model evaluation

Model performance was evaluated based on two tests commonly used in human speech perception studies, namely the AX and AXB discrimination tasks. The AX task is typically used to test whether participants can discriminate sounds (Nixon et al., 2018; Pisoni, 1973; Tomaschek, Truckenbrodt, & Hertrich, 2013, 2015). Two sounds are presented and participants are asked to decide whether they are the same or different. The AXB task is typically used to test which of two sounds (A or B) participants perceive a third, target, sound (X) as most similar to (MacKain, Best, & Strange, 1981). Three sounds are presented (AXB) and participants are asked to indicate whether the second sound is most similar to the first or third sound.

Test material We created 20-step continua for four pairs of German vowels and four pairs of fricatives. The pairs were produced in carrier words by the second author. Due to space restrictions, we examine one vowel pair /i/-/y/ and one consonant pair here /v/-/f/.

The vowel continua were created in Praat (Boersma & Weenink, 2014) by synthesising intermediate steps with interpolated formant frequencies between the two endpoint vowels (Winn, 2014). Vowel formants were: /i/: 280 Hz, 2400 Hz and 2910 Hz; /y/: 270 Hz, 1545 Hz and 1940 Hz. The fricative continua were created by linearly increasing/decreasing and adding the wave forms of the two endpoints in a step-wise manner. Visual inspection of the fricatives and intermediate stimuli and their spectra confirmed linear change of the spectrum.

Finally, discrete 25ms-by-0.47 Hz spectral components were then created for the test stimuli in the same manner as above for the training stimuli.

AX test

A form of the AX task commonly used with infants is the head turn paradigm (e.g. Werker & Tees, 1984). Infants hear a string of speech sounds and are trained to turn their head when the sound changes. Our AX test predicts the infant's decision to turn their head or not in this task. When the infant hears X on the current trial after A on a previous trial, do they respond to the change? Or, put another way, is X sufficiently surprising to warrant a head turn?

We model the continuum endpoints as A and each of the continuum steps as X. To determine whether X differs from A, we calculate the degree to which X activates A. *Activation* captures the amount of support from cues to a specific outcome. Activation is calculated by summing the weights between all present cues (in this case the cues for all time steps and mel frequency bands in the continuum step, X) to the outcome (in this case each individual mel frequency band in the endpoint, A). Calculating the activation individually for each of the 104 mel frequency bands in the endpoint stimulus (A) also allows us to generate predictions about which spectral frequencies are most important for discrimination.

AXB test

Our AXB test simulates performance in a two-alternative forced-choice task. The continuum steps represent the test items, X, and the endpoints represent the alternative choices A and B. For each mel frequency band, we determined whether activation was higher for the left or right endpoint stimulus (A or B). We then summed the number of winning frequency bands for each endpoint to get the probability of an A versus B response. The categorisation plots in Figure 1 (right panels) show the percentage of mel frequency bands with higher activation for A than B along the continuum.

Results

Model results were analysed using generalised additive mixed modelling (GAMM; Lin & Zhang, 1999; Wood, 2017). For the AX test, a Gaussian GAMM of activation was modelled with smooths of *continuum step*, *spectral frequency* and their interaction. For the AXB test, a binomial GAMM modelled endpoint stimulus selection (A vs. B) with a smooth of continuum step. Both models included random intercepts for speakers.

Figure 1 illustrates the GAMM model results for the vowels (top row) and fricatives (bottom row). The left two columns show the estimated effect of the interaction between continuum step (x-axis) and spectral frequency (mel/Hz; y-axis) on the activation of the left endpoint stimulus (left panels) and right endpoint stimulus (centre panels). Activation is shown on the z-axis and is colour-coded from cool to warm colours: dark blue represents low activation; green, medium; yellow, high activation. The right column illustrates the probability (y-axis) of selecting the left endpoint point stimulus in the AXB classification test for each step along the continuum (x-axis). Y-axis values were back-transformed from logit to probabilities.

In the model of the /i/ vowel (top left), in spectral frequencies between ∼1000 Hz and ∼3300 Hz, activation is high at the left of the continuum close to the target, /i/, and decreases towards the competitor, /y/. The inverse pattern occurs for the /y/ vowel (top centre). The further away the test stimulus is from the target, the lower the activation. The frequency

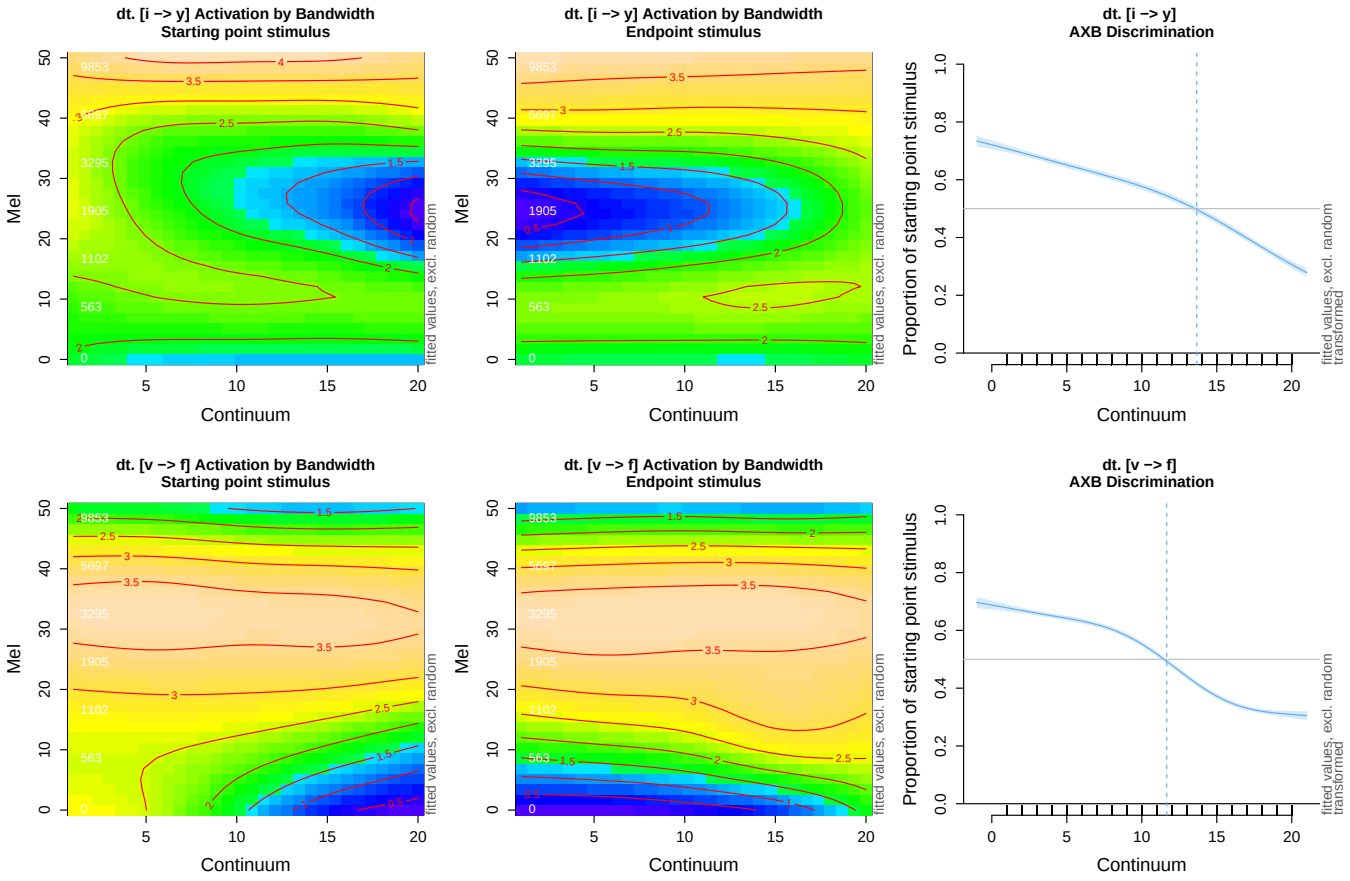

Figure 1: Results of model evaluation tests: AX (left and centre panels) and AXB (right panel) for vowel (/i/-/y/; top row) and fricative discrimination (/v/-/f/; bottom row.). Y-axis shows mel in black and Hertz in white.

bands with the greatest changes in activation correspond to the second and third formants, with higher and lower frequency ranges much less affected. Note that the pattern of activation is characterised by *low activation* in parts of the spectrum that support the competitor vowel (blue areas). This suggests *unlearning* of cues that do not predict the target outcome. The AXB test for the vowels shows an almost linear discrimination of the /i/-/y/ vowels.

For the consonants /v/-/f/, the biggest change in activation between target and competitor is below 500 Hz, i.e. the frequency band that represents the changes in fricative voicing. As with the vowels, activation is high at the target end and low at the competitor end of the continuum, suggesting unlearning of cues for a given target.

The AXB test for the /v/-/f/ contrast (bottom right panel) shows a sigmoidal classification function. This nonlinear change over the continuum is typical of consonant discrimination. While we do not have space to present them here in full, the other consonants and vowels tested showed a similar general pattern: relatively linear categorisation function of vowels and non-linear categorisation for consonants.

It is worth noting here again that, due to the discretisation of cues and outcomes, the model does not have a representation of the acoustic similarity or dissimilarity of the stimuli. The change in activation across the continuum emerges from the degree to which each continuum step *activates* the endpoint stimulus, not due to similarity per se. The activation in turn comes from the degree to which the various acoustic cues predicted one another during training. This allows the model to develop a nonlinear change over the continuum, as well as an asymmetry between the left and right target stimuli, despite linear acoustic change.

Discussion

The present study investigated an alternative account for how early first language speech sound acquisition could occur without assuming innate knowledge of phonological units, such as phonemes or phonetic features. We modelled speech sound aquisition as discriminative, error-driven learning of the acoustic speech signal. After training with incoming spectral amplitude components predicting upcoming spectral amplitude components from a corpus of spontaneous speech, the model simulated discrimination between speech tokens in a way that mimics human behaviour in perceptual tasks, such as infants' head-turn decisions in the head turn paradigm.

An aspect of the results that is worth highlighting is that

the discrimination – that is, the changes in activation across the continuum – occurs in the expected spectral frequency ranges. For vowels, the greatest effects were in the range ∼1000 Hz and ∼3300 Hz, the frequency range of the second and third vowel formants. Lip rounding (in the vowel /y/) lowers the formant frequencies (relative to /i/). For the fricatives, the greatest effects were at the lowest spectral frequencies, corresponding to the presence (in /v/) versus absence of voicing (in /f/). The emergence of differential activation patterns depending on the sound pair being discriminated suggests that the model learned which cues were important for discrimination for different speech sounds and, importantly, also *unlearned* cues that were not predictive.

A prediction that falls naturally out of the model is that changes in consonants and vowels are perceived differently. The AXB discrimination task showed a nonlinear sigmoidal function for the consonants, with relatively poor within-category discrimination, while vowels showed a more continuous, linear discrimination. The tendency for consonants to show sharper discrimination than vowels has been observed in human speech sound perception (Pisoni, 1973), although the effect has been found to be task dependent (Gerrits & Schouten, 2004; Massaro & Cohen, 1983). This may seem a surprising result, as neither the training nor the test involved any top-down identification of units of speech of any kind, including distinctions between consonants and vowels. Exactly how the model achieves this requires further investigation. One possibility is that due to coarticulation, gradual changes in spectral frequency during training developed relatively high expectations of encountering slightly different vowel formants. On the other hand, the cues in voiceless fricatives may have been poor predictors of voiced fricatives in training. Interestingly, there is an asymmetry between targets in the fricative results (comparing the left vs. centre panels), such that voiced fricatives are better predictors of voiceless fricatives than vice versa. This may be due to voiceless or final devoiced consonants in which the initial portion of the consonant has spillover voicing from the preceding vowel.

One question is whether the results really stem from learning the predictive relationship between cues and outcomes during the training phase and not simply from, for example, changes in acoustics across the continuum. The latter explanation seems unlikely for at least two reasons. Firstly, the acoustic changes are linear - changes occur in equal steps across the continuum. However, the model's predicted perceptual changes are nonlinear, at least in the case of the consonants. And secondly, activation patterns in the left and right endpoint stimuli are asymmetrical. Given that the acoustic differences are equal, if the model perception were acoustic only, we would expect the endpoints to be the inverse of each other. This asymmetry in the activations suggests an asymmetry in the cue-outcome weights. This suggests that the model has learned the informative cues and unlearned or downweighted cues that are predictive of competing outcomes.

In his review of current models of speech sound and word learning, Räsänen (2012) concludes that without assuming innate phonetic knowledge, learning words is possible if the speech signal is represented acoustically, as sequences of spectral features. However, models that first learn phone-like units and then learn words from phone sequences have had only limited success. The present study demonstrates a method for learning speech sounds acoustically without assuming phone-like units. At the same time, the model avoids some of the issues that statistical clustering models face.

Our aim here was to model the first few months of life, when even the passive vocabulary is minimal. At this age, due to the small or non-existent lexicon, infants presumably do not learn speech sounds by discriminating between lexical items. We propose instead that infants discriminate important events in the world based on any perceivable cues in the environment. In this study, we focused on the speech signal as both cues and outcomes. However, it is most likely that learning is multimodal, incorporating all available perceptual cues. As infants grow older, they also become better at using acoustic cues to predict events in the world other than the acoustic signal itself, such as feeding, social interactions, the appearance of important people, objects, foods, toys and so on - that is, they 'learn words'. Infants as young as six months have been shown to recognise many common words (Bergelson & Swingley, 2012). As the lexicon develops, lexical/semantic outcomes are likely to also become important outcomes that play a role in further developing discrimination of speech sounds. Further research is needed to incorporate other sensory modalities and the role of the developing lexicon.

In summary, the present study provides an alternative to statistical clustering models as an account of the first few months of speech sound acquisition. At the heart of the model is the idea that infants (and people generally) use currently available sensory information to predict important upcoming events in the world. Through feedback from prediction error, infants learn which acoustic cues are predictive of upcoming signal and which cues can be ignored. After training, the model was able to discriminate between speech sounds in ways very similar to human listeners. Furthermore, the model captured the specific spectral frequency ranges relevant to the sound pair in question.

Acknowledgements

We are grateful to four anonymous reviewers whose insightful and constructive comments helped to improve this manuscript. This research was supported by a collaborative grant from the Deutsche Forschungsgemeinschaft (German Research Foundation; Research Unit FOR2373 'Spoken Morphology', Project 'Articulation of morphologically complex words' , BA 3080/3-1) and an ERC Advanced Grant (Grant number 742545).

References

Allen, J. B. (2008). Nonlinear cochlear signal processing

and masking in speech perception. In *Springer handbook of speech processing* (pp. 27–60). Springer.

Arnold, D., & Tomaschek, F. (2016). The Karl Eberhards Corpus of spontaneously spoken Southern German in dialogues - audio and articulatory recordings. In C. Draxler & F. Kleber (Eds.), *Tagungsband der 12. tagung phonetik und phonologie im deutschsprachigen raum* (p. 9-11). Ludwig-Maximilians-Universität München. Retrieved from https://epub.ub.uni-muenchen.de/29405/

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). 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*, *12*(4), e0174623.

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2015). ndl: Naive discriminative learning [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ndl (R package version 0.2.17)

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Boersma, P., & Weenink, D. (2014). *Praat (version 5.5).*

Chapaneri, S. V., & Jayaswal, D. J. (2013). Efficient speech recognition system for isolated digits. *IJCSET*, *4*(3), 228–236.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Gerrits, E., & Schouten, M. (2004). Categorical perception depends on the discrimination task. *Perception & psychophysics*, *66*(3), 363–376.

Kapatsinski, V. (2018). *Changing minds changing tools: From learning theory to language acquisition to language change*. MIT Press.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society*, *61*(7), 381.

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, *2*(4), 369–390.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech communication*, *2*(1), 15–35.

Maye, J., & Gerken, L. (2000). *Learning phonemes without minimal pairs*. Proceedings of the 24th Annual Boston University Conference on Language Development.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3).

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

McMurray, B., & Hollich, G. (2009). Core computational principles of language acquisition: can statistical learning do the job? introduction to special section. *Developmental Science*.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the rescorla-wagner model. *Psychological bulletin*, *117*(3), 363.

Mohamed, Y. (2001). Pharyngealization in arabic: Modelling, acoustic analysis, airflow and perception. *Revue de La Faculté des Lettres El Jadida*, *6*, 51–70.

Nixon, J. S. (2020). Of mice and men: speech acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Nixon, J. S., Boll-Avetisyan, N., Lentz, T. O., van Ommen, S., Keij, B., Çöltekin, Ç., ... van Rij, J. (2018, June). Short-term exposure enhances perception of both between- and within-category acoustic information. In *Proceedings of the 9th International Conference on Speech Prosody* (pp. 114–118). Poznan, Poland.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*(s2).

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & psychophysics*, *13*(2), 253–260.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*(4), 760–793.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, *54*(9), 975–997.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Cur-*

rent research and theory (Vol. 2, pp. 64–99). New-York: Appleton-Century-Crofts.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294).

Shafaei-Bajestan, E., & Baayen, R. H. (2018, September). Wide learning for auditory comprehension. In *Interspeech* (pp. 966–970). Hyderabad.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the rescorla-wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321.

Soderstrom, M., Conwell, E., Feldman, N., & Morgan, J. (2009). The learner as statistician: three principles of computational success in language acquisition. *Developmental Science*, *12*(3), 409–411.

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final s in english with naïve discriminative learning. *Journal of Linguistics*, 1-39.

Tomaschek, F., Truckenbrodt, H., & Hertrich, I. (2013). Neural processing of acoustic duration and phonological german vowel length: Time courses of evoked fields in response to speech and nonspeech signals. *Brain and Language*, *124*(1), 117 - 131. doi: http://dx.doi.org/10.1016/j.bandl.2012.11.011

Tomaschek, F., Truckenbrodt, H., & Hertrich, I. (2015). Discrimination sensitivities and identification patterns of vowel quality and duration in german /u/ and /o/ instances. In A. Leemann, M.-J. Kolly, S. Schmid, & V. Dellwo (Eds.), . Frankfurt am Main / Bern: Lang.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*(1), 49–63.

Winn, M. (2014). *Gui-based wizard for creating realistic vowel formant continua from modified natural speech. version 30.* Retrieved from `www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt`

Wood, S. N. (2017). *Generalized additive models: an introduction with r*. Chapman and Hall/CRC.