Acoustic cue variability affects eye movement behaviour during non-native speech perception: a GAMM model

Jessie S. Nixon[1], *Catherine T. Best*[2]

[1]New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, New Zealand
[2]The MARCS Institute, Western Sydney University, Australia

jess.s.nixon@gmail.com, C.Best@westernsydney.edu.au

Abstract

Participants in the 'visual world' paradigm simultaneously process both auditory and visual cues in order to match speech to target images. Previous research has shown that when native speakers listen to speech that has high within-category variability in the contrastive dimension, auditory perceptual uncertainty increases, resulting in increased looks to competitor objects. This suggests a cross-modal effect, where reduced reliability in the auditory domain leads to increased search for evidence in the visual domain.

The present study investigated the effects of within-category acoustic variability on eye movements during the acquisition of a new acoustic dimension not present in the native language, namely English speakers' acquisition of lexical tone. All participants heard a bimodal distribution of stimuli, with distribution peaks at the prototypical pitch values for high and mid tones; however, presentation frequency differed between conditions: high-variance vs. low-variance. Based on previous research, we expected lower uncertainty and better learning in the low-variance condition.

GAMM models of eye movement data showed that within-category acoustic variance increases perceptual uncertainty in the auditory domain and hinders acquisition of a cue dimension. The results also show a cross-modal effect: lower relibility in the auditory domain leads to increased search for cues in the visual domain, even when visual cues are held constant across conditions.

Index Terms: speech perception, cross-modal effects, statistical learning, second language acquisition, Cantonese lexical tone, visual world eyetracking, generalised additive mixed models (GAMMs)

1. Introduction

The organisation of acoustic cues varies substantially across languages. Cue dimensions that are lexically contrastive in one language may not be contrastive in another. Therefore, acquisition of a new language often involves learning to substantially adjust cue weights (i.e. to adjust the degree to which various cues in the signal are utilised, consciously or unconsciously) for lexical contrasts. In some cases, this can pose significant challenges. Expert knowledge of statistical regularities in one's native language can lead to expectations that hinder non-native speech perception [1, 2]. Statistical properties that seem to play a role in shaping cue perception include the number of distribution peaks along a cue dimension [3, 4, 5, 6], acoustic distance between peaks in a bimodal distribution [7, 8] and within-category acoustic variance in a bimodal distribution [9, 10, 11].

Many recent studies have emphasised the role of variability in shaping and reshaping native and non-native sound systems. For example, early first language acquisition [12, 13] and second language acquisition [14, 15, 16, 17] seem to benefit from multiple speakers, compared to a single speaker in the training input. When there are multiple speakers, this increases variability in *non-contrastive* indexical dimensions, which seems to have the effect of highlighting the relative invariance of the contrastive dimensions. This is consistent with learning models, which posit that learning involves not only acquisition of knowledge, but also learning to ignore irrelevant cues [18, 19, 20].

An aspect that has received less attention is variability within *contrastive* dimensions. Acoustic studies suggest that this variability may be an important factor in discriminating sound contrasts. For example, for native English speakers, the third formant (F3) is generally the most reliable cue to the /l/ and /r/ contrast [21]. F3 values cluster around the /l/ and /r/ productions with relatively little spread or overlap. While there is some difference in the distribution of the second formant (F2) between /l/ and /r/, it is highly variable, with a high degree of overlap, and is therefore not a reliable cue for discriminating between /l/ and /r/ [22]. In addition, in recordings of native Japanese productions of English /l/ and /r/ [22], the F3 values are highly variable and largely overlapping for /l/ and /r/ productions. This increased variance and category overlap corresponds to reduced effectivenes of the cue for discriminating between the /l/ and /r/ tokens produced by these speakers.

Nixon and colleagues [10] investigated the effects of variability within contrastive dimensions on the temporal dynamics of perceptual uncertainty during native speech perception using the'visual world' eyetracking paradigm [23]. In this paradigm, participants see four pictures on the screen, hear a word and are instructed to click on the picture corresponding to the word. Nixon and colleagues found that, although the pictures were identical in both conditions, effects of acoustic variance affected visual processing. Effects emerged very early, in the first fixations of the trial. As auditory variability increased and speech cues became less reliable, listeners looked around more, presumably in search of visual information to provide further support for partially activated candidates. The idea that listeners were seeking additional evidence in the high-variability condition seems to suggest the appropriate conditions for adjusting cue weights, and perhaps increasing weights of previously downweighted cues.

The present study investigates whether such within-category acoustic variance also affects audio-visual processing during acquisition of a new acoustic dimension in a non-native language. This question is addressed by examining the effect of within-category acoustic variance on eye movements during native English speakers' acquisition of a pitch cue (fundamental frequency; *f0*) in a Cantonese lexical tone contrast. English does not use f0 as a lexical contrast, and tone can be notoriously difficult for beginning learners of non-tonal languages [24]. Based on previous studies [9, 10, 11], we expected greater weighting of the pitch cue over the course of the experiment - that is, better learning - in the low-variance, compared to the high-variance condition.

2. Method

2.1. Participants

Thirty-seven native English-speaking students from the University of Western Sydney who had not previously studied any tone language were recruited for the experiment for course credit[1]. Participants were tested individually in a sound-attenuated booth.

2.2. Experiment design and stimuli

Visual stimuli were black-on-white line drawings of eight common objects. Auditory stimuli were four minimal pairs of mid- and high-tone words (e.g. gon_mid and gon_high). They were real Cantonese nouns; however, the matching images were not the true meaning in Cantonese. The purpose of using common, very high-frequency objects was to reduce cognitive load in the language-learning task. All auditory stimuli were recorded by a male native speaker of Hong Kong Cantonese. Stimuli were then resynthesised into a 14-step pitch continuum (e.g. gon_mid to gon_high) using PRAAT [25]. One half of the continuum corresponded to the mid tone and one half to the high tone.

The number of times participants heard each token of the continuum followed a bimodal distribution, with the two peaks of the distribution corresponding to the prototypical f0 for the mid- and high-tone stimuli, respectively. All participants heard the same number of tokens; but the number of times they heard each token differed between conditions, with greater spread from the mean (statistical variance) in the *high-variance* versus the *low-variance* distribution (see Figure 1). The experiment consisted of 240 experimental trials, divided into six blocks of 40 trials, with breaks between the blocks. The order of presentation was pseudo-randomised for each participant.

2.3. Procedure

Participants sat at a viewing distance of 60 cm from a computer screen equipped with an SR Research Eyelink 1000 remote eyetracker. A chinrest and headrest were used to minimise movement. Stimulus presentation and data acquisition were conducted using SR Research Experiment Builder computer software with a sampling rate of 1000 Hz. The session began with ten practice trials to familiarize participants with the experimental procedure. None of the images or auditory stimuli from the experimental block appeared in the practice block. Each experimental trial began with a brief (1000 ms) presentation of four pictures, one in each quadrant of the screen. The purpose of the preview was to reduce noise in the data by reducing the time and likelihood of participants scanning the images at the beginning of the trial. The display always contained a target, a competitor and two distractor items. The target (e.g. gon_mid, 'pen') and competitor (e.g. gon_high, 'apple') had the same segmental syllable, but differed in tone. The location of each picture condition on the screen and their location relative to each other were randomised to avoid strategic effects.

[1]Participants were not explicitly asked whether they had studied a tone language, as this might influence the experiment results. Instead, they were asked to list all languages they spoke or had studied, and were screened if they did not meet this criterion.

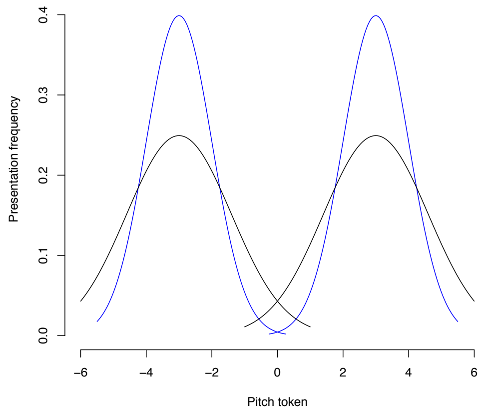

Figure 1: *Illustration of the presentation frequency distributions in the high-variance (black lines) and low-variance conditions (blue lines).*

The preview disappeared, followed by a gaze-contingent fixation cross to ensure participants were fixating the centre of the screen at the beginning of the critical trial period. The pictures then reappeared simultaneously with presentation of the auditory stimulus. Participants were instructed to select the picture corresponding to the word they heard by clicking on it with the mouse, and to guess if they did not know. They were given feedback ('correct'/'wrong') after each trial. Participants were told that this was a language-learning task, but were not informed about the pitch or tone manipulation or the target language.

3. Analysis

Eye movement data were analysed using *Generalised Additive Mixed Models* [26, 27, 28, GAMM] using the `mgcv` package (version 1.8.13) in R [29, version 3.3.0]. Generalised Additive Models (GAMs) are a type of Generalised Linear Model that use smooth functions to model nonlinear effects of continuous predictors. The 'mixed' in GAMMs refers to the inclusion of random effects in addition to fixed effects.

GAM is a well-established method of analysis that is increasingly being used in the cognitive and language sciences, and has been applied to EEG data [30, 31, 32, 33], eye movements in reading [34], reaction times [35, 36], articulography [37, 38], acoustic analysis [39], temporal clustering of sociolinguistic variants [40] and dialectology [41]. Recently, it has also been applied to eye movement data in the visual world paradigm [10]. GAMMs have also been used to analyse single-image eye movement data [42] and pupilometry [43].

GAMMs are a valuable method for analysing visual world fixation data for several reasons, including their ability to capture nonlinear changes in eye movements over the course of the trial and/or over the course of the experiment, the inclusion of random effects to deal with repeated measures, and methods for dealing with autocorrelation [10]. An important aspect of eye-tracking data is how fixations change over time. In experimental data sets, and especially time series data, autocorrelation can occur between data points [44]. In the `mgcv` package, functions have been implemented to deal with autocorrelation in GAM models.

Eye movement data were modelled as a continuous predictor of Euclidean distance of fixations from the centre of the target image. Because this gradient measure of distance includes data points where fixations have not reached the target image or fall between images, it is more likely to pick up on uncertainty effects, such as undershooting, hesitant or inaccurate oculomotor movements due to low activation or competing activations, compared to a categorical measure of within or outside the interest area [10]. All predictors of interest were entered into the initial model, and predictors that did not contribute to model fit were removed. Model comparison was conducted by means of χ^2 tests of fREML scores, using the `compareML` function in the `itsadug` package [45, version 2.2] in R. Because we were interested in the time course of fixations over the course of the trial, a continuous predictor of *time* was included. Data was downsampled to 50 Hz to reduce autocorrelation between data points. A 3200 ms window was selected for analysis, from 200 ms prior to to 3000 ms after auditory stimulus presentation. To test whether there was a learning effect over the course of the experiment, the model included a continuous predictor of trial, centred around 0 (*centred trial*). To determine whether participants were using pitch as a cue to distinguish between target and competitor images, the model included a continuous predictor of pitch, also centred around zero (*centred pitch*). The centred values ranged from -5.5 to 5.5, with the distribution peaks at -3 and 3. Distribution variance was modelled as a two-level factor, low-variance and high-variance. Previous research with the visual world paradigm has shown that the location of the target object on the screen significantly affects eye movement behaviour [46, 10]. Therefore, a smooth for *target position* over time was included as a control variable, a factor with four levels: top-left, top- right, bottom-left and bottom-right. Distance between the target and competitor was also included as a control variable. Random smooths for subject over time and subject over trial were included to account for differences between individual participants.

The initial model included intercepts for the two factor variables, variance condition and target position, and smooths (for each of the main effects, time, centred trial and centred pitch) and nonlinear regression lines[2] (for each two- and three-way interaction) for each level of condition. A smooth over time was also included for each level of target position. Random effects were modelled with shrunk factor smooths. After running the model, the model residuals were examined to check for autocorrelation. An AR1 model was included to account for autocorrelation in the residuals.

4. Results

Model comparisons showed that model fit was improved by including smooths for centred trial by condition ($p < .001$); target position over time ($p < .001$); and nonlinear regression smooths for time by trial by condition ($p < .01$); trial by pitch by condition ($p < .001$); time by target-competitor distance ($p < .05$); time by pitch by condition ($p < .01$); and trial by time by pitch by condition ($p < .001$). The model summary of the best fit model is shown in the Appendix.

Figure 2 shows the difference between the high-variance and low-variance conditions (high minus low) over time for four representative centred pitch values, 3.5 (top left panel), 1.5 (top right panel), -3.5 (bottom left panel) and -1.5 (bottom right panel). To examine the results of exposure to the two distributions, a trial near the end of the experiment (trial 230) is selected. Where the line is above 0 indicates when fixations were

[2] The partial effects tensors are modelled with the ti() function in the `mgcv` package.

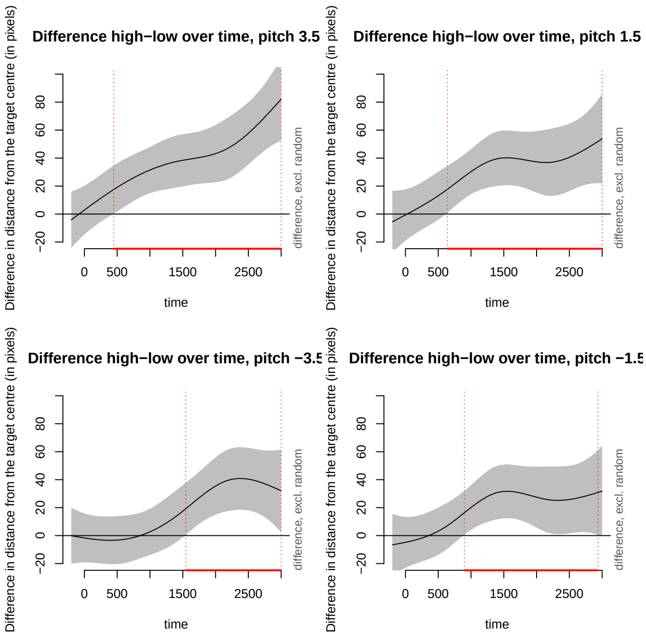

Figure 2: *Smooth of the difference between the high-variance and low-variance conditions over time for centred pitch values 3.5 (top left), 1.5 (top right), -3.5 (bottom left) and -1.5 (bottom right). Time (ms) is on the x-axis. The difference (in pixels) between conditions (high minus low) is on the y-axis. Trial is set to 230 (centred trial 109.5). Random effects are removed from these plots.*

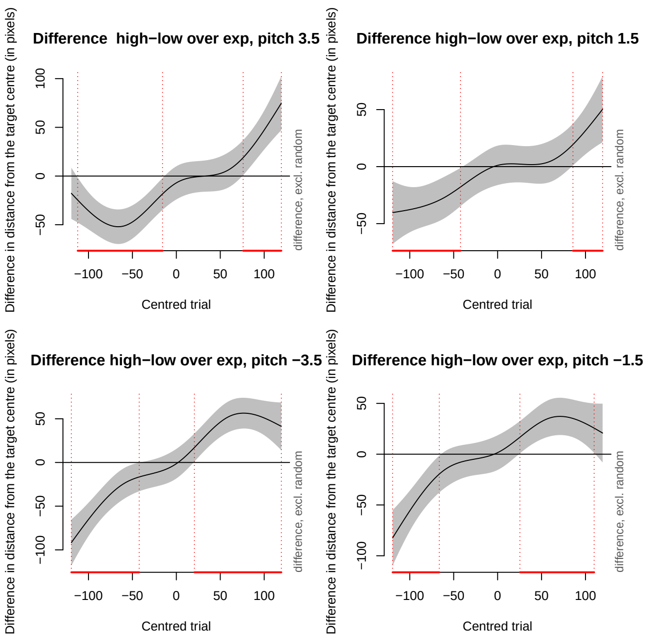

Figure 3: *Smooth of the difference between the high-variance and low-variance conditions over the course of the experiment for centred pitch values 3.5 (top left), 1.5 (top right), and -3.5 (bottom left) and -1.5 (bottom right). Trial is on the x-axis and is centred around 0. The difference (in pixels) between conditions (high minus low) is on the y-axis. Time is set to 2500 ms. Random effects are removed from these plots.*

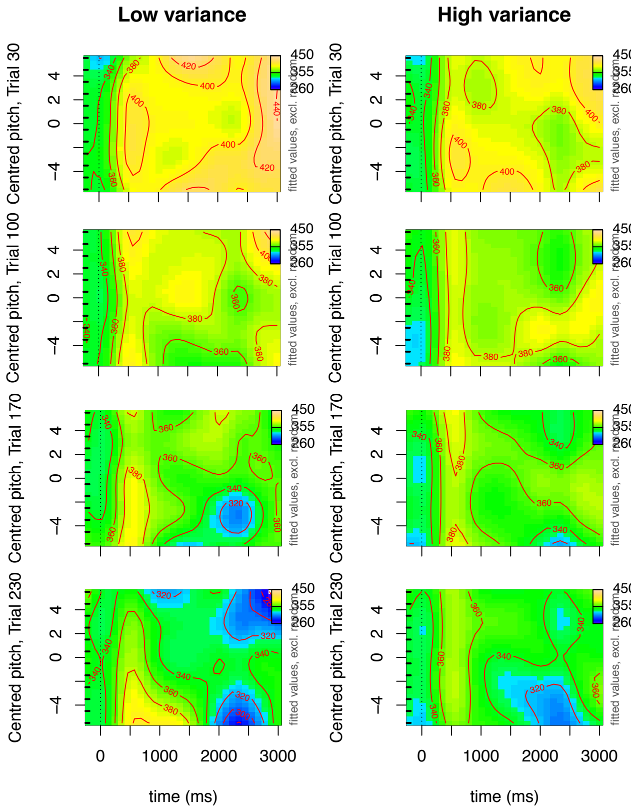

Figure 4: *Topographic maps of the model fit for the best fit model of Euclidean distance from the centre of target object in the low-variance (left panels) and high-variance conditions (right panels). Time (in milliseconds) is represented on the x-axis. Pitch is on the y-axis. Pitch is centred around 0, the category boundary. Positive values correspond the the high tone, negative values to the mid tone. Distribution peaks were at 3 and -3. Distance of fixations from the target object is plotted on the z-axis and is colour coded. Higher values (warmer colours) indicate greater distance; lower values (cooler colours) indicate a smaller distance. The key in the top-left corner shows the corresponding distance (in pixels) and z-limits. Random effects are excluded from this plot.*

significantly closer to the target in the low-variance compared to high-variance condition. For both high and mid tones, fixations are significantly closer to the target in the low-variance condition, compared to the high-variance condition. For the high tone (pitch 3.5 and 1.5), this effect emerges early and increases over the remainder of the trial period. Interestingly, there is a nonlinear effect of pitch over time: the difference between conditions is greater and emerges earlier for the stronger cue, the higher pitch value (3.5), compared to the pitch value closer to the category boundary (1.5). This is similar to the nonlinear effect of pitch value that Nixon and colleages [10] found for native Cantonese listeners, reflecting the ‘perceptual magnet effect’. For the mid tone (pitch -3.5 and -1.5), the low- and high-variance conditions diverge later than the high tones: around 1 second after auditory stimulus presentation for the mid tone, compared to around 500 ms for the high tones.

The difference between the high-variance and low-variance conditions over the course of the experiment is shown in Figure 3 for four representative pitch values (-3.5, -1.5, 1.5 and 3.5) at 2500 ms. Interestingly, at the beginning of the experiment, the distance from the target is greater in the low-variance condition than the high-variance condition for all pitch values. However, as the experiment progresses, the distance gets smaller, and by the end of the experiment, the distance from the target is significantly smaller in the low-variance condition.

The summed effects are shown in Figure 4. The figure shows a topographic plot of the Euclidean distance of fixations from the target object in the low-variance (left panels) and high-variance conditions (right panels). Time (ms) is on the horizontal axis. Centred pitch is on the vertical axis: positive pitch values correspond to the high tone and negative values to the mid tone. Distance from the target (in pixels) is on the z-axis and is colour-coded. Higher values (warmer colours) indicate fixations were further from the target image; lower values (cooler colours) indicate fixations were closer to the target image. The key in the top right corner indicates the corresponding values and z-limits. The panel rows show snapshots of trials throughout the experiment.

At the beginning of the experiment, fixations are further from the target in the low-variance condition, compared to the high-variance condition, for much of the trial. This effect lessens and has largely disappeared by trial 100. For the remainder of the experiment, fixations become gradually closer to the target in the low-variance condition. This becomes significant earlier in the low pitch values, as seen in trial 170. By the end of the experiment, fixations are significantly closer to the target in the low-variance condition for both the low and high pitch values.

5. Discussion

The present study investigated the effects of within-category acoustic variance on audio-visual processing during non-native acquisition of a new acoustic dimension, that is, pitch (f0) in a lexical tone contrast. Participants saw pictures of common objects and heard minimal word pairs, differing only in lexical tone. The tones were based on two Cantonese level tones, which are distinguished by pitch height. Auditory stimuli were sampled from pitch continua corresponding to the words. Stimuli were sampled according to a bimodal distribution. The critical manipulation was the statistical variance of the distribution, i.e. the amount of acoustic variability within the critical contrastive dimension, pitch. Participants heard either a *high-variance* or a *low-variance* distribution. Based on earlier studies investigating effects of variance in native language processing [9, 10], we predicted that acquisition of the pitch cue would be better in the low-variance condition. GAMM models of eye movements showed that when variance was low, participants learned to use the pitch cue better over the course of the experiment. The Euclidean distance between fixations and the centre of the target picture reduced over the course of the experiment and, by the end of the experiment, was lower in the low-variance condition, compared to the high-variance condition.

The results demonstrate an interesting interplay between auditory and visual cues. As variance in the auditory stimuli increased, distance of fixations from the target also increased. The increased distance is probably due to scanning the images or fixating the competitor. This suggests that as the reliability of the auditory cues decreased, participants increased their search for visual cues, looking around more for verification from the images. This was despite the fact that the image stimuli were kept constant between conditions. This suggests that as perceivers, we adapt to the current input, including cross-modal adjustments to make optimal use of multisensory input.

This result also provides new evidence that within-category acoustic variance shapes nonnative acoustic cue acquisition. Previous studies have shown that acoustic variance affects native speech perception, with increased variance leading to increased perceptual uncertainty [9, 10]. The present results show that the same mechanism can also help shape acquisition of a new acoustic dimension not present as a lexical contrast in the native language.

Cue variance has been investigated previously in native Japanese listeners' learning of the English /l/-/r/ contrast [47]. Many native Japanese listeners have trouble attending to the third formant (F3) cue - which native English listeners tend to use to distinguish /l/ and /r/ - and rely instead on the less reliable second formant (F2). Using a video game as training over several days, Lim and Holt[47] found that by presenting stimuli with high variability in the F2 cue dimension and low variability in the F3 dimension, participants' categorisation accuracy significantly increased and cue weighting shifted towards F3. The present study further contributes to our understanding of the effects of acoustic cue variance in several ways. Firstly, it directly compared the effects of high- vs. low-variance; the Lim and Holt study compared the effects of high-variance training to a control condition that did not involve exposure to English words. Secondly, the present study investigated the very early stages of acquisition , while the participants in the Lim and Holt study were proficient English speakers. Lim and Holt's participants had been studying English for at least 12 years and had lived in an English-speaking environment for up to 2.5 years. The present study investigated acquisition of a new cue dimension, not encountered before in a lexical contrast. Rather than improving an already partially acquired contrast, participants in the present study were experiencing both the language and the tonal contrast for the first time. Thirdly, the present study used approximately Gaussian distributions, whereas the Lim and Holt study used flat distributions - four steps of F2 and two steps of F3 - presented at equal frequency. Therefore the present study makes an important contribution by directly testing effects of distributional variance in a Gaussian distribution on acquisition of a new acoustic dimension.

While several recent studies have emphasised the facilitative effect of variability on learning [48, 16, 15, 17, 49, 12, 13], it is important to distinguish between within-category acoustic variance in the critical dimension and variability in non-contrastive dimensions. Variability can lower cue weighting. If that variability is in a contrastive dimension, it may hinder discrimination.

6. Acknowledgements

This research was made possible by an Endeavour Research Fellowship to the first author.

A. Appendix: Model summary

A. Parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	322.4100	4.6718	69.0122	< 0.0001
Condition=low variance	3.3199	5.5110	0.6024	0.5469
Target position = top right	15.4139	2.8761	5.3594	< 0.0001
Target position = bottom left	22.0030	2.8691	7.6690	< 0.0001
Target position = bottom right	39.1626	2.8718	13.6368	< 0.0001
B. Smooth terms	edf	Ref.df	F-value	p-value
s(Trial, Cond = low)	1.0354	1.0607	33.6331	< 0.0001
s(Time, Target position = top left)	6.2715	7.2789	2.3811	0.0098
s(Time, Target position = top right)	8.0449	8.7142	11.7981	< 0.0001
s(Time, Target position = bottom left)	1.0793	1.0968	0.0068	0.9499
s(Time, Target position = bottom right)	8.5727	8.9101	19.5311	< 0.0001
ti(Trial, Time, Cond = low	1.0069	1.0136	89.4874	< 0.0001
ti(Trial, Pitch, Cond = low	2.5555	3.0838	3.8280	0.0088
s(Time, Targ-Comp distance)	23.5304	26.4677	14.6500	< 0.0001
ti(Time, Pitch, Cond = low	5.1928	6.8190	3.2062	0.0024
ti(Time, Trial, Pitch, Cond = low	35.5535	45.8785	2.6620	< 0.0001
s(Trial, Subject)	65.8812	349.0000	0.3793	< 0.0001
s(Time, Subject)	192.5677	349.0000	1.6864	< 0.0001

Model summary for distance from target. Trial and Pitch are centred around 0. In smooth terms, s=smooth, ti=nonlinear regression line (tensor interaction).

B. References

[1] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[2] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, pp. 233–277, 1995.

[3] J. Maye and L. Gerken, *Learning phonemes without minimal pairs*. Proceedings of the 24th Annual Boston University Conference on Language Development, 2000.

[4] J. Maye, D. Weiss, and R. Aslin, "Statistical phonetic learning in infants: Facilitation and feature generalization," *Developmental Science*, vol. 11, no. 1, 2008.

[5] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, 2002.

[6] K. Wanrooij, P. Boersma, and T. L. van Zuijen, "Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study," *Frontiers in Psychology*, vol. 5., 2014.

[7] P. Escudero, T. Benders, and K. Wanrooij, "Enhanced bimodal distributions facilitate the learning of second language vowels," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, 2011.

[8] K. Wanrooij, P. Escudero, and M. E. Raijmakers, "What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning," *Journal of Phonetics*, vol. 41, no. 5, pp. 307–319, 2013.

[9] M. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. A. Jacobs, "Perception of speech reflects optimal use of probabilistic speech cues," *Cognition*, vol. 108, no. 3, pp. 804–809, 2008.

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

[11] ——, "Eye movements reflect acoustic cue informativity and statistical noise," in *ExLing 2015: Proceedings of the International Conference of Experimental Lingustics*, A. Botonis, Ed., 2015, pp. 54–57.

[12] G. C. Rost and B. McMurray, "Speaker variability augments phonological processing in early word learning," *Developmental Science*, vol. 12, no. 2, pp. 339–349, 2009.

[13] ——, “Finding the signal by adding noise: The role of non-contrastive phonetic variability in early word learning,” *Infancy*, vol. 15, no. 6, pp. 608–635, 2010.

[14] A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada, and Y. Tohkura, “Training Japanese listeners to identify English /r/ and /l/: Iv. some effects of perceptual learning on speech production,” *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2299–2310, 1997.

[15] J. S. Logan, S. E. Lively, and D. B. Pisoni, “Training Japanese listeners to identify English/r/and/l: A first report,” *The Journal of the Acoustical Society of America*, vol. 89, no. 2, 1991.

[16] S. E. Lively, J. S. Logan, and D. B. Pisoni, “Training Japanese listeners to identify English/r/and/l/,” *II: The role of phonetic environment and talker variability in learning new perceptual categories. The Journal of the Acoustical Society of America*, no. 1242., 1993.

[17] R. A. Yamada, “Effect of extended training on /r/ and /l/ identification by native speakers of Japanese,” *The Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2391–2391, 1993.

[18] R. A. Rescorla, “Probability of shock in the presence and absence of cs in fear conditioning,” *Journal of comparative and physiological psychology*, vol. 66, no. 1, pp. 1–5, 1968.

[19] ——, “Pavlovian conditioning: It’s not what you think it is.” *American Psychologist*, vol. 43, no. 3, p. 151, 1988.

[20] A. Wagner and R. Rescorla, “Inhibition in pavlovian conditioning: Application of a theory,” *Inhibition and learning*, pp. 301–336, 1972.

[21] J. D. O’Connor, L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, “Acoustic cues for the perception of initial/w, j, r, l/in english,” *Word*, vol. 13, no. 1, pp. 24–43, 1957.

[22] A. J. Lotto, M. Sato, and R. L. Diehl, “Mapping the task for the second language learner: The case of japanese acquisition of/r/and/l,” *From sound to sense*, vol. 50, no. 2004, pp. C381–C386, 2004.

[23] P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus, “Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models,” *Journal of Memory and Language*, vol. 38, no. 4, pp. 419–439, 1998.

[24] C. Kiriloff, “On the auditory perception of tones in mandarin,” *Phonetica*, vol. 20, no. 2-4, pp. 63–67, 1969.

[25] P. Boersma and D. Weenink, “Praat (version 5.5),” 2012.

[26] X. Lin and D. Zhang, “Inference in generalized additive mixed models using smoothing splines,” *Journal of the Royal Statistical Society*, vol. 61, no. 7, p. 381, 1999.

[27] S. Wood, *Generalized additive models: an introduction with R*. CRC press, 2006.

[28] ——, “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,” *Journal of the Royal Statistical Society (B)*, vol. 73, no. 1, pp. 3–36, 2011.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[30] C. de Cat, E. Klepousniotou, and H. Baayen, “Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English,” *Frontiers in Psychology*, vol. 6, p. 77, 2015.

[31] ——, “Electrophysiological correlates of noun-noun compound processing by non-native speakers of English,” *ComAComA 2014*, pp. 41–52, 2014.

[32] J. S. Nixon, J. van Rij, X. Q. Li, and Y. Chen, “Cross-category phonological effects on ERP amplitude demonstrate context-specific processing during reading aloud,” in *ExLing 2015: Proceedings of the International Conference of Experimental Lingustics*, A. Botonis, Ed., 2015, pp. 50–53.

[33] A. Tremblay and A. Newman, “Modelling non-linear relationships in ERP data using mixed-effects Regression with R examples,” *Psychophysiology*, vol. TBA, pp. 1–16, 2014.

[34] S. Vajjala, D. Meurers, A. Eitel, and K. Scheiter, “Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts,” 2016.

[35] L. B. Feldman, P. Milin, K. W. Cho, F. Moscoso del Prado Martín, and P. A. O‘Connor, “Must analysis of meaning follow analysis of form? a time course analysis,” *Frontiers in human neuroscience*, vol. 9, 2015.

[36] Pham, Hien, and H. R. Baayen, “Semantic relations and compound transparency: A regression study in CARIN theory,” *Psihologija*, vol. 46, no. 4, pp. 455–478, 2013.

[37] D. Arnold, P. Wagner, and H. Baayen, “Using generalized additive models and random forests to model German prosodic prominence,” *Proceedings of Interspeech 2013*, pp. 272–276, 2013.

[38] F. Tomaschek, M. Wieling, D. Arnold, and R. H. Baayen, “Word frequency, vowel length and vowel quality in speech production: an EMA study of the importance of experience.” in *Interspeech*, 2013, pp. 1302–1306.

[39] S. Kawase, *Examination of the role of native speech rhythm in non-native speech production and its perception (Doctoral dissertation, University of Western Sydney)*. Australia, 2017.

[40] M. Tamminga, C. Ahern, and A. Ecay, “Generalized additive mixed models for intraspeaker variation,” *Linguistics Vanguard*, vol. 2, no. s1, 2016.

[41] M. Wieling, S. Montemagni, J. Nerbonne, and R. H. Baayen, “Lexical differences between tuscan dialects and standard italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling,” *Language*, vol. 90, no. 3, pp. 669–692, 2014.

[42] J. van Rij, B. Hollebrandse, and P. Hendriks, “Children’s eye gaze reveals their use of discourse context in object pronoun resolution.” in *Empirical perspectives on anaphora resolution: Information structural evidence in the race for salience*, A. Holler, C. Goeb, and K. Suckow, Eds. Berlin, Walter de Gruyter, 2016.

[43] K. Lõo, J. van Rij, J. Järvikivi, and R. H. Baayen, “Individual differences in pupil dilation during naming task,” in *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, A. Papafragou, D. Grodner, D. Mirman, and J. Trueswell, Eds. Austin, TX: Cognitive Science Society, 2016, pp. 550–555.

[44] H. Baayen, S. Vasishth, R. Kliegl, and D. Bates, “The cave of shadows: Addressing the human factor with generalized additive mixed models,” *Journal of Memory and Language*, vol. 94, pp. 206–234, 2017.

[45] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, “itsadug: Interpreting time series and autocorrelated data using gamms,” 2016, r package version 2.2.

[46] D. Dahan, M. K. Tanenhaus, and A. P. Salverda, “The influence of visual processing on phonetically-driven saccades in the ‘visual world’ paradigm.” in *Eye movements: A window on mind and brain*, 2007, pp. 471–486.

[47] S.-j. Lim and L. L. Holt, “Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization,” *Cognitive science*, vol. 35, no. 7, pp. 1390–1405, 2011.

[48] A. R. Bradlow and T. Bent, “Perceptual adaptation to non-native speech,” *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[49] J. S. Allen, J. L. Miller, and D. DeSteno, “Individual talker differences in voice-onset-time.” *Journal of the Acoustical Society of America*, 2003.