The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception

Jessie S. Nixon[a], Jacolien van Rij[b], Peggy Mok[c], R. Harald Baayen[b], Yiya Chen[d]

[a] *Department of Pyschology, University of Potsdam, Potsdam, Germany*
[b] *Quantitative Linguistics Group, University of Tubingen, Germany*
[c] *Chinese University of Hong Kong, Hong Kong*
[d] *Leiden University Centre for Linguistics, Leiden, The Netherlands*

Abstract

Determining a speaker's message requires discrimination between discrete alternatives based on inherently noisy, non-discrete acoustic cues. This entails that there is always some degree of uncertainty in perception of speech. Despite well-documented effects of gradient sensitivity to within-category acoustic variation and growing interest in statistical effects in speech perception, very little is yet known about the time course of perceptual uncertainty in speech percpetion. Two visual world eyetracking experiments investigated how changes in acoustic cue values and the amount of within-category acoustic variation affect perceptual certainty during perception of Cantonese speech sound contrasts. Participants saw four pictures on screen and heard an auditory stimulus. Critical pictures were of word pairs that were identical except for initial consonants (Experiment 1), which were unaspirated (bou2, 'treasure') or aspirated (pou2 'shop'); or tones (Experiment 2), which were high (jin1, 'carpet') or mid (jin3, 'arrow'). Auditory stimuli consisted of a continuum of 12 tokens of increasing VOT (Experiment 1) or pitch (Experiment 2). The number of times participants heard each token followed a bimodal distribution. The amount of within-category variation differed between conditions: low-variance versus high-variance. Eye movements were monitored until participants selected a picture by clicking on it. The Euclidean distance of fixations from the target and competitor pictures was analysed using Generalised Additive Mixed Modelling. Results showed that the distance of fixations from target and competitor pictures over the course of the trial varied as a function of VOT value (Experiment 1) or pitch (Experiment 2), providing evidence for gradient, nonlinear sensitivity to cue values. Interestingly, the time course of these effects differed between the target distance and competitor distance models. Moreover, in both experiments, the effect of the acoustic cue value significantly interacted with how much acoustic variation participants heard. In the VOT models, fixations were closer to the competitor in the high-variance condition. However, in the pitch models, the category boundary was shifted and the opposite pattern emerged. This indicates that the shape of the acoustic cue distribution plays an essential role in perceptual processing. With little statistical variance, speech sound representations become more robust. Yet they also lead to greater uncertainty in the face of unexpected speech tokens. In addition, the pattern of effects over time suggests that the effect of statistical distribution (cue variance) suggests a global strategy in response to the level of uncertainty: as uncertainty increases, verification looks also increase.

Keywords:
discriminative learning, statistical learning, speech perception, Cantonese, lexical tone

1. Introduction

Human listeners rely on highly variable, non-discrete acoustic information to discriminate between the different possible messages a speaker might intend to convey in an utterance. The question of how acoustic variation affects perceptual uncertainty during speech processing is an intriguing one. Listeners use variation *between* speech sounds to discriminate between words and messages. For example, in English, voice onset time (VOT) is longer in voiceless sounds (e.g. the /p/ in *pat*) than voiced sounds (e.g. the /b/ in *bat*). VOT is the time between the release burst of the consonant and the onset of voicing in the vowel, and is the most important cue for distinguishing voiced from voiceless sounds in English. However, there is also a considerable amount of variation *within* speech categories. For example, the mean VOT of English /p/ is 58 ms (Lisker and Abramson, 1964), but /p/ can be produced with a range of VOTs. Acoustic variation can even occur in productions of the same word by the same speaker in the same phonetic context under controlled lab settings (Newman, Clouse, and Burnham, 2001) and increases greatly across speakers (Ladefoged and Broadbent, 1957), in different phonetic contexts (Nixon, Chen, and Schiller, 2015a) and even depending on word frequency (Gahl, 2008).

The high degree of variation in the acoustic signal means that there is nothing in the speech stream that conclusively points to particular meanings, words or even phonemes. The listener can only use cues to assess the likelihood that a speaker intended one message rather than another, meaning that there is always some degree of uncertainty in the process of speech perception. In addition to the issue of within-category acoustic variation, listeners also face the challenge of changes in the whole statistical distribution of acoustic cues in particular contexts, for example, when encountering a new speaker or accent. Recent evidence suggests that both variation in acoustic cues (McMurray, Tanenhaus, and Aslin, 2002; McMurray, Aslin, Tanenhaus, Spivey, and Subik, 2008a; McMurray, Tanenhaus, and Aslin, 2009) and changes in the statistics of cue distributions affect listeners' level of perceptual uncertainty during speech perception (Clayards, Tanenhaus, Aslin, and Jacobs, 2008; Escudero, Benders, and Wanrooij, 2011; Escudero and Williams, 2014; Wanrooij, Boersma, and van Zuijen, 2014; Wanrooij, Escudero, and Raijmakers, 2013; Liu and Kager, 2011). The present study aims to contribute to our understanding of perceptual uncertainty in speech perception by examining the time course of effects of a) variation in acoustic cues and b) the degree of variance in statistical distributions of acoustic cues in native Cantonese listeners. In this paper, we use the term *variance* to describe, in a given speech sample, the amount of acoustic variation there is *within* a speech category. This term refers to the degree to which acoustic values spread out from the mean of the distribution of that speech category. A variance of zero means that all values are identical.

Early accounts claimed that speech perception was 'categorical' in that listeners were unable to detect within-category acoustic variation, and only able to detect variation when it occurred across boundaries. Evidence in favour of this claim came from studies showing sharp categorisation functions between speech categories, and chance-level performance in detecting within-category acoustic differences (e.g. Liberman, Harris, Hoffman, and Griffith, 1957; Ferrero, Pelamatti, and Vagges, 1982; Schouten and van Hessen, 1992). However, more recently, abundant evidence has accumulated demonstrating listeners' remarkable sensitivity to fine-grained phonetic information, given the appropriate task (e.g. Andruski, Blumstein, and Burton, 1994; Dahan, Magnuson, Tanenhaus, and Hogan, 2001; Marslen-Wilson and Warren, 1994; Utman, Blumstein, and Burton, 2000; McMurray et al., 2008a, 2002, 2009).

Moreover, not only are listeners sensitive to gradient acoustic variation, they are able to rapidly adapt to context-specific changes in acoustic characteristics of speech, based on the effectiveness of a particular dimension for speech recognition (Idemaru and Holt, 2011, 2014). Relatedly, listeners are also sensitive to *frequency* distributions of acoustic cues. One line of research has investigated how the acoustic distance between speech categories affects categorisation accuracy. For example, several studies have shown that when trained with a unimodal distribution (no distance between categories), participants are less likely to categorise the endpoints of a distribution as different, compared to when they are trained with a bimodal distribution (Maye and Gerken, 2000; Maye, Weiss, and Aslin, 2008; Liu and Kager, 2011; Escudero and Williams, 2014; Wanrooij et al., 2014; Maye, Werker, and Gerken, 2002). Even when trained with a bimodal distribution, a greater distance between categories improves categorisation accuracy, compared to training with a bimodal distribution with a small distance between categories (Escudero et al., 2011; Wanrooij et al., 2013).

Much of the research in adult distributional learning has focused on the acquisition and development of non-native contrasts. For example, a series of recent studies has investigated the effects of statistical distributions on non-native perception of Dutch vowel contrasts (Escudero et al., 2011; Gulian, Escudero, and Boersma, 2007; Wanrooij et al., 2013). Motivated by the observation that infant and foreigner directed speech has a ‘stretched’ vowel space, Escudero et al. (2011) investigated effects of the acoustic interval between vowel categories in second language acquisition. They used *natural bimodal* (reduced acoustic interval; i.e. vowel categories were similar to each other) versus *enhanced bimodal* distributions (increased acoustic interval) to train Spanish learners to distinguish a Dutch vowel contrast. After two minutes of exposure natural bimodal or enhanced distributions, there was an increase in ‘correct’ categorisation, compared to the music (control) group. This increase only reached significance in the enhanced group.

Most studies of distributional learning in adults have used offline categorisation responses as the measure of learning. Categorisation measures provide information about the final outcome of the decision process; however, they do not provide information about online processing during perception itself. In discussions of effects on categorisation, it is often implicitly or explicitly assumed that assigning tokens to one category rather than two occurs because the two tokens were not discriminated. This assumption may not necessarily be justified. In a forced-choice categorisation task, regardless of the degree of uncertainty, or any gradient degree of goodness of fit with one category or another, the participant must make a binary choice. While it is interesting that factors such as cue distribution can affect even the final outcome of the decision process, examining the moment by moment online processing can tell us about how subtle differences in statistical distributions can affect the development of perceptual processes over time, prior to the decision process.

One interesting and innovative recent eyetracking study (Clayards et al., 2008) is, to the best of our knowledge, the only other study that has used online measures to investigate statistical processing of acoustic cues during perception of native speech contrasts. This study has examined how the *amount* of within-category acoustic variation affects perceptual certainty. Using the visual world paradigm (VWP; Allopenna, Magnuson, and Tanenhaus, 1998), Clayards et al. (2008) tested the hypothesis that greater variation in the acoustic signal would lead to greater perceptual uncertainty. Native English-speaking participants saw four pictures on screen, heard an auditory stimulus and were instructed to click on the picture of the word they heard. Critical picture stimuli consisted of pairs of words beginning with /b/ and /p/ (e.g. ‘beach’ and ‘peach’). Auditory stimuli consisted of a VOT continuum which spanned the word pair (e.g. from beach to peach). Presentation frequency of the tokens on the continuum always followed a bimodal distribution. However, the amount of within-category acoustic variation was manipulated between participants: participants heard either a high-variance or low-variance distribution of the acoustic stimuli.

In the analysis, the proportion of categorisation responses was calculated per participant per condition and for each token on the VOT continuum. Overall, the categorisation slope was shallower in the high-variance condition, indicating that with greater variation in the acoustic input, participants were less consistent in their assignment of cues to the contrastive categories. Eye movement data were also analysed for the six points on the continuum that had sufficient data points, three each for the /b/ and /p/ words. There was a significant effect of distribution condition for the /b/ words and a significant interaction between distribution condition and VOT token for the /p/ words. In both word types, the effect was carried by the VOT token closest to the category boundary; however, the trend was similar for all VOT tokens analysed: there were more looks to the competitor in the high-variance, compared to the low-variance condition. This provided evidence that the amount of variation in the acoustic signal has direct effects on speech perception: increased variance can lead to an increase in perceptual uncertainty.

Our understanding of how acoustic variance affects perceptual certainty could be enhanced by knowing at what point in time these effects come into play. While Clayards et al. (2008) examined the effects of acoustic cue variance on eye movements, the measure reported in their study was the proportion of looks over the whole trial. Information about the time course of effects is important for understanding the underlying mechanism. As listeners gain experience with the input distribution, does statistical information affect the early perceptual processes? Is uncertainty a global effect that influences eye movement behaviour from the onset of the trial? Or is the statistical information used only in the later decision process to discriminate between alternative candidates? The present study aims to address these questions by examining changes in eye movement patterns over the course of the trial, including nonlinear interactions between predictors over time.

Similarly, although listeners' ability to detect and respond to within-category variation is now well established, few studies have investigated the time course of its effects. One recent VWP study investigated 'lexical garden path' recovery in English (McMurray et al., 2009). This study used a VOT continuum to manipulate bilabial stop word-onsets, creating temporarily ambiguous words, such as 'barricade' versus 'parrakeet'. Although the study measured the time course of fixations, the main focus was to establish that sensitivity to VOT variation was gradient, rather than categorical. Therefore, the discussion of the time course mainly focused on establishing that effects of within-category differences in VOT persist over durations longer than a syllable, rather than establishing the point in time where different VOT values diverged.

The large majority of research investigating speech perception processes, in general, and sensitivity to cue values and cue distributions in particular, has been conducted on alphabetic, Indo-European languages, such as English. The present study examines speech perception by native speakers of a typologically very different language, Hong Kong Cantonese. Cantonese was selected for the present experiments in order to extend the investigation of perceptual uncertainty effects to a new set of speech sounds, which included both the previously-investigated temporal cue, VOT, as well as a suprasegmental cue, pitch (f0), in a lexical tone contrast. Cantonese has a complex tonal system, with six lexical tones (Bauer and Benedict, 1997; Wiener and Turnbull, 2015; Mok and Wong, 2010; Siddins and Harrington, 2015).[1] Three of these are level tones, in which the primary cue is pitch (f0) height. These level tones make Cantonese an ideal language for investigating distributional effects in tone processing. In addition to being a tonal language, Cantonese also differs from English in other important respects. Cantonese uses a logographic writing system, in which phonology is not explicitly represented. Each character represents a particular morpheme and is pronounced with a single syllable. The lack of explicit phonological representation influences the phonological awareness of Cantonese speakers, leading to more holistic processing and less awareness of low-level phonological changes (McBride-Chang, Bialystok, Chong, and Li, 2004). In addition, compared to English, due to its syllabic structure, Cantonese has a large number of homophones. This means that it is often necessary to rely on top-down context effects to a greater degree in Cantonese than in English. Such cross-linguistic differences call for investigation of typologically diverse languages in order to have a complete understanding of language-general mechanisms in speech perception.

The present study. The present study investigates the time course of perceptual uncertainty effects during perception of Cantonese tonal and segmental speech sound contrasts. Two manipulations were expected to affect percep-

[1] The number of tones is sometimes reported as nine, including the checked tones.

tual uncertainty: the location of an acoustic cue along the cue continuum, in particular the distance from the category boundary; and the distribution condition, that is, amount of within-category acoustic variance in the signal. These questions were tested with two sets of models. The first examined the Euclidean distance of fixations from the centre of the target picture, and the second examined the Euclidean distance of fixations from the centre of the competitor picture.

We tested four main hypotheses. Since we know of no other similar study of Cantonese speech perception using VWP, we based these hypotheses on studies in English. The first was that the fixations would be further from the target and closer to the competitor picture the closer the acoustic cue values were to the category boundary. This prediction was based on a number of previous studies in English that have shown gradient effects of acoustic cue values using a VOT continuum (e.g. McMurray et al., 2008a; McMurray, Clayards, Tanenhaus, and Aslin, 2008b; McMurray et al., 2009). The second was that fixations would be further from the target and closer to the competitor in the high-variance condition, compared to the low-variance condition, similar to the results of Clayards et al. (2008).

Our third and fourth hypotheses relate to the time course of effects, in particular the time course of effects of the acoustic cue value and of acoustic cue variance. McMurray and colleagues (McMurray et al., 2008b, 2009) found that when English-speaking participants were presented with auditory stimuli from a VOT continuum, divergences in eye movements to target pictures began around 600 ms after stimulus presentation. Therefore, we expected to see effects of acoustic cue value start to emerge around 600 ms after presentation.

Regarding the time course of effects of acoustic variance, as far as we are aware, the present research is the first to investigate this question in any language. Therefore the study is largely exploratory in this respect. The time course of various other effects during speech perception has been investigated using VWP. For example, McMurray et al. (2008b) asked at what point asynchronous cues are integrated during speech perception. Their results showed that word-initial cues (voicing and formant transitions) influenced eye movements to target pictures earlier than cues that occurred later in the signal (vowel length), providing evidence for continuous integration of acoustic cues as the speech signal unfolds. Another study investigated the time course of effects of lexically-guided retuning of a fricative contrast. Mitterer and Reinisch (2013) found that effects of retuning (f-biased versus s-biased training) occurred very early, around 200 ms after frication onset. They argued that this was evidence that retuning occurs at the perceptual level, rather than affecting higher-order decision processes. The present study differs from Mitterer and Reinisch (2013) in that it does not require adjustment of category boundaries. Rather, it investigates participants‘ responses to higher or lower levels of uncertainty.

Finally, as the VWP involves both auditory perception and a visual component, we controlled for the effects of the location of the pictures on the screen in our analysis. The pictures were randomly assigned to a screen position on each trial. We expect that the vertical (top-bottom) and horizontal (left-right) position of the target and competitor pictures on the screen will influence the distance of fixations from these respective pictures over time.

In addition to testing these hypotheses, we also present a statistical modelling method (Generalised Additive Mixed Modelling, GAMM; Wood, 2006, 2011) that is well suited to analysis of eyetracking data. This is not a new statistical method; it has been used in the analysis of a wide variety of experimental paradigms investigating cognition of language, as well as other fields. However, as far as we are aware, it has not previously been applied to the analysis of fixation data from the four-field visual world eyetracking paradigm. GAMMs are well suited to analysis of data with a time component, because they allow for analysis of changes of a variable over time. They provide solutions to some of the challenges of analysing time series data, such as autocorrelation. They also allow for analysis of complex interactions (including over time) and nonlinear random effects. A description of the modelling method and some of its benefits will be returned to in the Method section.

2. Experiment 1 Voice onset time

2.1. Method

Participants. Thirty-seven native Cantonese-speaking undergraduate students from the Chinese University of Hong Kong participated in the experiment for payment. Participants were tested individually in a quiet room.

Experiment design and stimuli. The experiment design and stimuli were based on those presented in Clayards et al. (2008). Visual stimuli were picture pairs whose names began with either bilabial stops (‘b’, ‘p’) or alveolar affricates (‘j’, ‘ch’). The two members of each word pair were identical except

Table 1: Presentation frequency per variant per condition: each variant represents one step on the VOT continuum

		Number of iterations											
	Variant	1	2	3	4	5	6	7	8	9	10	11	12
Distribution	**Low-variance**	0	6	54	108	54	6	6	54	108	54	6	0
condition	**High-variance**	6	24	54	60	54	30	30	54	60	54	54	6

for the initial consonants, which were either unaspirated (bou3, 'cloth'; jun1 'brick') or aspirated (pou3, 'shop'; chun1, 'village'). Pictures were black-on-white line drawings.

All auditory stimuli were recorded by a male native speaker of Hong Kong Cantonese. Stimuli were then resynthesised into a 12-step VOT continuum using the Pitch-Synchronous-Overlap-and-Add (PSOLA) method in PRAAT (Boersma and Weenink, 2012), using the unaspirated token as the target for resynthesis. Increasing steps of aspiration were added following the stop or affricate burst before the onset of the vowel. The consonant duration ranged from 0 ms to 88 ms for the stops and 40 ms to 260 ms for the affricates. The vowel portion of the recorded syllables ranged from 432 ms to 571 ms. The number of times participants heard each step followed a bimodal distribution, with the two peaks of the distributions corresponding to the prototypical mean VOT for the unaspirated and aspirated stimuli, respectively (Cheung and Wee, 2008; Ng and Wong, 2009). Ten native Cantonese speakers also participated in a perception test which verified the stimuli. Table 1 shows the presentation frequency of each step on the continuum. Each condition contained 456 tokens, 76 for each word pair. All participants heard the same number of tokens; only the number of times they heard each token varied between conditions: high-variance versus low-variance distributions.

The experiment consisted of 456 experimental trials, divided into six blocks of 76 trials, with breaks between blocks. The order of presentation was pseudo-randomised for each participant.

Procedure. Participants sat at a comfortable viewing distance from the computer screen and wore an SR Eyelink II head mounted eye-tracker with a sampling rate of 500 Hz. Stimulus presentation and data acquisition were conducted using SR Research Experiment Builder computer software (2011;

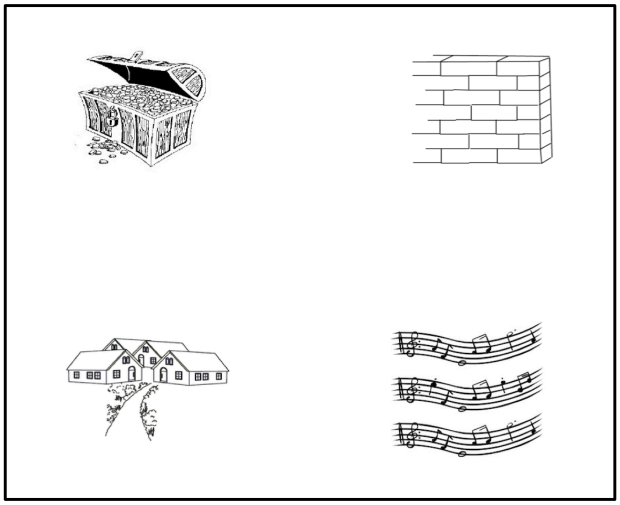


Figure 1: Sample screen display during stimulus presentation.

version 1.10.165). The session began with 12 familiarization trials in which participants saw the pictures and their corresponding written labels once each. This was followed by a practice block to familiarize participants with the experimental procedure. None of the experimental pictures or words were presented during the practice phase.

Each experimental trial began with drift correction to ensure accurate calibration of the equipment, followed by brief presentation (1000 ms) of four pictures, one in each quadrant of the screen (see Figure 1). The purpose of giving an advance preview of the stimuli was to reduce the time and likelihood of participants scanning the pictures at the beginning of the trial, and hence to reduce noise in the eye movement data. The display always contained two test items and two filler items. The location of the picture conditions on screen, as well as their relative location, was randomised to avoid strategic effects. The picture preview disappeared, replaced with a gaze-contingent fixation cross, which ensured participants were looking at the centre of the screen at the beginning of the critical trial period. The pictures reappeared and, simultaneously, one of the auditory stimuli was presented and participants chose the picture they thought most appropriate by clicking on it with the mouse. Eye movements were monitored from the onset of the preview until participants made a response. (Analysis was conducted on a shorter period, starting just prior to the auditory stimulus).

3. Analysis

Eye movement data were analysed using *Generalised Additive Mixed Modeling* (GAMM; Wood, 2006, 2011) using the `mgcv` package (version 1.8-7) conducted in R (version 3.2.2; R core team, 2015; www.r-project.org). GAMM is a type of Generalised Linear Modelling (GLM) that uses nonlinear smooth functions to model nonlinear effects for continuous predictors.

Generalised Additive Models[2] are a well-established method of analysis used with a wide range of psychological, psychophysiological and speech production data, ranging from EEG data (de Cat, Klepousniotou, and Baayen, 2014, 2015; Nixon, 2014; Nixon, van Rij, Li, and Chen, 2015b; Tremblay and Newman, 2014), reaction times (Feldman, Milin, Cho, Moscoso del Prado Martin, and O'Connor, forthcoming; Pham, Hien, and Baayen, 2013) and pupilometry (van Rij, Pya, van Rijn, Wood, and Baayen, in preparation) to articulography (Arnold, Wagner, and Baayen, 2013; Tomaschek, Wieling, Arnold, and Baayen, 2013) and dialectology (Wieling, Montemagni, Nerbonne, and Baayen, 2014). As far as we are aware, the present study is the first to apply GAMMs to the typical four-field visual world paradigm, although it has previously been to used in the analysis of single-field gaze data (van Rij, Hollebrandse, and Hendriks, in press).

There are several characteristics of GAMMs that make them particularly well suited to analysis of visual world paradigm eye movement data. Firstly, GAMMs drop the assumption of a linear relationship between dependent and independent variables. Assuming linearity when the relationship in the data is nonlinear can lead to failure to observe regularities or structure that do exist in the data (see Tremblay and Newman, 2014, for a discussion of the benefits of relaxing the linearity assumption in psychological research). Instead, GAMMs determine the linearity or degree of nonlinearity from the data itself. The method used for this is penalized iteratively re-weighted least squares (PIRLS; see Wood, 2006, for details). PIRLS determines the optimal linear or nonlinear equation for avoiding both over-fitting and over-generalizing of the model. Secondly, GAMMs allow for analysis of continuous variables and nonlinear interactions. This is an advantage for analysis of fixation data, as processing is often influenced by continuous predictors, such as time and,

[2]The 'mixed' in Generalised Additive Mixed Models refers to the inclusion of random effects, such as participant and item random effects in the present study, in addition to fixed effects. That is, a GAMM is a type of GAM that includes random effects.

in the present study, location on the acoustic continuum; importantly, often several predictors may interact. A third aspect of GAMMs that benefits VWP eye movement analysis is the inclusion of random effects. This allows the model to take into account that repeated measures are taken from participants and items without the need to average over them in the analysis. This is also an important means of reducing autocorrelation (see Baayen, van Rij, de Cat, and Wood, to appear; Baayen, Vasishth, Bates, and Kliegl, 2015, for a discussion of the benefits of GAMMs for reducing autocorrelation in language-related experimental data). Finally, a common problem in many experimental data sets, and particularly in data with a time series component, such as eye tracking, is that autocorrelation can occur between data points. In the `mgcv` package, methods have been implemented specifically to deal with autocorrelation (Baayen et al., to appear).

All predictors of interest were entered into a GAMM model. Predictors that did not contribute to model fit were eliminated. Model comparison was conducted using a χ^2 test of fREML scores in the `compareML` function in the `itsadug` package (version version 1.0.4; van Rij, Baayen, Wieling, and van Rijn, 2015) in R. Together with the model comparisons and model plots, the statistics provided by the model summaries were used to determine whether each predictor contributed to the variance explained by the model.

Fixation data were modelled as two separate continuous variables of Euclidean distance: distance from the centre of the target picture (*target distance*) and distance from the centre of the competitor picture (*competitor distance*). Figure 2 shows a sample trial as an illustration of the target distance measure. There are least two advantages to modelling the eye movement data in this way. Firstly, it allowed us to model the data as a gradient measure, rather than a binary variable with an arbitrary cut-off point. Because data points that fall short of the target picture or fall between two pictures are included, the distance measure is more likely to pick up on uncertainty effects, such as hesitant oculo-motor movements, undershooting the mark due to low activation or inaccurate movements due to competing activations. Secondly, the models are more robust, because more data is included. We initially ran models with the proportion of fixations on the target picture as the dependent variable. However, this led to artefacts in the early fixations due to insufficient data in the initial 200 ms of the trial. The distance measure solved this issue. Separate models were run for each of these dependent variables.

Because we were interested in the time course of processing over the

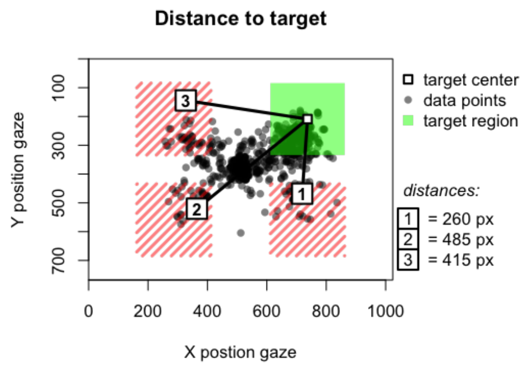


Figure 2: Illustration of the Euclidean-distance-from-target measure. This figure shows a random sample of data points from a trial with the target picture in the top right corner. Fixations 1, 2 and 3 are sample fixations from this trial. Note that the absolute X and Y coordinates on the figure axes are measured from the top left corner of the screen. However, the measure of interest (Euclidean distance) is measured from the centre of the target picture. For each fixation, the Euclidean distance (in pixels) from the centre of the target picture is calculated from the X (x-axis) and Y coordinates (y-axis). For a given fixation, a distance greater than 176 is outside the interest area and a distance of 125 or less is within the target picture interest area.

whole trial, from early perceptual processing to later decision processes, the predictor *time* was included. A 1400 ms time window from -200 ms (i.e. 200 ms prior to presentation of the auditory stimulus) to 1200 ms was selected for analysis. After this time, the number of data points became too few, as mean response time was approximately 1300 ms. An initial model was run with data downsampled to twenty milliseconds (50 Hz). However, inspection of the residuals of the first statistical model indicated that a moderate level of correlation remained between subsequent measurements. Therefore, to reduce autocorrelation further, forty millisecond (25 Hz) time bins were used.

VOT (Experiment 1) and *pitch* (Experiment 2) were modelled as continuous variables, centred around 0. The centred values ranged from -4.5 to 4.5, with the distribution peaks at -2.5 and 2.5. Distribution condition was modelled as a factor with two levels, low variance and high variance. As control variables, the location of the target on the screen was included in the target distance models, and location of competitor in the competitor distance models. This was a factor variable with four levels: top-left, top-right, bottom-left and bottom right. Changes over the course of the experiment were investigated by including a predictor of *trial*. However, this did not improve model fit, so was removed from the analysis.

The initial model included intercepts for condition (low- vs. high-variance) and target position, a nonlinear interaction[3] of centred VOT (or pitch) by condition over time and a nonlinear regression line[4] of target position over time. After running the models, the residuals were examined to determine the degree of remaining autocorrelation. We included an AR1 model to account for autocorrelation in the residuals with the *rho* parameter, which measures how much the residuals of the current data point are determined by the residuals at the previous data point. In GAMM models, *shrunk factor smooths* can be used to model random effects. They are the nonlinear equivalent of by-subject and by-item random slopes and intercepts in an LMM.

4. Results

4.1. Target distance model: distance of fixations from the target picture

4.1.1. Random effects

The best-fit model for target distance (Appendix A) includes trends over time as random effects per participant per target item. Random effects were modelled as a separate smooth for each participant-item pair in order to capture participants' variable responses to the different items. Each *random wiggly curve* represents the difference in eye movement behaviour over time for a particular participant for a particular item compared to the average.

4.1.2. Effects of voice onset time value on target distance

The best-fit model included a smooth of centred VOT over time (Appendix A), which significantly contributed to variance explained in the model (F(65.706, 476634.3)=98.5). Estimated effects of VOT over time are shown in the top row of Figure 3. In the figure, time is represented on the horizontal axis. Centred VOT is on the vertical axis. Category means are at VOT -2.5 (for the unaspirated stimuli, e.g. bou2) and 2.5 (for the aspirated stimuli, e.g. pou2). The distance of fixations from the centre of the target picture is plotted on the z-axis, represented by colour codes. Higher values (shown in yellow) indicate a relatively greater distance from the target; lower values

[3]In the mgcv package, this type of nonlinear interaction is modelled with the te() function. It includes all main effects and interactions.

[4]This nonlinear regression line is modelled with the ti() function. In the mgcv package, the ti() function can be used to model partial effects, including nonlinear regression lines and nonlinear interactions without the main effects or lower-level interactions.

(shown in blue) indicate a relatively shorter distance. The key at the bottom left of each panel shows the corresponding pixel values and z-limits for each model plot. Note that the range is different between the target distance plot (top row) and the competitor distance plots (bottom row): the target plot ranges between 80 and 320 pixels, while competitor plots range between 200 and 440 pixels. The scale is the same. Random effects are excluded from these plots. A plot of the raw data for target distance in Experiment 1 is provided in Appendix E (upper panel). To assist with interpretation, particularly for readers who are unfamiliar with topographic plots, Appendix G provides an illustration of the mapping between the topographic plot and a line plot of the raw data.

The plot indicates that changes in eye movements over the course of the trial occur differently at different points on the VOT continuum. Over the course of the trial period, the pattern of eye movements increasingly reflects the differences in VOT values, with differential fixation behaviour at central and outer regions of the continuum. Prior to and for the first 200 ms after presentation of the auditory stimulus, the plot shows a flat distribution. Fixations are consistently around 280 pixels from the target; that is, the distance between the centre of the target and the fixation cross. At around 200 ms, the eyes begin to move away from the fixation cross. From around 400 ms, the distance steadily decreases until the end of the trial. Differences between VOT values begin to emerge around 400-500 ms. The decrease in distance from the target occurs more rapidly at the distribution peaks and peripheries, compared to the central values. The difference in distance from the target remains throughout the trial, with a consistently greater distance for the central VOT values, compared to the outer values from around 450 ms until the end of the trial.

4.1.3. Effects of distribution condition on target distance

The VOT-by-condition interaction was not significant. Initial models, which did not include a rho parameter, hinted that there might be an effect of distribution condition. However, once autocorrelation was reduced by including rho, the χ^2 test of fREML scores showed that including an interaction with distribution condition no longer significantly improved fit. In the upper panel of Figure 3 condition is collapsed.

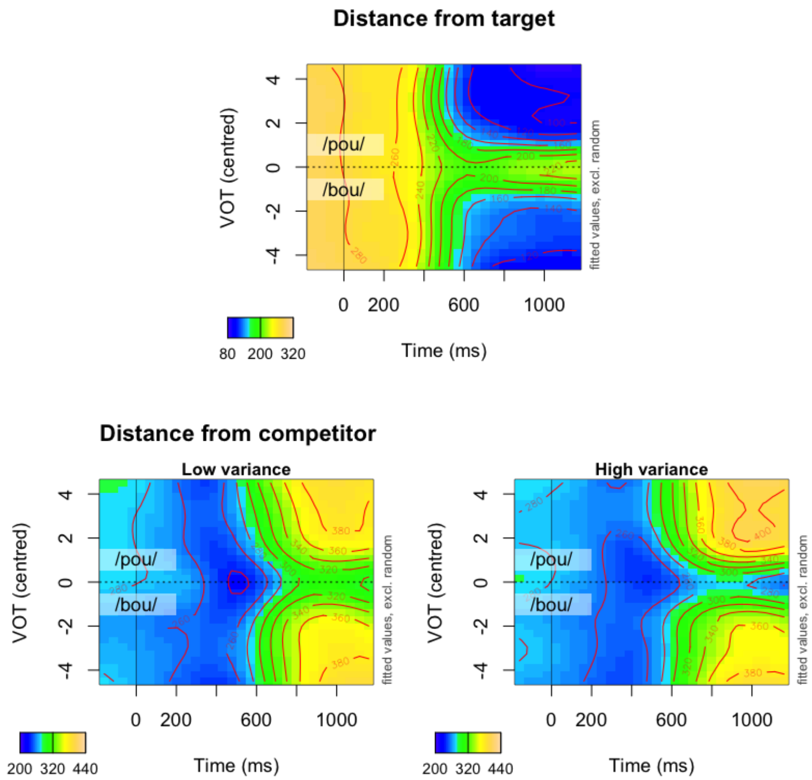


Figure 3: Topographical maps for the VOT models in Experiment 1. Top row: model fit for the best fit model of Euclidean distance from the target picture. The predictor Target Position is 'top left' in this plot (see the left panel of Figure 4 for the effects of Target Position). Bottom row: model fit for the best fit model of Euclidean distance from the competitor picture for the low-variance (left panel) and high-variance conditions (right panel). The predictor Competitor Position is 'top left' in these plots (see the right panel of Figure 4 for the effects of Competitor Position). All plots: Estimated effects are in pixels. Time (in milliseconds) is represented on the x-axis. Voice onset time (VOT) is on the y-axis. VOT is centred around 0, the category boundary. The negative VOT values correspond to unaspirated stimuli (e.g. bou), the positive values to aspirated stimuli (e.g. pou). Category means are at VOT -2.5 and 2.5, respectively. Distance is plotted on the z-axis, represented by colour codes. Higher values (yellow areas) indicate a relatively greater distance; lower values (blue areas) indicate a relatively smaller distance. The key in the bottom left corner shows corresponding pixel values and the z-limits. Note that the range differs between the surface plots for target and competitor model plots; target plots (top row): 80 to 320 pixels; competitor plots (bottom row) 200 to 440 pixels. (The scale is the same). Random effects are excluded from these plots.

4.1.4. Effect of target position on target distance

Target picture position was included in the model as a control variable. If participants had search strategies, such as left-to-right or top-to-bottom scanning, then the eyes would be likely to fall on the target more quickly when the target occurred in certain positions on the screen. Including these effects would strengthen the ability of the model to capture our predictors of interest by accounting for this variation. The model summary shows that target position had a significant effect on the distance of fixations from the target over time (top-left: $F(3.979, 476634.3) = 321.5$; top-right: $F(3.941, 476634.3) = 254.7$; bottom-left: $F(1.002, 476634.3) = 895.8$; bottom-right: $F(3.990, 476634.3) = 360.9$). The left panel of Figure 4 shows the effect of target position over time. Time is on the x-axis, target distance on the y-axis. Each position on the screen is represented by a coloured line according to the key in the top right corner of the plot. The plot shows substantially different distances, depending on the target position. Fixations are closest to the target when the target is in the top left corner, and furthest when it is in the bottom right corner. The effect emerges immediately in the first fixation, around 150-200 ms, and continues until late in the trial, around 800 ms. The eyes locate the target more quickly when it is in the top left of the screen; otherwise the eyes may initially move further away from the target compared to the initial position on the fixation cross. Note that this is true on average, but does not entail that this occurs on every trial. Indeed, given the size of the effect, it is unlikely that it occurs on every trial.

4.2. Competitor distance model: distance of fixations from the competitor picture

Apart from investigating the effects of uncertainty on how accurately participants fixated the *target*, we were also interested in how perceptual uncertainty affects the degree to which participants were drawn towards the *competitor* picture. We therefore ran models looking at the distance of fixations from the competitor picture. This measure corresponds to Clayards et al. (2008), in which the by-trial proportion of fixations on the competitor object was reported. The models included the same predictors as the target distance models, only the dependent variable was the distance of fixations from the competitor picture, and competitor position on the screen replaced target position. A visualisation of the raw data for competitor distance in Experiment 1 is shown in Appendix E (lower panel).

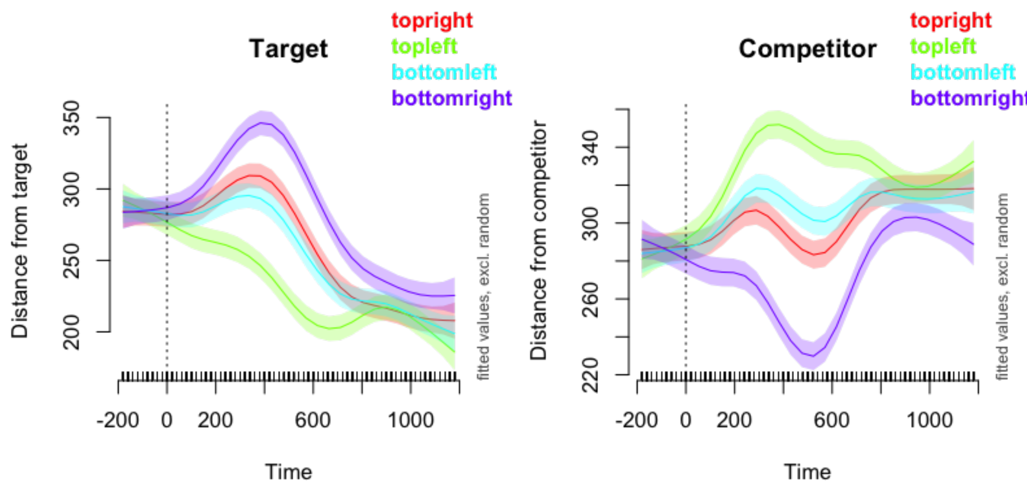


Figure 4: Model fit for the effect of target position in the best fit model for the Euclidean distance from the target (left panel) and the effect of competitor position on the Euclidean distance from the competitor (right panel) in Experiment 1. Time (ms) is on the x-axis. Distance from the target (left panel) or competitor (right panel) is on the y-axis. Each position on the screen is represented by a line, colour-coded according to the legend in the top right corner. The predictor Condition is set to low-variance; VOT is set to -0.5. As the models did not include an interaction between target/competitor position and VOT or target/competitor position and condition, the estimated effects of position are the same for low and high variance and for the different VOT values. Error bars are 95% confidence intervals (indicating the uncertainty around the model estimates).

4.2.1. Effects of voice onset time value on competitor distance

The model summary for competitor distance (Appendix B) shows the interaction of VOT by condition over time. The baseline (low-variance) condition is shown in the lower left panel of Figure 3. For all VOT values, the distance from the competitor first shows a dip (blue area), then steadily increases over time. Comparison of the estimated distance from the target and competitor pictures in this time period suggests that the eyes initially move toward the competitor, before rejecting it and moving towards the target.

The effect of VOT starts to emerge in the first fixations of the trial, around 150 ms to 300 ms after stimulus presentation. The distance from the competitor decreases for the outer and mean VOT values earlier than for the central VOT values, as the eyes move towards the competitor object. After this initial period, the distance from the competitor is smallest at the central values. This pattern suggests that when the VOT is near the category boundary, it takes participants longer to move their eyes away from the fixation cross for the first fixation of the trial. At all VOT values, the initial fixations tend to move towards the competitor object, before rejecting it and moving towards the target. At the central values, this is process seems to be delayed, with eye movements both towards and away from the competitor occurring later at the central values than at the mean and outer values. That is, the short distance from the competitor (blue area) starts later and continues until later in the trial at the central VOT values. The difference in competitor distance between central and outer VOT values remains throughout the trial. At the outer VOT values, the distance from the competitor steadily increases, starting from around 550 ms (green then yellow areas). Near the category boundary, although the distance increases, it does not reach the same level as the outer VOT values. This suggests that a greater degree of uncertainty remains for the central VOTs right until the end of the trial.

4.2.2. Effects of distribution condition on competitor distance

As noted above, there was a significant effect of VOT by condition over time. Including a VOT-by-condition interaction significantly improved model fit, compared to a model without condition (χ^2(5.0)=8.663, p $<$ .004). This effect is shown in the models plots (lower panels of Figure 3), which show the distance of fixations from the centre of the competitor object in the low-variance (left panel) versus the high-variance condition (right panel). The effect of distribution condition seems to emerge mainly at the central VOTs at the beginning and end of the trial, where fixations are closer to the competitor in the high-variance condition than in the low-variance condition. In the early fixations, the effect of VOT is flatter in the high-variance compared to the low-variance condition. In the low-variance condition, the eyes take longer to move away from the fixation cross at the central values compared to the more peripheral values. However, this effect is absent in the high-variance condition, in which the eyes move towards the competitor object at around the same time for all VOT values. From around 500 ms onwards, fixations are closer to the competitor object around the central VOT values in the high-variance compared to the low-variance condition.

4.2.3. Effects of the position of the competitor on the screen on competitor distance

The model summary shows that competitor position had a significant effect on the distance of fixations from the competitor over time (top-left: $F(3.969, 476712.6) = 118.975$; top-right:$F(3.736, 476712.6) = 87.236$; bottom-left $F(3.799, 476712.6) = 84.505$; bottom-right $F(3.939, 476712.6) = 120.162$). The results are shown in the right panel of Figure 4. The general pattern is the inverse of the effects of target position in the target distance models. The fixations are closest to the competitor picture when it is in the top left corner, and furthest when it is in the bottom right corner.

4.3. Discussion

Experiment 1 investigated the effects of perceptual uncertainty on eye movements towards target and competitor pictures during perception of Cantonese words beginning with aspirated and unaspirated consonants. Two causes of uncertainty were investigated. On the one hand, this experiment investigated the time course of effects of changes in the acoustic cue value, VOT, during speech perception. This manipulation was the same for all participants. Greater perceptual uncertainty was predicted as cues approached the category boundary. On the other hand, the experiment investigated the effects of within-category acoustic variance. That is, the presentation frequency of the different acoustic cue values. Based on the results of Clayards et al. (2008), we predicted that fixations would fall closer to the target and further from the competitor for participants in the low-variance condition, compared to the high-variance condition.

4.3.1. *Effects of time*

Overall, the GAMM models for Experiment 1 showed that fixations became closer to the target and further from the competitor over time. However, this was a nonlinear trend. In the target distance model, there was an initial period of relative stability, followed by a steady convergence on the target. In the competitor distance model, there was a *decrease* in distance from the competitor in the early period around 200-400 ms, as fixations initially approached the competitor for a period before moving away from it. After this period, fixations began to steadily approach the target.

4.3.2. *Effects of voice onset time value*

Both the target distance and the competitor distance models showed a nonlinear effect of VOT value on participants' perceptual uncertainty. In the target distance model, at the outer VOT values, fixations began to rapidly approach the target picture by around 500 ms; by around 700-800 ms, fixations were within the target picture interest area, on average. However, at the more central VOT values, a substantial amount of uncertainty remained throughout the trial. The distance from the target remained substantially greater near the category boundary than at the outer VOTs right until the end of the trial. Conversely, in the competitor distance models, the distance from the competitor was generally smaller at the central VOT values, compared to the outer values. This effect of VOT on distance to the competitor emerged very early, in the first fixations of the trial. Near the category boundary, it took longer for the eyes to move away from the fixation cross. After this delay, fixations were closer to the competitor at the category boundary for the rest of the trial.

Interestingly, the effect of VOT value seemed to emerge mainly between the central values and the distribution peaks. The exaggerated acoustic information in the outer cue values did not seem to greatly benefit participants in terms of the time it took to fixate the target. Another interesting observation is that these effects are quite symmetrical. This is surprising given that within-category acoustic variance is *asymmetrical* in language. In Cantonese bilabial stop production (as in English), the variance in unaspirated stops is much lower than in aspirated stops. The standard deviation of unaspirated stops in syllable production is less than 6 ms, compared to more than 21 ms in aspirated stops (Ng and Wong, 2009). Given that there is more than three times as much variation in aspirated stimuli in speech, we might expect that listeners are more tolerant of variation in aspirated stimuli in the experiment setting. For example, we might expect to see steeper slopes on the unaspirated side in the plots. But this was not the case.

4.3.3. Effects of acoustic cue variance

The target distance models did not show any significant effects of distribution condition. The competitor distance models, on the other hand, did show a significant interaction with VOT over time. The model plots indicate that the biggest differences between conditions occur at the central VOTs, near the category boundary. In the low-variance condition, the eyes seem to take longer to move away from the fixation cross at the central values at the beginning of the trial. Later in the trial, after about 600-700 ms, fixations are closer to the competitor in the high-variance condition, compared to the low-variance condition. This result is line with our hypothesis that the greater degree of within-category acoustic variance would lead to greater uncertainty in the high-variance condition. The result is also consistent with the findings of Clayards et al. (2008), which showed that the overall proportion of fixations on the competitor versus the target was greater in their high-variance condition. One of the aims of this study was to extend the investigation to examine the time course of effects. The competitor distance model shows that the effect of distribution emerges early, affecting the very first fixations, and continues over the course of the trial.

This early effect could be attributed to changes in early perceptual processing of the acoustic information as a result of the distributional input. However, given that there was no effect of trial in this experiment, it is unlikely that the effect stems from 'perceptual learning' such that there were shifts in the category boundary. Another possibility is that participants adopt a global strategy in response to the level of uncertainty. As uncertainty increases, participants look around more in search of additional evidence to support their selection. Participants tend to fixate the competitor before moving to the target. They do this more and later in the trial in the high-variance condition. This suggests that these fixations are part of a kind of verification process. As competition between target and competitor increases, it takes longer to reject the competitor in favour of the target.

4.3.4. Effects of target and competitor position

An interesting observation that comes out of this study is the effect of the location of the target and competitor on the screen. Fixations were substantially closer to the target when the target was in the top left corner of the screen, and further when it was located in the bottom right; conversely, fixations were further from the competitor when the competitor picture was located in the top left corner of the screen, and closer when it was located in the bottom right. These effects are probably the result of scanning strategies during the preview period and the early part of the trial. If participants had a particular scan path that favoured the top-left over the bottom-right, this would enable them to locate the target and reject the competitor better when it was in the top-left position and least when it was in the bottom right.

Though we know of no other study that has reported this effect in the visual world paradigm, a bias for initial fixations to move to the left is known in scene perception research (Dickinson and Intraub, 2009; Ossandon, Onat, and Koenig, 2014). This left-to-right, top-to-bottom pattern closely matches the direction of eye movements during reading. However, the extent to which reading direction contributes to the effect is unclear. Cross-linguistic studies of scene and face perception have reported mixed results (Chokron and De Agostini, 2000; Gilbert and Bakan, 1973; Heath, Rouhana, and Abi Ghanem, 2005; Nicholls and Roberts, 2002; Vaid and Singh, 1989) suggesting that there may be a language-independent effect that is modulated by the direction of reading.

Regarding the time course of effects, both the target and competitor position effects were present for most of the trial, beginning with the first fixation. However, the time course is slightly different for target position and competitor position. For target position, when the target is in the top left, the distance steadily decreases from the first fixation onwards. When the target is in the bottom right, in contrast, the first fixations tend to move sharply away from the target in the first fixations, perhaps landing on the competitor, or a distractor picture. The distance continues to increase until around 400 ms. At this time, the participant presumably realises that they have made an error and prepares to launch another saccade. But this error sets the participant back substantially, and although the distance decreases steadily from this point, the lines only come together again around 800 ms, towards the end of the trial.

For competitor position, the overall effect is roughly the inverse of the effect of target position: fixations are furthest from the competitor when it is the top left, and come closest when it is in the bottom right. However, there are also differences in the time course, compared to the effect of target position. While the lines of the four positions in the target position plot are roughly parallel for a large part of the trial, in the competitor position plot, the effect is closer to a mirror image. The first fixations move towards the competitor when it is in the bottom right and away from it when it is in the top left and this pattern continues well into the trial. The probable reason for this difference in the time course between target and competitor is that when fixations land on the target picture, they are much more likely to stay there for the rest of the trial. On the other hand, if early fixations land on the competitor picture, they are likely to move away again after a time. The plot shows that the eyes start moving away from the competitor at around 400 to 550 ms, depending on its location.

5. Experiment 2 Tones

5.1. Method

Participants. Thirty-nine native Cantonese-speaking undergraduate students from the Chinese University of Hong Kong participated in the experiment. An additional six participants were recruited, but were excluded from analysis due to the eyetracker unexpectedly quitting before the end of the experiment (four participants) and inability to calibrate (two participants).

Experiment design and stimuli. The experiment design was the same as Experiment 1, except that different stimulus items were used. Visual stimuli were picture pairs whose names were word pairs that were either high level tone (e.g. jin1 'carpet'; gun1 'crown') or mid level tone (jin3 'arrow'; gun3 'can'). The two members of each word pair had the same segmental syllable. Initial consonants were either velar stops ('g') or alveolar affricates ('j'). Auditory stimuli were produced by the same speaker as Experiment 1. The stimuli were then resynthesised in PRAAT (Boersma and Weenink, 2012), using the mid tone as the target, to create a 12-step f0 continuum with equal semitone steps ranging from 86 Hz to 129 Hz. Syllable duration ranged from 357 ms to 491 ms, of which the mean initial consonant duration was 41 ms for the stops and 61 ms for the affricates.

Procedure. The procedure was identical to Experiment 1.

6. Analysis

Analysis was conducted using the same variables as Experiment 1, except that the acoustic cue was a continuum of pitch (f0) values, instead of VOT values.

7. Results

7.1. Target distance model: distance of fixations from the target picture

7.1.1. Random effects

As in Experiment 1, the models for Experiment 2 included by-participant by-item random wiggly curves over time (Appendix C). Random effects were modelled as separate smooths for each participant-item pair.

7.1.2. Effects of pitch value on target distance

Model comparisons showed that model fit was improved by including a nonlinear interaction of pitch by condition over time. The model summary for target distance is shown in Appendix C. A visualisation of the raw data is provided in Appendix F (upper panel). The effect of pitch value over time is illustrated in the model plots for the baseline (low-variance) condition (left panel of Figure 5). The distance of fixations from the target picture is plotted on the z-axis, represented by colour codes. Higher values (shown in yellow) indicate a relatively greater distance from the target; lower values (shown in blue) indicate a relatively shorter distance. Category means are at -2.5 (for the mid-tone stimuli, e.g. gon3) and 2.5 (for the high-tone stimuli, e.g. gon1).

The plot shows a very similar pattern to the results for the VOT model. Changes in eye movements over the course of the trial occur differently for different pitch values. Until around 200 ms, the plot shows a flat distribution, as participants are looking at the fixation cross. Then the eyes begin to move away from the fixation cross. After about 400 ms, target distance starts to decrease steadily.

As in the VOT model, differences between pitch values begin to emerge around 400-500 ms after presentation of the auditory stimulus. In addition, the target distance remains greater at the central values, compared to the outer values, for the rest of the trial.

However, there are also differences compared to the VOT model. The plot for the pitch model is not entirely symmetrical. The greatest distances from the target are actually centred just above 0, at about 0.5, rather than at 0, as expected. This suggests that the category boundary in the stimuli may have been slightly lower than participants' own category boundary estimates.

7.1.3. Effects of distribution condition on target distance

Unlike the VOT models of target distance, in which there was no effect of condition, the interaction of pitch by condition over time significantly con-

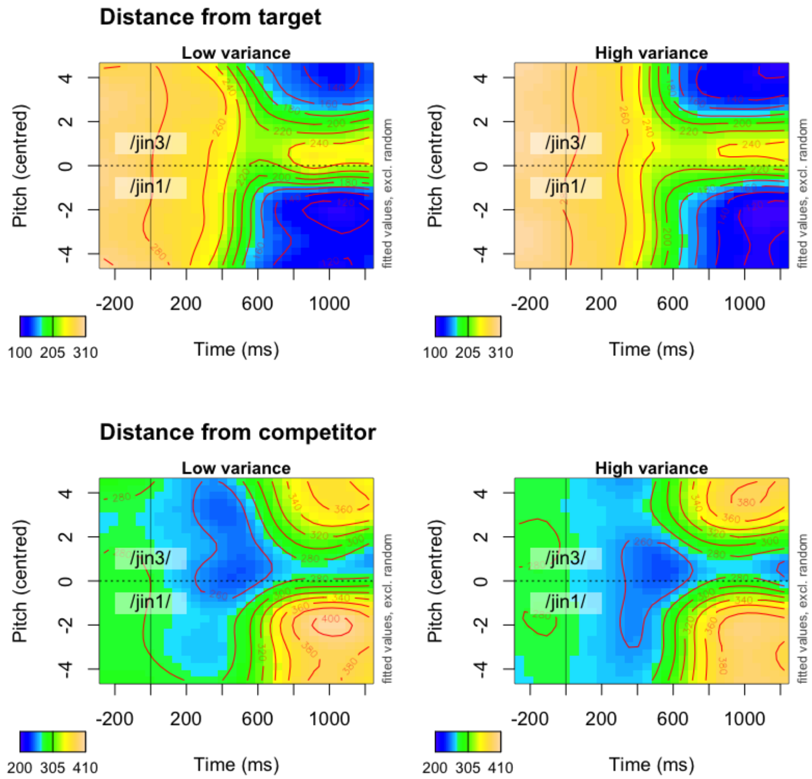


Figure 5: Topographical maps for the pitch models in Experiment 2. Top row: model fit for the best fit model of Euclidean distance from the target picture for the low-variance (left panel) and high-variance conditions (right panel). The predictor Target Position is 'top left' in this plot (see the left panel of Figure 6 for the effects of Target Position). Bottom row: model fit for the best fit model of Euclidean distance from the competitor picture for the low-variance (left panel) and high-variance conditions (right panel). The predictor Competitor Position is 'top left' in these plots (see the right panel of Figure 6 for the effects of Competitor Position). All plots: Estimated effects are in pixels. Time (ms) is represented on the x-axis. Pitch is on the y-axis. Pitch is centred around 0, the category boundary. The negative pitch values correspond to mid-tone stimuli (e.g. jin3), the positive values to high-tone stimuli (e.g. jin1). Category means are at centred pitch values -2.5 and 2.5, respectively. Distance is plotted on the z-axis, represented by colour codes. Higher values (yellow areas) indicate a relatively greater distance; lower values (blue areas) indicate a relatively smaller distance. The key in the bottom left corner shows corresponding pixel values and the z-limits. Note that the range differs between the surface plots for target and competitor model plots: 100-310 for the target plots; 200-410 for the competitor plots. (The scale is the same). Random effects are excluded from these plots.

tributed to model fit in the pitch models for target distance ($\chi^2(5.0)$=41.812, p < .001). In the upper panel of Figure 5, differences in the distance from the target appear between the low-variance condition (upper left panel) and the high-variance condition (upper right panel). The differences are most apparent at the central pitch values, beginning at around 700 ms. There is greater distance from the target in the low-variance compared to the high-variance condition. This result was counter to our expectations. Based on the results of Clayards et al. (2008), we hypothesised greater distance in the high-variance condition. A possible reason for this effect may be that the stimulus category boundaries differed from participants' initial category boundary estimates, as noted above. In the high-variance condition, because participants had more experience with these central values, this may have given them the opportunity to adjust their category boundaries and bring them in line with the distribution. Unlike in the VOT models, there are also differences at the category means. Fixations are further from the target for the high tone (positive pitch values) and closer to the target for the mid tone (negative pitch values) in the low-variance condition, compared to the high-variance condition.

7.1.4. Effects of target position on target distance

The effects of target location in the pitch model are very similar to those seen in the VOT models. The model summary shows a significant effect of target position on target distance over time (top-left: F(3.974, 507685.1) = 261.29; top-right: F(2.847, 507685.1) = 260.67; bottom-left: F(1.156, 507685.1) = 676.26; bottom-right: F(3.979, 507685.1) = 273.96). The effects are shown in the left panel of Figure 6. Fixations are closest to the target when the target occurs in the top left corner of the screen, and furthest when the target is located in the bottom right of the screen.

7.2. Competitor distance model: distance of fixations from the competitor picture

As with the VOT models, we were interested not only in the target fixations, but also in how much fixations were drawn to the competitor during tone perception. The model summary for competitor distance is shown in Appendix D. A visualisation of the raw data for competitor distance in Experiment 2 is shown in Appendix F (lower panel).

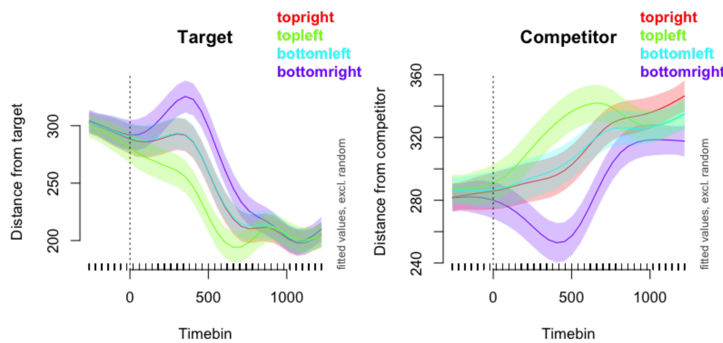


Figure 6: Model fit for the the effect of target position in the best fit model for Euclidean distance from the target (left panel) and the effect of competitor position on the distance from the competitor (right panel) in Experiment 2. Time is on the x-axis. Distance from the target (left panel) or competitor (right panel) is on the y-axis. Each position on the screen is represented by a line, colour-coded according to the legend in the top right corner. The predictor Condition is set to low-variance; pitch is set to -0.5. As the models did not include an interaction between target/competitor position and pitch or target/competitor position and condition, the estimated effects of position are the same for low and high variance and for the different pitch values. Error bars are 95% confidence intervals (indicating the uncertainty around the model estimates).

7.2.1. Effect of pitch value on competitor distance

The model for competitor distance included a nonlinear interaction of pitch by condition over time. The effects of pitch over time are shown in the baseline (low-variance) condition (lower left panel of Figure 5). In the early fixations, seems to be asymmetrical. As expected, fixations are closer to the competitor at the central values. But they are also closer to the competitor at the very high pitch values. This effect of the peripheral pitch values is smaller in the mid tones, so that there is an overall bias towards the mid tone. This effect appears around 200-400 ms. From around 600 ms, there is a steady increase in the competitor distance at the outer pitch values; however, the competitor distance remains shorter the closer the pitch is to pitch values just above the category boundary, at centred pitch values 0.5-1. We see the same asymmetry that appeared in the target distance models.

7.2.2. Effects of distribution condition on competitor distance

In the model for competitor distance, the interaction between condition and pitch over time significantly contributed to model fit, compared to a model without condition ($\chi^2(5.0)$=69.970, p < .001). The effect of distribution condition is shown in the model plots (lower panels of Figure 5). As noted above, in the low-variance condition, the effect of pitch cue value emerges from around 200-400 ms. Fixations move towards the competitor early in the first fixations near the category boundary. These fixations occur earlier in the low-variance condition (left panel), compared to the high-variance condition (right panel). Additionally, at the central values, the competitor distance is smaller in the low-variance condition, compared to the high-variance condition in this period. The competitor distance remains shorter in the low-variance condition right up until near the end of the trial.

7.2.3. Effect of competitor location on competitor distance

The model summary for competitor distance shows a significant effect of competitor position over time (top-left: F(3.967, 507729.8) = 127.84; top-right: F(3.700, 507729.8) = 105.73; bottom-left: F(3.799, 507729.8) = 130.65; bottom-right: F(3.808, 507729.8) = 111.86). This result follows a very similar pattern to the VOT models of competitor distance, and roughly the inverse of the effect of target position on target distance. As shown in the right panel of Figure 6, the competitor distance is greatest when the competitor is in the top left corner, and smallest when it is in the bottom right corner.

7.3. Discussion

Like Experiment 1, Experiment 2 investigated the effects of perceptual uncertainty on eye movements towards target and competitor pictures during Cantonese speech perception. While Experiment 1 investigated a temporal cue, voice onset time, in a segmental contrast, aspiration, Experiment 2 investigated a suprasegmental cue, pitch (f0), in a lexical tone contrast. The same two types of uncertainty effects were investigated: differences in the acoustic cue value, in this case pitch, and differences in the amount of acoustic cue variance (low-variance versus high-variance). As in Experiment 1, greater perceptual uncertainty was expected as cues approached the category boundary, compared to more peripheral pitch values, and in the high-variance compared to the low-variance condition. Perceptual certainty was investigated in two separate models. The first examined the distance from the centre of the target picture; and the second, the distance from the centre of the competitor picture.

7.3.1. *Effects of time*

The overall trend of fixations over time in the GAMM models for Experiment 2 was remarkably similar to Experiment 1. Generally, fixations became closer to the target and further from the competitor over time, but this followed an initial small *decrease* in distance from the competitor in the early period. The eyes initially moved towards the competitor in the first fixations of the trial, before steadily moving away from it.

7.3.2. *Effects of pitch value*

The effect of time was modulated by pitch value. At the outer pitch values, fixations began to rapidly converge on the target picture by around 500-600 ms, and by around 700-800 ms, fixations were within the target picture interest area, on average. However, as the pitch values approached values just above the category boundary, the distance from the target gradually increased. At the values 0.5-1, fixations were substantially further from the target compared to the outer values. This pattern of increased target distance suggests that participants' category boundaries were centred around the values 0.5-1, rather than 0.

While the bulk of the pitch value effect occurs as values approach these values just above the category boundary, there is also an interesting effect towards the periphery of the mid tone, which appears in the lower half of the plot, in the later part of the trial. There is a peak where fixations are closest to the target that emerges between 800-1200 ms and which occurs at the distribution peak for the mid tone (pitch -2.5). Fixations are closest to the target at the distribution peak, and become few towards the edge of the distribution. This differs from the positive pitch values, as well as the VOT models. The fact that this effect appears in the tone models, but not in the VOT models may reflect language-specific properties of the phonological system. The consonant system in Cantonese has only two levels of aspiration: aspirated and unaspirated. However, in the tonal system there are three level tones. This experiment investigated only the high and mid level tones, but there is also a low level tone. Although it does not occur in this experiment, this low tone seems to be having an affect. As the outer regions of the mid tone begin to slip into low tone territory, the distance from the target increases slightly, suggesting that activation of this low tone may be creating an additional cause of uncertainty.

The presence of the low tone at the lower boundary of the mid tone seems to have had an additional effect. Towards the end of the trial, an asymmetry emerges in the target distance. The pattern of fixations suggests that the participants' category boundaries are approximately half a continuum step higher than the experimental boundary. This may be due to the pressure of the low tone. This is supported by evidence from production data showing that there is less variation in the pitch height of the mid tone (Siddins and Harrington, 2015), presumably due to pressure from the surrounding tones. The effect does not occur in the high tone, which has no tone above it.

7.3.3. Effects of acoustic cue variance

In Experiment 2, there was a significant interaction between distribution condition and pitch over time in both the target distance and the competitor distance models. In the target distance model, the effect of distribution condition was greatest near the category boundary, and emerged around 700 ms. There was also a similar effect at the category boundary in the early fixations, around 200-400 ms. Contrary to expectations, at the central values, the distance from the target was greater in the low-variance condition than the high-variance condition. A similar effect was found in the competitor distance models, where distance was shorter in the low-variance condition. Based on the results of Clayards et al. (2008), we predicted greater competitor distance in the low-variance condition.

This result is probably due to a mismatch between the experimental distribution and participants' initial category boundaries, as noted above. The VOT models suggest that low-variance input leads to clearer, more certain perception. However, in the pitch experiment, the experimental category boundaries appear to be slightly lower than participants' initial estimated boundaries. This leads to quite different effects of the distribution. When participants encounter an input distribution that does not match their expectations, this leads to greater uncertainty in the low-variance condition.

Effects of cue variance also emerged at the category means. In both groups, there seemed to be a bias toward the mid tone (negative centred pitch values): fixations were more likely to be closer to the target and further from the competitor for the positive pitch stimuli than the negative pitch stimuli. This effect was stronger in the low-variance condition. The pattern lends further support to the idea that the low-variance condition leads to less flexible representations.

8. General Discussion

The present study investigated the temporal dynamics of perceptual uncertainty during Cantonese speech perception. Participants saw pictures of word pairs consisting of aspirated and unaspirated counterparts (Experiment 1) or mid and high tone counterparts (Experiment 2) and heard an auditory stimulus sampled from acoustic cue continua corresponding to the word pairs. Two experimental manipulations were expected to affect participants' level of perceptual uncertainty. The first manipulation was the acoustic cue value; i.e. the location of the cue along the acoustic continuum between speech sounds. The second manipulation was the degree of within-category acoustic variance. Participants heard either a relatively large amount of variation (the *high-variance* distribution condition) or relatively little variation in acoustic stimuli (the *low-variance* distribution condition). Eye movements to the pictures were monitored until participants selected a picture by clicking on it. For each experiment, two sets of models were run. The first examined the distance of fixations from the target picture, and the second examined the distance from the competitor picture.

We expected to see gradient effects in the distance of fixations from the target and competitor pictures, depending on the location of the cue along the continuum, with fixations landing further from the target as the cue approached the category boundary (McMurray et al., 2009). We also expected that fixations would be further from the target in the high-variance, compared to the low-variance condition (Clayards et al., 2008). One of the most interesting aspects of the study was the investigation of the time course of effects. Given that the time course of statistical distribution effects has not previously been investigated, the temporal aspects of the present study were largely exploratory. The time course of other effects during speech perception have been investigated using a similar experimental methods. For example, Mitterer and Reinisch (2013) investigated the time course of effects in lexically-guided adaptation. They found effects in the first fixations of the trial. Eye movements were affected by the fricative type on a particular trial (s-final versus f-final) as well as training condition (f-biased versus s-biased). Both effects emerged in roughly the same time window. They interpreted this effect as evidence that lexically-guided adaptation affects the very early perceptual processes rather than higher-order decision processes.

8.1. *Effect of time*

Analysis of eye movement data using Generalised Additive Mixed Modelling (GAMM) revealed that the distance of fixations both from the target picture and from the competitor picture in Experiment 1 followed a nonlinear trajectory over time. Overall, the eyes tended to move towards the target and away from the competitor over time. However, this pattern was not constant over the whole trial. Up until around 200 ms after presentation of the auditory stimulus, the model plots show that target distance remained steadily around 280 pixels, as the eyes focused on the fixation cross. At around 200 ms, the eyes began to move away from the fixation cross. In the early part of the trial, between 200 ms and 400 ms, there was an initial small *decrease* in distance from the competitor, indicating that fixations initially moved towards the competitor in this period, before steadily moving away from it. This suggests that if participants fixate the competitor picture, the most likely point in time that they will do so is in the first fixations of the trial. Finally, from around 400 ms onwards, the distance of fixations from the target steadily decreased and distance from the competitor increased until the end of the trial. The time course of effects in Experiment 2 was essentially the same as Experiment 1. Fixations initially remained on the fixation cross, then shifted briefly towards the competitor before moving steadily towards the target picture for the remainder of the trial.

8.2. *Effects of acoustic cue value*

Models for both target distance and competitor distance showed that the acoustic cue value had a nonlinear effect on participants' perceptual certainty. The distance of fixations from the target and competitor over the course of the trial varied as a function of VOT value (Experiment 1) or pitch (Experiment 2). As predicted, in the VOT experiment, the target distance increased as VOT values approached the category boundary. This is consistent with the results of earlier studies that have found gradient effects of VOT value in discrimination of stop contrasts (e.g. McMurray et al., 2008a, 2002). Conversely, in the competitor distance models, the distance from the competitor was smaller at the central VOT values, compared to the outer values, providing further support for the conclusion that uncertainty increased as cue values approached the category boundary. The same nonlinear effect of cue value was also found in Experiment 2, with target distance increasing and competitor distance decreasing as the pitch value approached what seemed to be participants' initial category boundary, just above the boundary set in the experiment. This shows that the same kind of gradient sensitivity that has been shown in VOT perception also applies to perception of pitch height during tone perception. Although gradient sensitivity to pitch height in Cantonese has been investigated in offline identification and discrimination tasks (e.g. Francis, Ciocca, and Ng, 2003), as far as we are aware this is the first investigation of native Cantonese tone perception using eye movement data, which provides a measure of participants' uncertainty over and above their final category judgment. The results additionally demonstrate that this is a nonlinear effect.

As for the time course of the cue value effects on target distance, changes in eye movements over time occurred differently at different points on the VOT/pitch continuum. Differences between VOT values in Experiment 1 began to emerge around 400-500 ms after stimulus presentation. This was consistent with a previous study that examined proportions of fixations on the target picture object during English voiced-voiceless stop discrimination (McMurray et al., 2009). At the outer regions of the VOT continuum, after a period of relative stability, fixations began to rapidly approach the target picture from around 500 ms. The eyes generally reached the target picture interest area by about 700-800 ms, on average. However, at the central VOT values, a substantial amount of uncertainty remained throughout the trial. The distance from the target remained considerably greater near the category boundary than at the outer VOTs right until the end of the trial.

There were some intriguing differences in the time course between the target distance and competitor distance models. Specifically, the competitor distance effects emerged earlier in the trial, compared to the target distance effects. In the competitor distance models, the effect of VOT starts to emerge around 150 ms to 300 ms after stimulus presentation, compared to around 500 ms in the target distance models. The competitor distance decreases for the outer VOT values earlier than for the central VOT values. This suggests that when the VOT is near the category boundary, it takes participants longer to move their eyes away from the fixation cross for the first fixation of the trial. The early effects in the competitor models are probably due to participants fixating the competitor mostly in the first fixation or two, after which time they reject it in favour of the target. It is interesting that even in these very early 'error' fixations, the acoustic cue value affects the speed with which the eyes move towards the competitor.

The overall pattern of effects in Experiment 2 was very similar to Experiment 1. However, the pattern was shifted upwards. While the largest effect of VOT in Experiment 1 occurs near the category boundary, centred pitch 0, the largest effect of pitch value in Experiment 2 centres around 0.5-1, rather than 0. This suggests that participants' category boundaries were higher than those specified in the stimulus distributions. In addition, in Experiment 2, the effect of pitch value on target distance emerged earlier than the VOT effect in Experiment 1, in the first fixations of the trial. There is also another interesting difference between the VOT and pitch cue effects. There appears to be little effect of cue value at the edges of the VOT cue continuum or in the positive pitch values (i.e. the high tone). However, in the lower half of the plot for pitch (Figure 5), distance from the target starts to increase again at the edge of the continuum. This is probably due to an influence of the low level tone. While the present experiment investigated only the high and mid level tones, Cantonese also has a third level tone, the low tone. The pitch height of the low and mid tones is closer together than the pitch of the mid and high tones. It is likely that at the lower edge of our continuum, participants began to have activation from this low tone, adding another source of uncertainty to the eye movements. Indeed, acoustic studies of production data show that the variance in the mid tone is much less than either the high or low tones (Siddins and Harrington, 2015), probably as a result of pressure from the surrounding low and high tones. This also seems to have had a knock-on effect on the perception of the category boundary in the present experiment. There is an asymmetry in the fixation distance in the later part of the trial. Participant category boundaries seem to be shifted up by half a step relative to the stimuli category boundary. Since there is no tone higher than the high tone, this crowding effect is absent at the top edge of the continuum. And since there are only two levels of aspiration (aspirated and unaspirated) in Cantonese consonants, the effect is absent in the VOT models also.

8.3. Effects of distribution condition

A very interesting pattern of effects emerged for distribution condition. Based on the results of Clayards et al. (2008), we hypothesised that the fixations would fall further from the target and closer to the competitor in the high-variance, compared to the low-variance condition. In Experiment 1, the effect of distribution was not significant in the target distance models. However, the competitor distance models showed a significant nonlinear interaction between condition and VOT over time. The finding of an effect of cue variance replicated the findings of Clayards et al. (2008), but with a continuous measure of competitor distance rather than fixation proportions. In a visual world eyetracking experiment, Clayards et al. (2008) presented native English listeners with a 12-step VOT continuum and pictures of English /b/ and /p/ words, presented in either a high- or a low-variance condition. Their results showed that categorisation accuracy and the proportion of fixations on the competitor depended on the degree of variance. The same overall pattern of results that Clayards et al. (2008) found in English voiced and voiceless stops was found in the present study in Cantonese words beginning with aspirated and unaspirated stops and aspirated and unaspirated affricates (Experiment 1). This finding lends further support to the idea that listeners are sensitive to the amount of acoustic variance in the signal and that increased variance leads to increased perceptual uncertainty.

Clayards et al. (2008) hypothesised that the largest differences in looks to the competitor object between the low-variance and high-variance conditions would be at the VOT values closest to the category boundaries. However, due to a smaller number of participants in their experiment and a different method of analysis, the relatively small number of trials at the most central VOT values meant that there was insufficient power to test this prediction for all VOTs. One of the aims of present experiment was to test this hypothesis by including these central acoustic values in the analysis. With the increased power of GAMMs, along with a larger number of participants, we were able to evaluate the fixations at these VOT values. Clayards and colleagues' predictions were upheld. The greatest differences emerged at the central VOT values.

Another aim of the present study was to uncover the time course of perceptual uncertainty effects by analysing changes in eye movement behaviour over the course of the trial. While Clayards et al. (2008) reported between-condition differences in the proportion of fixations collapsed over the trial, we were interested in when these differences emerged and how they changed over the course of the trial. Using a continuous measure of distance and using GAMMs for analysis enabled us to also investigate the temporal effects. Effects of distribution condition emerged very early, in the first fixations of the trial and increased later in the trial, with maximal effects after around 500 milliseconds.

It is interesting to note the different time course of effects that emerged in the present study by examining eye movements to both the target and competitor pictures separately. In previous eye movement studies that have used a VOT continuum to investigate acoustic cue processing, where analysis has focused on fixations to the target (e.g. McMurray et al., 2009), VOT effects emerged around 600 ms. In studies that have included both target and competitor by analysing the proportion of looks to each category, e.g. /b/ vs. /p/ (e.g. McMurray et al., 2008b), the effects seem to emerge earlier. In the present study, effects of the VOT value emerged in the target distance models around 500-600 ms after stimulus presentation. In the competitor distance models, the cue value effect emerged early, with fixations further from the target at the category boundary in the first fixations of the trial, between 150-300 ms.

In Experiment 2, unlike in the VOT models, the interaction between condition and pitch over time had a significant effect on target distance. As in the VOT models, differences between conditions were most obvious at the central pitch values, emerging around 500-600 ms. However, in the pitch models, the competitor distance was greater in the low-variance condition than the high-variance condition. This result was counter to our predictions. Based on the results of Clayards et al. (2008), we had expected to see greater distance from the target in the high-variance condition.

We believe that this result may be related to the the asymmetry in the eye movements with respect to the category boundary. It seems that in the pitch experiments the mid point between the two peaks of the distribution was lower than participants' category boundary estimates. Under these conditions, the fixations were further from the target in the low-variance condition. Around the category mean and periphery of the high tone, starting from around 200 ms until late in the trial, fixations were further from the target in the low-variance condition, compared to the high-variance condition. Conversely, around the category mean and periphery of the mid tone fixations were closer to the target in the low-variance condition, compared to the high-variance condition. The effect started slightly later in the mid tone, around 400-500 ms. In the low-variance condition, fixations were closer to the target when it was a mid tone (negative pitch values) and further from the target when it was a high tone (positive pitch values). If participants' initial category boundaries were higher than the boundaries set in the experiment, they would hear more tokens as mid tone. This effect seems to have been stronger in the low-variance condition. This pattern suggests that a low-variance distribution may lead to more robust categories, but that this in turn leads to a trade-off when tokens deviate from the expected values. Deviations from these expectations are more surprising, and therefore lead to a greater level of uncertainty and difficulty discriminating these tokens.

In addition, differences between these two experiments may also be partially attributed to acoustic differences between stimuli. In general, tones seem to be more susceptible to perceptual error and represented less precisely, compared to consonant contrasts, such as the VOT cue (Cutler and Chen, 1997; Taft and Chen, 1992) and, at least in Mandarin, are more mutable than either consonants or vowels (Wiener and Turnbull, 2015). In fact, the overall level of perceptual uncertainty seems to have been higher in the tone experiments, compared to the VOT experiments, as indicated by the range of cue values over which target distance was relatively high. In the VOT experiments, the biggest effects of VOT occur in the central three to four steps of the continuum, with largely reduced effects in the outer values. In the pitch experiments, the effects spread over up to five steps of the continuum. This suggests that participants had less precise category boundaries for tones than for the consonants. This may have given a further disadvantage to participants in the low-variance condition when it came to processing tokens towards the edges of their distribution.

One surprising finding of this study was that we did not see learning effects over the course of the experiment. That is, the effect of trial was not significant. This is interesting from the point of view of the effects of acoustic variance conditions. Since the distributional effects are expected to occur through a learning process, we expected to find changes in the pattern of eye movements over time, as participants gained experience with the distributions. This was not the case. The effect of cue variance was constant throughout the experiment. This points to a more global strategy that participants adopt in response to uncertainty. Namely, to look around more under conditions of increased uncertainty. A strategy such as this can explain the very early effects in the competitor models, as well as the lack of trial effects.

The present results show that for a given acoustic cue, the degree of variance has an immediate effect on the degree to which the cue is used for discrimination. The cues used in the present study were contrastive cues in the listeners' native language. This raises the question of how variance affects other acoustic cues present in the speech signal, such as indexical cues. In principal, the way that listeners learn to use and process these two types of cues is presumably affected by the same mechanisms. At the beginning of life, infants presumably know little about which types of cues are contrastive and which cues are indexical.[5] But experience of the way in which certain variations in speech covary with speakers, while other variations occur consistently over many speakers provides information from which infants can learn to distinguish between indexical and contrastive cues. Therefore the same mechanism that enables learners to acquire contrastive dimensions may also enable them to lower the weighting cues not relevant to the task at hand.

The relationship between these contrastive and non-contrastive cues may be vital to the process of acquiring speech categories. Rost and McMurray (2010) demonstrated a crucial role for indexical cue variation in infant language acquisition. In a series of experiments in which phonetic cues were varied or held constant, 14-month-olds were able to acquire the voicing contrast only when indexical speaker cues were varied. Statistical information in VOT values themselves within the same speaker was not sufficient for learning, but variance in *non-contrastive indexical dimensions* in the multi-speaker condition enabled infants to extract the relative invariance in the contrastive VOT dimension. This is consistent with the assumption in learning models that learning involves not only acquisition of knowledge, but also learning to ignore cues that are not effective discriminators (Baayen, Hendrix, and Ramscar, 2013).

One question is whether the effects of these experiments would generalise to new phonetic environments. For example, during or following exposure to high-variance aspiration or pitch in the present study, would participants also display high-uncertainty behaviour in response to unmanipulated stimuli? The present design did not allow for testing this kind of generalisation, as all stimuli were in the same variance condition and there were no separate training and test phases. That is, the whole experiment was both training and test. However, Idemaru and Holt (2011, 2014) have shown that when listeners were presented with a reliable cue (VOT) and a less reliable cue (f0) in one of two voicing contrasts, *beer-pier* and *deer-tear*, listeners lowered their use of the less reliable cue for discrimination between the word pair, but the effect did not generalise to the other place of articulation.

While the present results investigated individual cues in isolation, real-world speech rarely varies by a single cue. For example, Lisker (1986) identi-

[5] There is evidence that some information about the native language is learned in the womb, such as recognising the mother's voice and recognising some prosodic properties of the native language. However, even if learned before birth, this knowledge comes from experience with the ambient language.

fied as many as 16 different cues that affect native English listeners' identification responses to the voiced-voiceless contrast in stops, such as *rabid-rapid*. Jongman and colleagues (Jongman, Wayland, and Wong, 2000; McMurray and Jongman, 2011) found 20 cues involved in English fricative discrimination. So, the process of raising or lowering the weighting of particular cue values normally occurs in the context of multiple cues. These cues all compete for relevance in relation to the particular goals of the listener. Presumably any detectable cue can potentially contribute to the process of discrimination, and the size of the contribution depends in part on its variance. However, covariance with other cues has also been shown to be an important factor and may even work to counter the effects of variance and improve discrimination. For example, both voice onset time and vowel length covary with speaking rate. Toscano and McMurray (2012) found that, rather than normalising for speaking rate, listeners may instead use vowel length in combination with VOT as a cue to the voicing distinction in stops. The combination of the two cues together reduces the uncertainty that would result from variance in the single cue.

Cue weighting has been investigated in categorisation of non-linguistic auditory stimuli. Holt and Lotto (2006) presented participants with two categories distinguished by two acoustic dimensions (centre frequency and modulation frequency). In a pre-test, each dimension was tested separately to establish the appropriate step size for the continuum that would achieve an accuracy rate of 70%. However, when cues were combined, participants exhibited a bias towards use of the centre frequency cue for discrimination (Experiment 1). This bias remained even when the between-category acoustic distance for centre frequency was reduced (Experiment 2). However, when the within-category acoustic variance of modulation frequency was reduced, the relative cue weighting for modulation frequency increased (Experiment 3). Idemaru and Holt (2011, 2014) additionally showed that listeners track covariance of acoustic cues and dynamically adjust weighting of cues in response to changes in cue covariance.

Toscano and McMurray (2010) provided a demonstration of how listeners can adjust the relative weights of different cues in the signal based on their distributional statistics, using Mixture-of-Gaussians simulations. Importantly, when simulations were based on multidimensional distributions, where each cue lay on a separate dimension, the models failed to account for cue integration effects. Only when cues were integrated in a cue-weighting updating learning model, did the model reflect the interaction of effects from the two cues found in behavioural data. This suggests that the effects do not emerge purely from the statistics alone and that the learning process itself plays an important role.

The present results open up several questions for further investigation. This study involved native Cantonese listeners, who, with a lifetime of experience with the language, presumably had well-established categories for the contrasts investigated. We found that the informativity of the input can have immediate effects on processing these established categories. An interesting question is whether and how the degree of within-category variance affects acquisition of new speech categories, either in infant first language learners or in adult second language learners.

The present work focused on within-category variance. Another factor that is likely to affect speech category acquisition and processing is the acoustic interval - the acoustic interval between categories. As discussed in the introduction, it has been proposed that certain properties of speech that are particular to speech with infants help them to acquire their native phonology. Studies have shown that speech with infants tends to have increased acoustic intervals, compared to speech with adults, at least for some speech contrasts. This kind of distribution has been mimicked, at least in L2 acquisition (Escudero et al., 2011; Wanrooij et al., 2013). But infant speech also has increased variance, compared to speech with adults. Further work is needed to tease apart the effects of these two properties.

Author note The authors are grateful to Meghan Clayards and an anonymous reviewer, whose insightful comments on previous versions substantially improved this article. We also wish to thank students at the Chinese University of Hong Kong who participated in the study, and audiences at talks and conferences where this work has been presented who provided helpful feedback and suggestions. This work was supported by the European Research Council under the ERC Starting Grant [206198] to YC and an Alexander von Humboldt Award to RHB. JSN was responsible for the experiment design, implementation, stimuli creation, data collection and manuscript preparation. JSN and JvR conducted the analysis and interpretation with significant contributions from RHB. PM hosted JSN at CUHK and provided lab equipment and participant lists.

References

Allopenna, P. D., Magnuson, J. S., Tanenhaus, M. K., 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of memory and language 38 (4), 419–439.

Andruski, J. E., Blumstein, S. E., Burton, M., 1994. The effect of subphonetic differences on lexical access. Cognition 52 (3), 163–187.

Arnold, D., Wagner, P., Baayen, H., 2013. Using generalized additive models and random forests to model german prosodic prominence. Proceedings of Interspeech 2013, 272–276.

Baayen, H., Vasishth, S., Bates, D., Kliegl, R., 2015. Out of the cage of shadows. arXiv preprint arXiv:1511.03120.

Baayen, R. H., Hendrix, P., Ramscar, M., 2013. Sidestepping the Combinatorial Explosion: An Explanation of n-gram Frequency Effects Based on Naive Discriminative Learning, 329–347.

Baayen, R. H., van Rij, J., de Cat, C., Wood, S. N., to appear. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In: Speelman, D., Heylen, K., Geeraerts, D. (Eds.), Mixed Effects Regression Models in Linguistics. Springer, Berlin.

Bauer, R. S., Benedict, P. K., 1997. Modern cantonese phonology. Vol. 102. Walter de Gruyter.

Boersma, P., Weenink, D., 2012. Praat.

Cheung, W. H., Wee, L.-H., 2008. Viability of VOT as a parameter for speaker identification: Evidence from Hong Kong. Current Issues in the Unity and Diversity of Languages: Collection of the papers selected from the CIL 18 held at Korean University, in Seoul, 21–26.

Chokron, S., De Agostini, M., 2000. Reading habits influence aesthetic preference. Cognitive Brain Research 10 (1), 45–49.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., Jacobs, R. A., 2008. Perception of speech reflects optimal use of probabilistic speech cues. Cognition 108 (3), 804–809.

Cutler, A., Chen, H.-C., 1997. Lexical tone in Cantonese spoken-word processing. Perception & Psychophysics 59 (2), 165–179.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., Hogan, E. M., 2001. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes 16 (5-6), 507–534.

de Cat, C., Klepousniotou, E., Baayen, H., 2014. Electrophysiological correlates of noun-noun compound processing by non-native speakers of English. ComAComA 2014, 41–52.

de Cat, C., Klepousniotou, E., Baayen, H., 2015. Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. Frontiers in Psychology 6, 77.

Dickinson, C. A., Intraub, H., 2009. Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias? Attention, Perception, & Psychophysics 71 (6), 1251–1262.

Escudero, P., Benders, T., Wanrooij, K., 2011. Enhanced bimodal distributions facilitate the learning of second language vowels. The Journal of the Acoustical Society of America 130 (4).

Escudero, P., Williams, D., 2014. Distributional learning has immediate and long-lasting effects. Cognition 133 (2), 408–413.

Feldman, Milin, P., Cho, Moscoso del Prado Martin, F., O'Connor, forthcoming. Frontiers in Psychology.

Ferrero, F., Pelamatti, G., Vagges, K., 1982. Continuous and categorical perception of a fricative-affricate continuum. Journal of Phonetics.

Francis, A. L., Ciocca, V., Ng, B. K. C., 2003. On the (non) categorical perception of lexical tones. Perception & psychophysics 65 (7), 1029–1044.

Gahl, S., 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. Language 84 (3), 474–496.

Gilbert, C., Bakan, P., 1973. Visual asymmetry in perception of faces. Neuropsychologia 11 (3), 355–362.

Gulian, M., Escudero, P., Boersma, P., 2007. Supervision hampers distributional learning of vowel contrasts. Proceedings of the international congress of phonetic sciences.

Heath, R., Rouhana, A., Abi Ghanem, D., 2005. Asymmetric bias in perception of facial affect among Roman and Arabic script readers. Laterality: Asymmetries of Body, Brain, and Cognition 10 (1), 51–64.

Holt, L. L., Lotto, A. J., 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. The Journal of the Acoustical Society of America 119 (5), 3059–3071.

Idemaru, K., Holt, L. L., 2011. Word recognition reflects dimension-based statistical learning. Journal of Experimental Psychology: Human Perception and Performance 37 (6), 1939.

Idemaru, K., Holt, L. L., 2014. Specificity of dimension-based statistical learning in word recognition. Journal of Experimental Psychology: Human Perception and Performance 40 (3), 1009.

Jongman, A., Wayland, R., Wong, S., 2000. Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America 108 (3), 1252–1263.

Ladefoged, P., Broadbent, D. E., 1957. Information conveyed by vowels. The Journal of the Acoustical Society of America 29 (1), 98–104.

Liberman, A. M., Harris, K. S., Hoffman, H. S., Griffith, B. C., 1957. The discrimination of speech sounds within and across phoneme boundaries. Journal of experimental psychology 54 (5), 358.

Lisker, L., 1986. “Voicing” in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. Language and speech 29 (1), 3–11.

Lisker, L., Abramson, A. S., 1964. A cross-language study of voicing in initial stops: acoustical measurements. Word 20.

Liu, L., Kager, R., 2011. How do statistical learning and perceptual reorganization alter Dutch infant's perception to lexical tones? In: ICPhS. Vol. 17. pp. 1270–1273.

Marslen-Wilson, W., Warren, P., 1994. Levels of perceptual representation and process in lexical access: words, phonemes, and features. Psychological review 101 (4), 653.

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

Maye, J., Weiss, D., Aslin, R., 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. Developmental Science 11 (1).

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

McBride-Chang, C., Bialystok, E., Chong, K. K., Li, Y., 2004. Levels of phonological awareness in three cultures. Journal of Experimental Child Psychology 89 (2), 93–111.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., Subik, D., 2008a. Gradient sensitivity to within-category variation in words and syllables. Journal of Experimental Psychology: Human Perception and Performance 34 (6).

McMurray, B., Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., 2008b. Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic Bulletin & Review 15 (6), 1064–1071.

McMurray, B., Jongman, A., 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological review 118 (2), 219–246.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., 2002. Gradient effects of within-category phonetic variation on lexical access. Cognition 86 (2), B33–B42.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., 2009. Within-category VOT affects recovery from 'lexical' garden-paths: Evidence against phoneme-level inhibition. Journal of Memory and Language 60 (1), 65–91.

Mitterer, H., Reinisch, E., 2013. No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. Journal of Memory and Language 69 (4), 527–545.

Mok, P. P.-K., Wong, P. W.-Y., 2010. Production and perception of the rising tones in Hong Kong Cantonese. In: the 9th Phonetics Conference of China, Tianjin.

Newman, R. S., Clouse, S. A., Burnham, J. L., 2001. The perceptual consequences of within-talker variability in fricative production. The Journal of the Acoustical Society of America 109 (3).

Ng, M. L., Wong, J., 2009. Voice onset time characteristics of esophageal, tracheoesophageal, and laryngeal speech of cantonese. Journal of Speech, Language, and Hearing Research 52 (3), 780–789.

Nicholls, M. E., Roberts, G. R., 2002. Can free-viewing perceptual asymmetries be explained by scanning, pre-motor or attentional biases? Cortex 38 (2), 113–136.

Nixon, J. S., 2014. Sound of Mind: electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing. (Doctoral dissertation, University of Leiden. Netherlands).

Nixon, J. S., Chen, Y., Schiller, N. O., 2015a. Multi-level processing of phonetic variants in speech production and visual word processing: evidence from mandarin lexical tones. Language, Cognition and Neuroscience 30 (5), 491–505.
URL http://dx.doi.org/10.1080/23273798.2014.942326

Nixon, J. S., van Rij, J., Li, X. Q., Chen, Y., 2015b. Cross-category phonological effects on ERP amplitude demonstrate context-specific processing during reading aloud. In: Botonis, A. (Ed.), ExLing 2015: Proceedings of the International Conference of Experimental Lingustics. pp. 50–53.

Ossandon, J. P., Onat, S., Koenig, P., 2014. Spatial biases in viewing behavior. Journal of Vision 14 (2), 20–20.

Pham, Hien, Baayen, H. R., 2013. Semantic relations and compound transparency: A regression study in CARIN theory. Psihologija 46 (4), 455–478.

Rost, G. C., McMurray, B., 2010. Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. Infancy 15 (6), 608–635.

Schouten, M., van Hessen, A. J., 1992. Modeling phoneme perception. i: Categorical perception. The Journal of the Acoustical Society of America 92 (4), 1841–1855.

Siddins, J., Harrington, J., 2015. Does vowel intrinsic f0 affect lexical tone? In: ICPhS. pp. 27–43.

Taft, M., Chen, H.-C., 1992. Judging homophony in Chinese: The influence of tones. Advances in Psychology 90, 151–172.

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

Toscano, J. C., McMurray, B., 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science 34 (3), 434–464.

Toscano, J. C., McMurray, B., 2012. Cue-integration and context effects in speech: Evidence against speaking-rate normalization. Attention, Perception and Psychophysics 74 (6), 1284–1301.

Tremblay, A., Newman, A., 2014. Modelling non-linear relationships in ERP data using mixed-effects Regression with R examples. Psychophysiology TBA, 1–16.

Utman, J. A., Blumstein, S. E., Burton, M. W., 2000. Effects of subphonetic and syllable structure variation on word recognition. Perception & Psychophysics 62 (6), 1297–1311.

Vaid, J., Singh, M., 1989. Asymmetries in the perception of facial affect: Is there an influence of reading habits? Neuropsychologia 27 (10), 1277–1287.

van Rij, J., Baayen, R. H., Wieling, M., van Rijn, H., 2015. itsadug: Interpreting time series, autocorrelated data using gamms. R package version 0.6 (development version).

van Rij, J., Hollebrandse, B., Hendriks, P., in press. Children's eye gaze reveals their use of discourse context in object pronoun resolution. In: Holler, A., Goeb, C., Suckow, K. (Eds.), Empirical perspectives on anaphora resolution: Information structural evidence in the race for salience.

van Rij, J., Pya, N., van Rijn, H., Wood, S., Baayen, R., in preparation. Pupil dilation to study cognitive processing: challenges and solutions for time course analyses.

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

Wanrooij, K., Escudero, P., Raijmakers, M. E., 2013. What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. Journal of Phonetics 41 (5), 307–319.

Wieling, M., Montemagni, S., Nerbonne, J., Baayen, R. H., 2014. Lexical differences between tuscan dialects and standard italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. Language 90 (3), 669–692.

Wiener, S., Turnbull, R., 2015. Constraints of tones, vowels and consonants on lexical selection in mandarin chinese. Language and Speech, 0023830915578000.

Wood, S., 2006. Generalized additive models: an introduction with R. CRC press.

Wood, S., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73 (1), 3–36.

Appendix A. Model summary Distance from Target Experiment 1

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	218.1934	2.4675	88.4269	< 0.0001
Condition=high variance	1.8210	3.1827	0.5722	0.5672
Target Position=bottom right	24.5794	1.1176	21.9920	< 0.0001
Target Position=top left	-19.5370	1.1153	-17.5176	< 0.0001
Target Position=top right	6.3936	1.0787	5.9272	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Time, VOT)	65.7065	67.7241	98.4949	< 0.0001
ti(Time, Target Pos=bottom left)	1.0020	1.0021	895.7533	< 0.0001
ti(Time, Target Pos=bottom right)	3.9897	3.9996	360.9250	< 0.0001
ti(Time, Target Pos=top left)	3.9793	3.9991	321.4577	< 0.0001
ti(Time, Target Pos=top right)	3.9414	3.9965	254.7427	< 0.0001
s(Time, SubjectTarget)	1827.0807	2145.0000	11.2009	< 0.0001

Appendix B. Model summary Distance from Competitor Experiment 1

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	328.7309	2.3943	137.2987	< 0.0001
Condition=high variance	1.1495	3.2648	0.3521	0.7248
Competitor Position=bottom right	22.3582	1.1010	20.3079	< 0.0001
Competitor Position=top left	-24.9015	1.1028	-22.5794	< 0.0001
Competitor Position=top right	5.4476	1.1286	4.8268	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Time, VOT, Cond=low variance)	53.4448	60.7526	14.0871	< 0.0001
te(Time, VOT, Cond=high variance)	50.5629	59.2332	55.6577	< 0.0001
ti(Time, Comp Pos=bottom left)	3.7986	3.8286	84.5055	< 0.0001
ti(Time, Comp Pos=bottom right)	3.9394	3.9480	120.1620	< 0.0001
ti(Time, Comp Pos=topleft)	3.9687	3.9731	118.9750	< 0.0001
ti(Time, Comp Pos=top right)	3.7356	3.7714	87.2358	< 0.0001
s(Time, SubjectTarget)	1707.9899	2143.0000	8.8713	< 0.0001

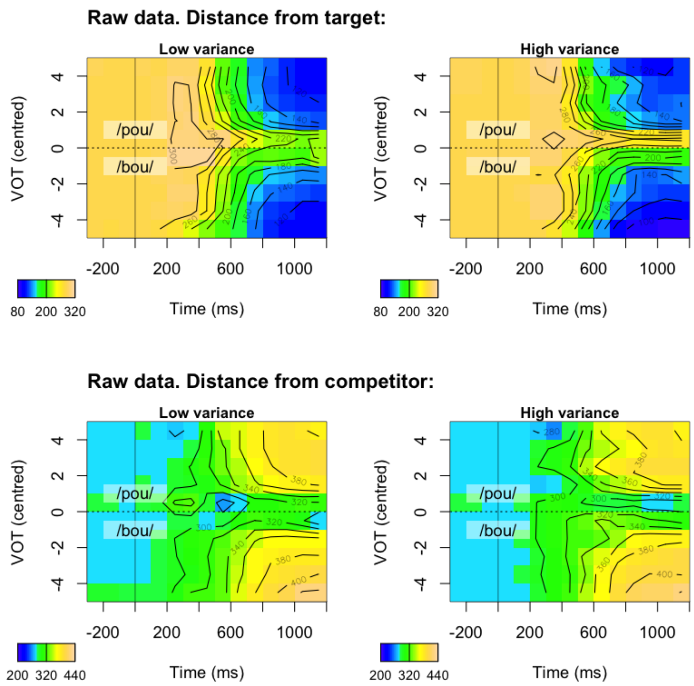
Appendix C. Model summary Distance from Target Experiment 2

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	233.5211	2.7040	86.3600	< 0.0001
Conditionw	0.2531	3.9043	0.0648	0.9483
Target Pos=bottom right	13.1427	1.2193	10.7791	< 0.0001
Target Pos=top left	-15.0850	1.2133	-12.4332	< 0.0001
Target Pos=top right	-1.2133	1.1551	-1.0504	0.2935
B. smooth terms	edf	Ref.df	F-value	p-value
te(Time, pitch, Cond=low variance)	62.0047	66.4747	87.4145	< 0.0001
te(Time, pitch, Cond=high variance)	63.7441	68.0654	81.5326	< 0.0001
ti(Time, Target Pos=bottom left)	1.1556	1.1963	676.2567	< 0.0001
ti(Time, Target Pos=bottom right)	3.9791	3.9969	273.9594	< 0.0001
ti(Time, Target Pos=top left)	3.9738	3.9958	261.2926	< 0.0001
ti(Time, Target Pos=top right)	2.8467	3.3261	260.6682	< 0.0001
s(Time, SubjectTarget)	873.1670	1049.0000	14.9350	< 0.0001

Appendix D. Model summary Distance from Competitor Experiment 2

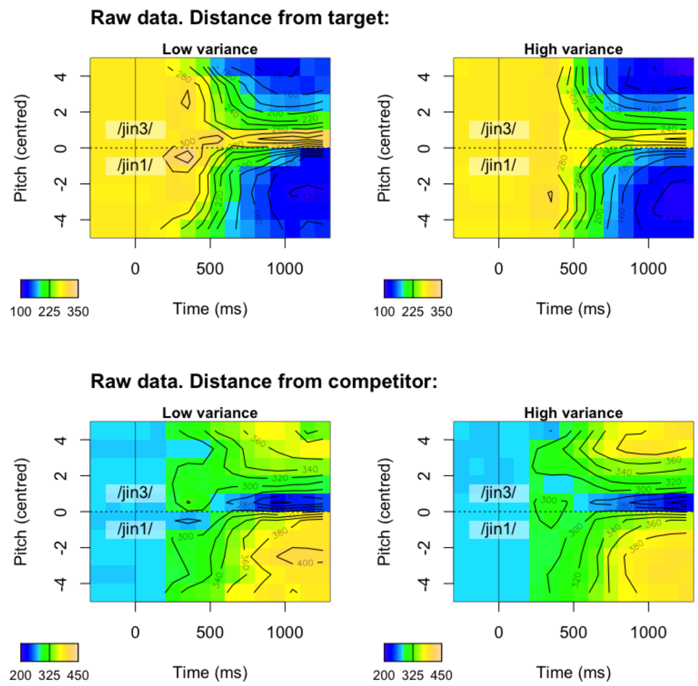
A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	324.2913	2.4692	131.3371	< 0.0001
Condition=high variance	0.0748	3.4645	0.0216	0.9828
Competitor Position=bottom right	9.5556	1.1813	8.0890	< 0.0001
Competitor Position=top left	-23.4422	1.1812	-19.8457	< 0.0001
Competitor Position = top right	-1.8628	1.2080	-1.5420	0.1231
B. smooth terms	edf	Ref.df	F-value	p-value
te(Time, pitch, Cond=low variance)	50.0866	57.7930	82.9908	< 0.0001
te(Time, pitch, Cond=high variance)	51.9104	60.6567	78.8503	< 0.0001
ti(Time, Comp Pos=bottom left)	3.7987	3.8324	130.6464	< 0.0001
ti(Time, Comp Pos=bottom right)	3.8077	3.8361	111.8629	< 0.0001
ti(Time, Comp Pos=top left)	3.9669	3.9721	127.8361	< 0.0001
ti(Time, Comp Pos=top right)	3.7000	3.7449	105.7310	< 0.0001
s(Time, SubjectTarget)	848.9467	1049.0000	12.6510	< 0.0001

Appendix E. Raw data Experiment 1



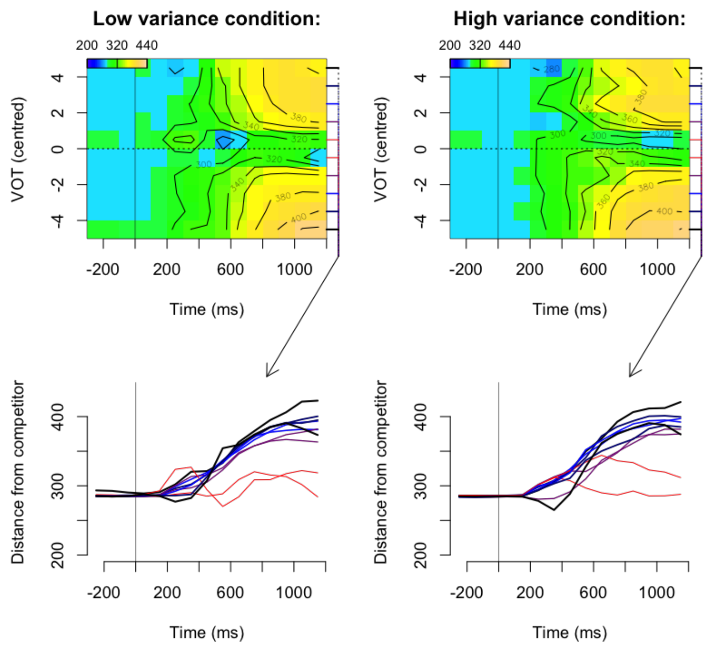
Raw data for target distance (top row) and competitor distance (bottom row) over time per VOT value in the low-variance (left panels) and high-variance conditions (right panels) in Experiment 1. Data was aggregated to 10 Hz (100 ms intervals) for the purposes of plotting. Time is on the x-axis. Centred VOT value is on the y-axis. Category means are at VOT -2.5 (for the unaspirated stimuli, e.g. bou2) and 2.5 (for the aspirated stimuli, e.g. pou2). Distance from the target/competitor is on the z-axis, represented by colour codes. Higher values (shown in yellow) indicate a relatively greater distance; lower values (shown in blue) indicate a relatively shorter distance. The key at the bottom left of each panel shows the corresponding pixel values and z-limits for each model plot. Note that the height range differs between the target and competitor: the target plots range between 80 and 320 pixels, whereas the competitor plots range between 200 and 440 pixels. (The scale is the same). To assist with interpretation of the topographical plots, an illustration showing the relation of the topographical plots of to line plots of the same raw data is provided in Appendix G.

Appendix F. Raw data Experiment 2



Raw data for target distance (top row) and competitor distance (bottom row) over time per pitch value in the low-variance (left panels) and high-variance conditions (right panels) in Experiment 2. Data was aggregated to 10 Hz (100 ms intervals) for the purposes of plotting. Time is on the x-axis. Centred pitch value is on the y-axis. Category means are at pitch -2.5 and 2.5. Distance of fixations from the target/competitor is on the z-axis, represented by colour codes. Higher values (shown in yellow) indicate a relatively greater distance; lower values (shown in blue) indicate a relatively shorter distance. The key at the bottom left of each panel shows the corresponding pixel values and z-limits for each model plot. Note that the height range differs between the target and competitor. (The scale is the same).

Appendix G. Illustration of the relation between topographic plots and line plots.



This illustration is intended to assist with interpretation, particularly for readers who are unfamilar with topographic plots. The plots show the raw data for Competitor Distance in Experiment 1. The same data are represented in two ways. In all panels, time is on the x-axis. In the topographic plots (upper panel), centred VOT value is plotted on the y-axis. In the line plots (lower panel), in contrast, centred VOT value is represented as individual, colour-coded lines. For each value of centred VOT, the lines at the right edge of the topographic plot panels indicate the line colour in the line plot and the corresponding location on y-axis of the topographic plot. In the topographic plots, distance from the competitor is plotted on the z-axis, represented by colour codes. Higher values (shown in yellow) indicate a relatively greater distance; lower values (shown in blue) indicate a relatively shorter distance. The key at the top left of each panel shows the corresponding pixel values and z-limits for each model plot. In the line plots, in contrast, distance from competitor is represented on the y-axis.