Accuracy-latency association in discrimination of L2 vowel contrasts

Yizhou Wang[a], Rikke L. Bundgaard-Nielsen[b], Brett J. Baker[a], Olga Maxwell[a]

[a]School of Languages and Linguistics, the University of Melbourne
[b]MARCS Institute for Brain, Behaviour and Development, University of Western Sydney
yizwang3@unimelb.edu.au, rikkelou@gmail.com, bjbaker@unimelb.edu.au
omaxwell@unimelb.edu.au

Abstract

This study investigates the efficacy of using response time-based metrics for assessing difficulty levels in L2 vowel discrimination. L1 Mandarin listeners discriminated six Australian English vowel contrasts with different cross-linguistic phonological statuses (cross-boundary vs within-category, according to PAM [1]). Results suggest that latency metrics, similarly to accuracy measures, can indicate the level of difficulty in AXB discrimination. Results also show that the correlation between accuracy and latency metrics is conditioned by the phonological status of the L2 vowel contrasts being tested. Strong accuracy-latency associations exist in cross-boundary contrasts, but no clear correlations are found for within-category pairs.

Index Terms: vowel discrimination, latency, accuracy

1. Introduction

It is well known that second language (L2) listeners at times encounter difficulty in discriminating target language phones that are non-contrastive in their native (L1) phonology inventory, e.g., it is notoriously difficult for Japanese listeners to discriminate the English /r/-/l/ contrast [2], [3]. This difficulty can be explained in terms of patterns of cross-language assimilation of L2 phones, according to the Perceptual Assimilation Model (PAM) [1] and its extension for L2 listening (PAM-L2) [4]. The central premise of PAM/PAM-L2 is that an L2 phone will be assimilated, or phonologically mapped, to an L1 category that is perceptually similar. For a pair of L2 phones, discrimination can be particularly difficult if both phones are perceived as equally good exemplars of a single L1 category, forming a Single-category (SC) pair. Discrimination can be relatively difficult when both L2 phones are perceived as exemplars of a single L1 category, but with different levels of phonetic similarity, i.e., they form a Category-goodness (CG) pair. If a pair of L2 phones are perceived as exemplars of two distinct L1 categories, they form a Two-category (TC) assimilation pair, for which good discrimination performance is predicted. Lastly, if one L2 phone is assimilated to an L1 category, but the other is not successfully mapped to any L1 category, they form an Uncategorised-categorised (UC) pair, which should also lead to good discrimination performance. TC and UC pairs thus represent *cross-boundary* contrasts, while CG and SC pairs represent *within-category* differences. In general, the discriminability of L2 contrasts follows a hierarchy based on PAM/PAM-L2 types, such that TC/UC > CG > SC [5].

While PAM/PAM-L2 has been shown to be a successful model in predicting the discriminability of non-native phones by analysing cross-language category mapping, support for the model has come primarily from studies relying on accuracy measures, i.e., percent correct (Pc), whilst some PAM-based studies also use latency measures as additional measures [6]–[8]. Accuracy is, however, an offline metric indicating end-point performance, and it cannot capture difficulty encountered during online processing, except as an inference.

To address this shortcoming, theories such as Automatic Selective Perception (ASP) [9], [10] call for the inclusion of latency measures such as response time (RT) and event-related potential (ERP) as additional indices for evaluating perceptual performance. In particular, the ASP model posits that automatic and attentional processing constitute a continuum of "effort", which can be defined in terms of response latency measures, e.g., mean response time (MRT). In particular, the ASP model predicts that discriminating within-category contrasts (i.e., CG/SC in PAM) requires a mode of processing which is relatively slow and susceptible to adverse listening conditions. In contrast, discriminating cross-boundary pairs (i.e., TC/UC in PAM) uses a fast mode of speech processing, which is robust even in adverse conditions.

In addition, computational cognition research [11]–[13] suggests response time variability (RTV, which is defined as the intra-participant standard deviation of response times) is also an important metric for evaluating discrimination performance. While MRT captures the central tendency of discrimination RT, RTV reflects perceptual instability during the decision-making process. In the present study, we aim to explore the association of three discrimination performance metrics, including accuracy (Pc), automaticity (MRT), and instability (RTV) in L1 Mandarin listeners' discrimination of six Australian English (AusE) vowel contrasts, /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/, which potentially form different PAM/PAM-L2 contrasts. For instance, Mandarin has only one low vowel /a/, which can have various allophones including [ɐː] (in open syllables), [ɐ] (in closed syllables), [æ] (after /j/ and before /n/) [14]. Therefore, AusE contrasts /æ/-/ɐ/ and /ɐː/-/ɐ/ can be difficult for Mandarin listeners because these vowels are not contrastive phones in Mandarin. Similarly, /iː/-/ɪ/ can also be challenging since Mandarin does not have duration contrasts [14]. Additionally, Mandarin does not have mid front monophthongs /e/ or /eː/, and therefore AusE /e/ and /eː/ can be deemed as "unfamiliar" vowels for Mandarin listeners. Lastly, Mandarin has a mid-vowel /ɤ/ which is potentially similar to AusE /ɜː/ in perception. Therefore, the six contrasts cover a wide range of phonological scenarios involving both "familiar" and "unfamiliar" vowels for Mandarin listeners. More specifically, this paper aims to address two research questions:

(1) Can response latency measures (MRT and RTV), like accuracy (Pc), serve as effective indices of discrimination

difficulty for L2 vowel pairs of different PAM/PAM-L2 assimilation types?

(2) How are accuracy and latency metrics correlated with each other across and within different PAM/PAM-L2 assimilation pairs?

2. Methods

2.1. Participants

The participants were twenty native Mandarin speakers (M_{age} = 26.4, *SD* = 3.2; 17 females), who were post-graduate students at an Australian university with a mean Length of Residence (LoR) in Australia of 1.5 years (*SD* = 1.4). All spoke English as an L2 and had an average English learning history of 15.6 years (*SD* = 4.0) in classroom-based settings prior to their arrival in Australia. All participants originally came from PR China, and all were native speakers of Standard Mandarin [14]. None reported fluency in a third language. Research suggests that L2 speech proficiency is influenced by language dominance [15], i.e., the relative linguistic command in each language, and therefore all participants completed a Bilingual Language Profile (BLP) questionnaire [16]. On average, their Mandarin dominance score (*M* = 200, *SD* = 10) was significantly higher than English (*M* = 90, *SD* = 14), *t* = 23.4, *p* < .0001, indicating that our participants were not balanced bilingual speakers.

2.2. Stimuli

Two female native speakers of AusE were recorded producing the eight relevant target vowels in /hVba/ carrier pseudowords, which minimise coarticulatory effects and provide a controlled phonological structure [17]–[19]. The speakers were instructed to produce the vowels in a clear citation style [17], [20], [21], and they produced each stimulus pseudoword multiple times. All stimuli were inspected in the phonetic software Praat [22], and the "best" instances were selected to be used in the perception experiments on the basis of voice and phonetic quality. The average acoustic measurements are reported in Table 1.

Table 1. *Acoustic properties of the vowel stimuli.*

AusE V	Dur (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)
/ɐː/	221	841	1247	3168
/ɐ/	78	894	1313	3037
/æ/	110	793	1616	3133
/eː/	225	645	2208	3152
/e/	88	638	2105	2968
/ɜː/	207	519	1626	2827
/iː/	171	441	2779	3378
/ɪ/	72	465	2481	3276

2.3. Experimental design

To answer the research questions, we designed two tasks. The first experiment was a perceptual assimilation task, which examines L2-to-L1 similarities, where the results will inform the PAM/PAM-L2 type for each vowel contrast. In this task, participants first watched an instruction video explaining the purpose and structure of the task. They were then instructed to complete the task in a quiet environment wearing headphones. The experiment consisted of 80 randomised trials (eight vowels × two speakers × five repetitions). On each trial, participants first heard an AusE /hVba/ token and were then asked to categorise the target vowel into a native language (Mandarin) category. We adopted the *whole-system* approach [17], [18] and provided all phonotactically attested L1 categories as categorisation options, including all Mandarin monophthongs, diphthongs, and triphthongs: /a, i, y, u, o, ɤ, wə, yə, jə, ja, wa, aj, aw, əw, əj, jaw, waj, iəw, wəj/. Responses were made with the help of Mandarin keywords labelled on a virtual keyboard with *Pinyin* orthography. After each categorisation, participants heard the stimulus again and were prompted to provide a goodness rating based on the perceived similarity between the stimulus and the response on a seven-point scale, where 1 = "very different", 7 = "very similar".

The second experiment was an AXB discrimination task, which is the conventional paradigm for testing discrimination in studies based on PAM/PAM-L2 [5], [18], [23]. Six AusE vowel pairs, /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/, were tested. On each trial, participants heard three consecutive /hVba/ stimuli and were asked to respond whether the middle stimulus (X) was more similar to the first stimulus (A) or the third one (B). In each triplet, A and B represented two different L2 vowel categories while X was phonologically identical to either A or B. Participants made responses by pressing the key "F" (X = A) or key "J" (X = B) on their keyboard. We presented participants with four counterbalanced triplet types (AAB, ABB, BAA, and BBA), ensuring that the participants were unable to predict the correct answer. In all trials, stimulus A and B were produced by one speaker, and X produced by the other. The task had 192 trials in total (6 pairs × 4 triplets × 2 speakers × 4 repetitions).

To ensure that the AXB task taps into phonological processing, rather than phonetic processing, we set the interstimulus interval (ISI) at 1000 ms [24]. On each trial, participants responded within 2000 ms, and response RTs were systematically measured from the offset of the last stimulus presentation to the time of response. Three metrics are calculated for each participant to assess their discrimination performance, including Pc (i.e., percentage correct), MRT (i.e., the average response time in correct trials), and RTV (i.e., the standard deviation of response times in correct trials). The experiments were developed and delivered using PsyToolkit 3.0 [25], [26].

3. Results

3.1. Assimilation results

The categorisation responses were averaged for participants and summarised in Table 2. By applying the 70% criterion in PAM-based studies [5], [17], [18], six out of eight AusE vowels were deemed as categorised: three AusE vowels, /ɐː/, /ɐ/, and /æ/, were consistently categorised as instances of the Mandarin low vowel /a/; AusE vowels /iː/ and /ɪ/ were consistently categorised as instances of the Mandarin high front vowel /i/, and AusE /ɜː/ was categorised as an instance of the Mandarin mid vowel /ɤ/. AusE /eː/ and /e/ were not assimilated into any Mandarin vowel category consistently, leaving them as uncategorised phones. Both were perceived as similar to Mandarin vowels /ej/ and /aj/, and to a lesser degree, /je/ and /ɤ/.

The assimilation status of individual vowels determines the PAM/PAM-L2 types of each vowel pair in /æ/-/ɐ/, /ɐː/-/ɐ/, /iː/-/ɪ/, /æ/-/e/, /eː/-/æ/, and /ɜː/-/e/. Since the vowels in the first three pairs were categorised into a single Mandarin vowel, they form either an SC or CG pair, depending on whether there is a perceptible difference in the goodness rating measure. Both

AusE /æ/ and /ɐ/ were perceived as relatively poor exemplars of Mandarin /a/ with a mean goodness score of 4.75 (*SD* = 1.21) and 4.78 (*SD* = 1.17), respectively. We built a linear mixed-effects model (LMM) at the decision level with random slopes and intercepts set for all participants. Then, a Wald Chi-squared test compared the mean scores, and there was no significant difference, $\chi^2(1) = 0.092$, $p = .762$. Therefore, we deemed /æ/-/ɐ/ as an SC pair.

Table 2. *Assimilation matrix of the AusE vowels: Percentage of responses.*

V	/a/	/i/	/ɤ/	/ej/	/aj/	/je/
/ɐ:/	96	1	3			
/ɐ/	96		2		1	
/æ/	76	1	3	5	14	
/e:/	1	1	10	46	30	12
/e/	4	1	9	45	33	8
/ɜ:/			99		1	
/i:/		87	2	9		1
/ɪ/		73	3	19		5

Next, AusE /ɐ:/ was perceived as a relatively good exemplar of Mandarin /a/ with a mean goodness score of 5.49 (*SD* = 1.04). For /ɐ:/-/ɐ/, we built another LMM which revealed a significant difference of goodness, $\chi^2(1) = 10.873$, $p < .001$. We therefore deemed /ɐ:/-/ɐ/ as a CG pair. Lastly, while both AusE /i:/ and /ɪ/ were both categorised as Mandarin /i/, the former was perceived as a better exemplar than the latter (mean goodness = 5.04, *SD* = 1.16, and mean goodness = 4.10, *SD* = 1.25, respectively). When analysed with an LMM, the difference was significant, $\chi^2(1) = 13.438$, $p < .001$. Therefore, /i:/-/ɪ/ formed another CG pair. For the last three pairs (/æ/-/e/, /e:/-/æ/, and /ɜ:/-/e/), the assimilation patterns resulted in one uncategorised vowel (/e:/ or /e/) and one categorised vowel (/æ/ or /ɜ:/), and thus they formed three UC pairs. Based on the prediction that discrimination accuracy should form a hierarchy that TC/UC > CG > SC [5], we predict that Mandarin listeners will show good discrimination performance in AusE /æ/-/e/$_{UC}$, /e:/-/æ/$_{UC}$, and /ɜ:/-/e/$_{UC}$, relatively good discrimination performance in AusE /ɐ:/-/ɐ/$_{CG}$ and /i:/-/ɪ/$_{CG}$, and poor performance in AusE /æ/-/ɐ/$_{SC}$.

3.2. Discrimination results

The listeners' performance in the AXB discrimination task is summarised in Table 3. The lowest accuracy was found in the SC pair /æ/-/ɐ/ (Pc = 61%), and the highest accuracy was found in the UC pair /ɜ:/-/e/ (Pc = 88%). For all six pairs, the accuracy measure (Pc) differed significantly when checked by an LMM, $\chi^2(5) = 78.208$, $p < .001$. Similarly, we built an LMM for MRT, which also revealed a significant effect of contrast, $\chi^2(5) = 81.809$, $p < .001$. We similarly found a significant effect for RTV, $\chi^2(5) = 45.55$, $p < .001$. To summarise, all AusE vowel pairs differed significantly in terms of the three metrics obtained.

A series of Tukey-adjusted *post hoc* tests (see Table 4) revealed significant differences between AusE vowel pairs of different PAM/PAM-L2 contrast types. Additionally, the three metrics, i.e., Pc, MRT, and RTV, showed some level of both commonality and complementarity: All three metrics detected a significant difference between /æ/-/ɐ/$_{SC}$ and /e:/-/æ/$_{UC}$, /æ/-/ɐ/$_{SC}$ and /ɜ:/-/e/$_{UC}$, /ɐ:/-/ɐ/$_{CG}$ and /ɜ:/-/e/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /ɜ:/-/e/$_{UC}$. In addition, the Pc metric captured differences that were not shown in the MRT and RTV patterns, between /æ/-/ɐ/$_{SC}$ and /i:/-/ɪ/$_{CG}$, /æ/-/ɐ/$_{SC}$ and /æ/-/e/$_{CG}$, and between /ɐ:/-/ɐ/$_{CG}$ and /æ/-/e/$_{UC}$. The MRT metric uniquely captured significant differences between /ɐ:/-/ɐ/$_{CG}$ and /i:/-/ɪ/$_{CG}$, /i:/-/ɪ/$_{CG}$ and /e:/-/æ/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /æ/-/e/$_{UC}$, which the Pc metric did not capture. Similarly, RTV captured the difference between /ɐ:/-/ɐ/$_{CG}$ and /e:/-/æ/$_{UC}$, /i:/-/ɪ/$_{CG}$ and /e:/-/æ/$_{UC}$, and between /i:/-/ɪ/$_{CG}$ and /æ/-/e/$_{UC}$, where the Pc metric did not show a significance. Importantly, all observed significant differences were consistent with the prediction of discrimination performance based on PAM/PAM-L2, such that TC/UC > CG > SC [5].

Table 3. *AXB discrimination results. Standard deviations are shown in parentheses.*

Pair	PAM	Pc (%)	MRT	RTV
/æ/-/ɐ/	SC	61 (15)	633 (203)	382 (77)
/ɐ:/-/ɐ/	CG	71 (13)	565 (167)	387 (118)
/i:/-/ɪ/	CG	74 (11)	693 (183)	417 (98)
/æ/-/e/	UC	79 (11)	467 (127)	289 (83)
/e:/-/æ/	UC	83 (11)	500 (176)	327 (100)
/ɜ:/-/e/	UC	88 (13)	528 (158)	290 (118)

Table 4. *P-values in pairwise comparisons. Significant comparisons are set in bold face.*

Contrast	1	2	3	4	5
1 /æ/-/ɐ/$_{SC}$	-				**Pc**
2 /ɐ:/-/ɐ/$_{CG}$	.065	-			
3 /i:/-/ɪ/$_{CG}$	**.007**	.973	-		
4 /æ/-/e/$_{UC}$	**< .001**	.138	.519	-	
5 /e:/-/æ/$_{UC}$	**< .001**	**.007**	.065	.887	-
6 /ɜ:/-/e/$_{UC}$	**< .001**	**< .001**	**.001**	.121	.668
1 /æ/-/ɐ/$_{SC}$	-				**MRT**
2 /ɐ:/-/ɐ/$_{CG}$	.394	-			
3 /i:/-/ɪ/$_{CG}$	.553	.007	-		
4 /æ/-/e/$_{UC}$	**< .001**	.074	**< .001**	-	
5 /e:/-/æ/$_{UC}$	**.004**	.460	**< .001**	.935	-
6 /ɜ:/-/e/$_{UC}$	**< .001**	**.003**	**< .001**	.888	.338
1 /æ/-/ɐ/$_{SC}$	-				**RTV**
2 /ɐ:/-/ɐ/$_{CG}$	.999	-			
3 /i:/-/ɪ/$_{CG}$	.752	.856	-		
4 /æ/-/e/$_{UC}$	**.005**	**.003**	**< .001**	-	
5 /e:/-/æ/$_{UC}$	.262	.177	**.008**	.670	-
6 /ɜ:/-/e/$_{UC}$	**.006**	**.003**	**< .001**	.999	.688

3.3. Correlation between performance metrics

The second aim of the present study is to analyse how and to what extent the three implemented discrimination metrics (Pc, MRT, and RTV) correlate with each other. To address this, we carried out a series of Pearson's *r* correlation analyses based on the standardised score (*z*-score) of the three metrics within each vowel pair and across different vowel pairs, see Table 5. The results showed that the two latency-based metrics, MRT and RTV, were positively and significantly (p's < .01) correlated in four out of six vowel pairs (except /æ/-/ɐ/$_{SC}$ and /e:/-/æ/$_{UC}$). When the vowel pairs were pooled together based on phonological status, both within-category pairs (SC/CG) and cross-boundary pairs (UC) showed a very high correlation coefficient ($r = .595$ and $r = .648$, respectively, p's < .001). When all six AusE vowel pairs were pooled together, the correlation between MRT and RTV was still very high ($r = .695$, $p < .001$). In other words, the MRT-RTV correlation

tended to be similarly robust for within-category (SC/CG) as well as cross-boundary pairs (UC).

Table 5. *Correlation between metrics.*

Contrast	Pc ~ MRT	Pc ~ RTV	MRT ~ RTV
1. /æ/-/ɐ/$_{SC}$	0.036	–.063	0.43
2. /ɐ:/-/ɐ/$_{CG}$	–.180	–.188	.696***
3. /i:/-/ɪ/$_{CG}$	0.166	0.423	.649**
4. /æ/-/e/$_{UC}$	–.539*	–.694***	.803***
5. /e:/-/æ/$_{UC}$	–.488*	–.098	0.243
6. /ɜ:/-/e/$_{UC}$	–.493*	–.459*	.697***
SC/CG pairs	0.003	0.057	.595***
UC pairs	–.520***	–.416***	.648***
All pairs	–.428***	–.347***	.695***

Surprisingly, our analyses revealed no significant correlation between Pc and MRT for the within-category pairs (p's > .05,) or when the SC and CG pairs combined (p > .05). However, the Pc-MRT correlation was significant for all three UC pairs, i.e., /æ/-/e/, /e:/-/æ/, and /ɜ:/-/e/ (r = –.539, –.488, and –.493, p = .014, .029, and .027, respectively). When the three UC pairs were pooled together, the Pc-MRT correlation was significant (r = –.520, p < .001). This suggests that the Pc-MRT correlation is significant for cross-boundary (UC) pairs but not within-category (SC/CG) pairs. Lastly, we found no significant correlations between Pc and RTV in the within-category pairs (p's > .05) nor for the SC and CG pairs combined (p > .05). But we did find significant Pc-RTV correlations for the UC pairs /æ/-/e/ (r = –.694, p < .001) and /ɜ:/-/e/ (r = –.459, p = .042). When all three UC pairs were pooled together, we again found a significant Pc-RTV correlation (r = –.416, p = .001). These results show that both Pc-MRT and Pc-RTV tend to be negatively correlated for cross-boundary (UC) pairs, but no clear association is found for within-category (SC/CG) pairs. In other words, the strength of accuracy-latency association in L2 vowel discrimination depends on the phonological status, i.e., the PAM/PAM-L2 type of the vowel pair in question.

4. General discussion

The present paper addresses a number of theoretical and methodological questions. Firstly, we examined whether discrimination latency measures such as MRT and RTV, like Pc, can serve as effective indices of discrimination difficulty for L2 vowel pairs of different PAM/PAM-L2 assimilation types. Our results suggest that all three metrics can be used to indicate the discrimination difficulty of L2 vowel discrimination, as predicted by the PAM/PAM-L2 theory [1], [4]. By analysing six AusE vowel contrasts and their pairwise comparisons (Table 4), the results suggest that the accuracy and latency metrics have commonalities and complementarities. Accuracy (Pc) can at times capture between-pair differences that the latency metrics (MRT and RTV) do not show a difference for, and conversely sometimes latency metrics detect differences that are not visible in the accuracy measure.

Generally, our results support the ASP model's [9], [10] prediction that L2 vowels of different phonological statuses, i.e., perceptual assimilation types, will lead to different latency patterns: L2 listeners show accurate and fast speech processing for cross-boundary contrasts, while discrimination of within-category pairs is more difficult and less automatic. Typically, SC and CG pairs tend to rely on L2 listeners directing their cognitive resources to the reminiscent phonetic details, which is deliberately discouraged by the long ISI value in the AXB task [24]. On the contrary, UC pairs, which represent cross-boundary contrasts, are sufficiently discriminated based on coarse-grained phonological coding, which is resistant to the decay of short-term sensory memory in a long ISI design [24].

We also explored the effectiveness of including RTV as an additional latency metric, as it is suggested in computational cognition research [11], [12] to be informative of neural instability during the discrimination process. Thus, our findings also extend the ASP model's premise that the cross-boundary L2 contrasts are processed in a more *stable* manner than within-category contrasts. Methodologically, the results suggest that discrimination research should use both accuracy and latency since they can tap into different aspects of discrimination performance. Clearly, the resolution of the accuracy measure depends on the number of trials. For instance, with 32 trials testing an L2 vowel pair in discrimination, we could obtain a resolution of 100%/32 or 3.125%, and when fewer trials were tested, we would obtain a more coarse-grained measure at the participant level. On the contrary, the latency metrics MRT and RTV are gradient in nature, and their baseline levels can be determined by a control condition that should not be difficult to listeners, e.g., a TC/UC pair.

The second research question asks how accuracy and latency metrics are associated and correlated with each other across and within different L2 vowel pairs. Our results suggest that the two latency-based metrics, MRT and RTV, tend to be positively correlated irrespective of the phonological status of the L2 vowel pair. This finding is consistent with previous research, which reports that MRT and RTV tend to covary in an approximately linear manner in two-choice tasks [27]. More strikingly, we observed that Pc-MRT and Pc-RTV correlations were conditioned by the phonological status of the L2 vowel pairs, i.e., their perceptual assimilation patterns, see Table 5. For within-category (SC/CG) pairs, we did not find significant correlations between accuracy and latency, but we found robust negative correlations for cross-boundary (UC) pairs, such that high accuracy measures accompany high response speed and low instability. This indicates that MRT and RTV in easy conditions, i.e., for cross-boundary (UC, potentially also TC) pairs, sufficiently reflect the cognitive demands of the decision process during an AXB discrimination task. But in difficult conditions, i.e., for within-category (CG and SC) pairs, they cannot effectively measure the cognitive demand.

Recall that the AXB task has counterbalanced speaker order and triplet types, and in case a listener responds randomly they will still achieve 50% accuracy. However, the latency measure estimated from these trials will reflect properties of guesses rather than the targeted cognitive process in discrimination procedures. Although it is difficult to tease apart guesses from the dataset based on behavioural measures, it is reasonable to assume that lower accuracy measures should accompany higher proportions of guesses. On the contrary, the proportion of guesses should be relatively low in easy conditions, i.e., for cross-boundary (UC) pairs, and MRT and RTV should effectively reflect the speed and stability of the discrimination process. This finding reveals a disadvantage of latency as compared to accuracy: Guesses due to an inability to discriminate a contrast will contaminate the latency metrics [28] evaluated from an AXB task. In general, accuracy and latency metrics can still complement each other, especially when a range of difficulty levels are tested: Accuracy might fail to capture more nuanced differences when participants tend to achieve ceiling level accuracy; On the contrary, latency might not effectively reflect cognitive demands when the accuracy is relatively low.

5. Acknowledgements

We want to thank the twenty participants for their time. Special thanks also go to Debbie Loakes and Catherine Roberts for their help in generating the stimuli.

6. References

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

[2] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/r-l/ by Japanese adults learning English," *Percept. Psychophys.*, vol. 36, no. 2, pp. 131–145, 1984.

[3] K. S. MacKain, C. T. Best, and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Psycholinguist.*, vol. 2, pp. 369–390, 1981.

[4] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language experience in second language speech perception*, O.-S. Bohn, Ed. Amsterdam: John Benjamins, 2007, pp. 13–34.

[5] M. D. Tyler, C. T. Best, A. Faber, and A. G. Levitt, "Perceptual assimilation and discrimination of non-native vowel contrasts," *Phonetica*, vol. 71, no. 1, pp. 4–21, 2014.

[6] P. A. Hallé and C. T. Best, "Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 2899–2914, 2007.

[7] C. T. Best and P. A. Hallé, "Perception of initial obstruent voicing is influenced by gestural organization," *J. Phon.*, vol. 38, no. 1, pp. 109–126, 2010.

[8] A. Kilpatrick, R. L. Bundgaard-Nielsen, and B. J. Baker, "Japanese co-occurrence restrictions influence second language perception," *Appl. Psycholinguist.*, vol. 40, no. 2, pp. 585–611, 2019.

[9] W. Strange and V. L. Shafer, "Speech perception in second language learners," in *Phonology and second language acquisition*, J. G. Hansen-Edwards and M. L. Zampini, Eds. Amsterdam: Benjamins, 2008, pp. 153–192.

[10] W. Strange, "Automatic selective perception (ASP) of first and second language speech: A working model," *J. Phon.*, vol. 39, no. 4, pp. 456–466, 2011.

[11] R. Ratcliff, "A theory of memory retrieval," *Psychol. Rev.*, vol. 85, pp. 59–108, 1978.

[12] R. Ratcliff, P. L. Smith, S. D. Brown, and G. McKoon, "Diffusion Decision Model: Current issues and history," *Trends Cogn. Sci.*, vol. 20, no. 4, pp. 260–281, 2016.

[13] R. Ratcliff and G. McKoon, "The Diffusion Decision Model: Theory and data for two-choice decision tasks," *Neural Comput.*, vol. 20, no. 4, pp. 873–922, 2008.

[14] S. Duanmu, *The phonology of Standard Chinese*. Oxford, UK: Oxford University Press, 2007.

[15] J. E. Flege, I. R. A. MacKay, and T. Piske, "Assessing bilingual dominance," *Appl. Psycholinguist.*, vol. 23, no. 4, pp. 567–598, 2002.

[16] D. Birdsong, L. M. Gertken, and M. Amengual, "Bilingual Language Profile: An easy-to-use instrument to assess bilingualism. COERLL, University of Texas at Austin, TX," *https://sites.la.utexas.edu/bilingual/*, 2012.

[17] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, "Vocabulary size matters: The assimilation of second language Australian English vowels to first-language Japanese vowel categories," *Appl. Psycholinguist.*, vol. 32, no. 1, pp. 51–67, 2011.

[18] R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Tyler, "Vocabulary size is associated with second-language vowel perception performance in adult learners," *Stud. Second Lang. Acquis.*, vol. 22, pp. 433–461, 2011.

[19] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across German, French, and American English vowels: Phonetic context effects," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1111–1129, 2007.

[20] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 272–284, 2002.

[21] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech, Lang. Hear. Res.*, vol. 29, no. 4, pp. 434–446, 1986.

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9–10, pp. 341–345, 2001.

[23] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 775–794, 2001.

[24] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, pp. 35–44, 1985.

[25] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behav. Res. Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[26] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teach. Psychol.*, vol. 44, no. 1, pp. 24–31, 2017.

[27] E. J. Wagenmakers, R. P. P. P. Grasman, and P. C. M. Molenaar, "On the relation between the mean and the variance of a diffusion model response time distribution," *J. Math. Psychol.*, vol. 49, no. 3, pp. 195–204, 2005.

[28] R. Ratcliff and F. Tuerlinckx, "Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability," *Psychon. Bull. Rev.*, vol. 9, no. 3, pp. 438–481, 2002.