NON-NATIVE VOWEL PERCEPTION IN A 4IAX TASK: THE EFFECTS OF ACOUSTIC DISTANCE

Alba Tuninetti[1,2], James Whang[3] & Paola Escudero[1,2]

[1]MARCS Institute for Brain, Behaviour & Development, Western Sydney University
[2]ARC Centre of Excellence for the Dynamics of Language
[3]Saarland University
a.tuninetti@westernsydney.edu.au, research@jameswhang.net, paola.escudero@westernsydney.edu.au

ABSTRACT

The effects of acoustic versus phonetic similarity in non-native vowel perception have been the focus of many second language (L2) speech perception models, examining how non-native sounds are perceived and assimilated by listeners. These models use perceptual discrimination tasks (e.g., AX, AXB) that may elicit different modes of perception depending on the memory load and linguistic experience required in each. This paper examined how native Australian-English (AusE) speakers perceived naturally-produced Dutch vowels in a 4IAX task, less commonly used but believed to elicit more continuous perception and bypass conscious linguistic processing. Participants listened to six Dutch vowel pairs spoken by varying speakers and chosen for their acoustic distance from AusE vowels. Results showed that /ɪ-i/ was the least accurately discriminated compared to other vowel pairs, confirming predictions that L1-L2 acoustic distance is a driving force in non-native speech perception and suggesting that linguistic experience may affect perception during the 4IAX task.

Keywords: speech perception, L2, 4IAX, acoustic distance

1. INTRODUCTION

Models of second language (L2) speech perception, such as the Perceptual Assimilation Model (PAM-L2; [3]) and the Second Language Linguistic Perception model (L2LP; [7], [20]), examine how L2 learners perceive non-native or L2 sounds as a function of the similarity between their first language (L1) and the non-native language. This similarity is operationalized on the basis of phonetic or acoustic distance between the two languages and is usually tested in discrimination tasks that require listeners to compare various sounds.

In many perceptual discrimination tasks, listeners are tasked with responding whether two speech sounds are the same or different (AX) or whether one sound belongs to the category of two other sounds (AXB). For example, Alispahic et al. [2] used an XAB discrimination task to examine how native Australian English (AusE) and native Spanish speakers perceived naturally produced Dutch vowels. Their results showed that the acoustic distance between the contrasts was more predictive of participants' performance compared to vowel inventory size of the participants' L1. These results contrast with Iverson and Evans [10] who proposed that German and Norwegian speakers are better than Spanish and French speakers in identifying British English vowels because German/Norwegian vowel inventories are supersets of English. Recent work has used variants of the AXB task to examine discrimination profiles in cross-linguistic tasks to elicit categorical perception based on existing linguistic knowledge [5, 6, 19].

However, the results of AXB tasks have been shown to vary with inter-stimulus interval (ISI), lexical or linguistic knowledge, or synthetic versus natural stimuli [15]. A less-used perceptual discrimination task that is thought to impose fewer demands on short-term memory and allow for more continuous (as opposed to categorical) perception is the 4-interval forced choice (4IAX) task [16]. Listeners are asked to discriminate between two *pairs* of stimuli and choose which *pair* is different. Importantly, this task is believed to allow listeners to bypass their linguistic knowledge, such as the influence of language background or lexical access [16]. However, work with native English and native Japanese speakers showed that the frequency of certain L1 phonemes influenced their sensitivity to a continuum of VCV sequences in a 4IAX task [12]. This suggests that language-specific experience can influence tasks that are thought to elicit more 'acoustic' rather than phonetic category-based perception showing that low-level auditory processing can be influenced by language exposure.

Additionally, even though synthesized versus naturally-produced tokens can elicit different performance in perceptual discrimination tasks [17], the existing 4IAX literature has largely used synthesized stimuli to examine categorical versus continuous perception [12, 16]. However, recent work has shown that naturally-produced stimuli may be perceived and processed differently very early

during auditory processing compared to synthetic stimuli. Specifically, listeners are unable to filter out speaker characteristics (such as fundamental frequency, or F0) during an auditory oddball discrimination paradigm with naturally-produced speech tokens [18] contrasting existing results with synthetic stimuli [11]. This suggests that listeners do not automatically disregard speaker information during low-level auditory processing with natural speech tokens.

During natural speech perception, listeners must also cope with input from multiple speakers in real-world listening situations. This variability in speakers is usually normalized (or filtered out) during speech processing to focus on the linguistically-meaningful information (e.g., vowels, consonants, words) [1]. Indeed, participants' success ignoring speaker variability during certain speech perception tasks suggests they are skilled at extracting the relevant information [2, 5, 6]. However, given that differences have been found in how listeners normalize synthetic versus natural speech [11, 18] and, as discussed above, the 4IAX literature has largely used synthetic stimuli, it is currently unclear what the 4IAX design might reveal about pre-phonetic speaker normalization processes. Examining how listeners perceive naturally-produced non-native speech sounds in this task would help clarify the effects of linguistic experience during perceptual discrimination, since it is more akin to what listeners encounter daily.

The current study addresses this issue by presenting native AusE listeners with naturally-produced, isolated Dutch vowels in a 4IAX task. Vowels produced by three different speakers were used to examine how listeners use speaker variability during categorical perception. Additionally, we were interested in probing how this distance in 'non-linguistic' (speaker identity) and 'linguistic' (vowel identity) information interacts during non-native discrimination tasks, since acoustic distance varies greatly between and within speakers of the same language, as we know both types of information affect non-native (or L2) speech perception [2, 18]. Therefore, in addition to including multiple speakers, Dutch vowel pairs were predicted to be 'difficult' or 'easy' to discriminate accurately by AusE speakers by measuring the acoustic distances between native AusE speaker productions and the Dutch vowels within each pair [2]. Difficult tokens were defined as vowels more similar to (or with the least acoustic distance from) an AusE vowel, and would therefore be perceived as belonging to the same category, leading to worse discrimination performance. Easy vowels were those that were the least similar to (or had the most acoustic distance from) an AusE vowel, and would be perceived as belonging to separate (or non-existent) AusE categories, leading to better discrimination performance [2]. Finally, if listeners can filter out speaker variability, we would not expect to see effects of speaker information [11]. However, if speaker information is processed in pre-phonetic stages [18], then we would expect to see an effect of speaker in discrimination accuracy.

2. METHODS

2.1. Participants

Participants were 8 native Australian English (AusE) speakers from an undergraduate institution in the greater Western Sydney region (M_{age} = 28.57; 6 females). They signed informed consent and reported no hearing or language impairments.

2.2. Stimuli

Stimuli were four naturally produced Dutch vowels from the Adank, Smit, and van Hout [1] corpus: /a/, /ɪ/, /i/, and /ɛ/. The vowels were extracted from monosyllabic Dutch syllables /sas/, /sɪs/, /sis/ and /sɛs/. The six pairs of contrasts were constructed by pairing each vowel with every other vowel: /a-ɛ/, /a-ɪ/, /a-i/, /ɛ-ɪ/, /ɛ-i/, and /ɪ-i/. These vowels were chosen based on their similarity to and acoustic distance from AusE vowels [2].

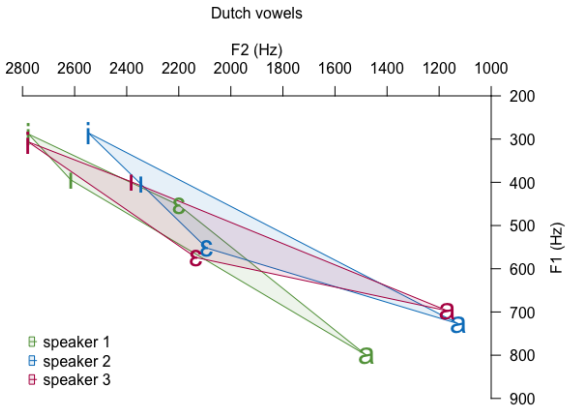
In addition, each token was produced by three different female native Dutch speakers. These speakers were chosen out of 10 possible speakers from the corpus by: (1) computing the F2:F1 ratios (Hz) of the Dutch vowels; (2) mapping their productions of these vowels onto the AusE vowel space (taken from [6] participant productions); and (3) selecting those who produced the target vowels the furthest away from the mean of all possible Dutch speakers (10 in total). This was done specifically to ensure that AusE listeners would not necessarily categorize the non-native vowels in a native category. See Table 1 for stimuli characteristics and Figure 1 for a graphic representation of the vowel space for all four Dutch vowels used as stimuli.

Table 1: Stimuli characteristics: F1, F2 means, F2:F1 ratio for each speaker.

Speaker	Vowel	F1 mean	F2 mean	F2:F1
1	a	800.5	1479.0	1.846
	ɛ	453.5	2198.5	4.848
	ɪ	394.5	2614.5	6.627
	i	288.5	2777.0	9.626
2	a	727.5	1125.5	1.547

	ɛ	552.0	2093.0	3.792
	ɪ	404.0	2345.0	5.804
	i	285.5	2548.0	8.925
3	a	697.5	1167.5	1.673
	ɛ	574.0	2134.5	3.719
	ɪ	399.0	2381.5	5.969
	i	306.5	2778.5	9.065

Figure 1: Dutch speaker vowel space for all vowels.



2.3. Design

All participants heard all six vowel contrasts produced by all three possible speaker pairs. This resulted in 18 possible combinations per participant (six vowel contrasts by three speaker pairs). For each stimulus pair, there were eight possible combinations for the 4IAX task, such that each participant heard a total of 144 trials (eight 4IAX combinations by 18 possible speaker and vowel pairs). See Table 2 for an illustration of the design.

Table 2: Example of 4IAX task design.

Trial	Speaker & Vowel
AA : AB	S1 /a/ - S2 /a/ : S1 /a/ - S2 /ɛ/
AA : BA	S1 /a/ - S2 /a/ : S2 /ɛ/ - S1 /a/
BB : AB	S2 /ɛ/ - S1 /ɛ/ : S1 /a/ - S2 /ɛ/
BB : BA	S2 /ɛ/ - S1 /ɛ/ : S2 /ɛ/ - S1 /a/
AB : AA	S1 /a/ - S2 /ɛ/ : S1 /a/ - S2 /a/
BA : AA	S2 /ɛ/ - S1 /a/ : S1 /a/ - S2 /a/
AB : BB	S1 /a/ - S2 /ɛ/ : S2 /ɛ/ - S1 /ɛ/
BA : BB	S2 /ɛ/ - S1 /a/ : S2 /ɛ/ - S1 /ɛ/

2.4. Procedure

Participants were seated in a sound-attenuated booth in front of a computer screen with a keyboard for responses. Stimuli were presented binaurally with Etymotic earphones at 70 dB SPL. They were told they would hear four speech sounds that comprised two pairs and that they needed to respond which vowel pair was *different*. If they believed the first pair they heard was different, they pressed the <z> on the keyboard; if they believed the second pair was different, they pressed the letter <m> on the keyboard. Trials were presented randomly. The inter-stimulus interval (ISI) between stimuli *within* the pair was 200 ms; the ISI *between* pairs was 500 ms [12]. Participants had 1000 ms after presentation of the second pair to make a response at which point 'no response' was recorded and the next trial began automatically. Participants completed the task in approximately 20 minutes.

3. RESULTS

Because we were interested in whether listeners would discriminate between non-native (Dutch) vowel contrasts based on their similarity to their native language (AusE), planned t-tests were conducted to examine if participants responded above chance for each vowel contrast. Contrasts /ɪ-ɛ/ (M = .48) and /ɪ-i/ (M = .47) were not discriminated above chance (*t*'s < 1); all other contrasts were discriminated above chance (all *p*'s < .005; $M_{a\text{-}ɛ}$ = .77, $M_{a\text{-}ɪ}$ = .76, $M_{a\text{-}i}$ = .84, $M_{i\text{-}ɛ}$ = .78).

We were also interested in how vowel contrast and speaker variability interacted. Our data comprised categorical responses, so participant's accuracy for each trial was entered into a mixed-effects logit model with fixed effects for our variables of interest: vowel pair and speaker pair [12] and subject and trial type as random effects. There was a main effect of vowel pair, such that participants performed worse on the /ɪ-i/ (ß = -1.62, SE = .40, *z* = -4.05, *p* < .001) and /ɪ-ɛ/ (ß = -0.97, SE = .39, *z* = -2.47, *p* = .014) contrasts. Participants showed the best performance on the /a-ɪ/ contrast (ß = 1.30, SE = 0.49, *z* = 2.66, *p* = .008). Additionally, when listening to speaker pair 2 (blue) and 3 (red), participants showed overall better discrimination compared to the other speaker pairs (ß = 1.65, SE = 0.53, *z* = 3.11, *p* = .002); this was qualified by an interaction with vowel pair. Specifically, participants had worse performance for the /ɪ-ɛ/ (ß = -1.51, SE = 0.65, *z* = -2.32, *p* = .017) and /i-ɛ/ (ß = -1.65, SE = 0.69, *z* = -2.39, *p* = .02) contrasts for speaker pair 2 and 3. No other interactions were significant.

4. DISCUSSION

Our result showed that our predictions were confirmed for vowel difficulty, such that those Dutch vowels that were perceived to belong to the same native AusE vowel category showed the worst performance. The vowel pairs /ɪ-i/ and /ɪ-ɛ/ were the most difficult for participants to perceive as

'different'. The difficulty with /ɪ-i/ is most likely due to listeners categorizing both of the Dutch vowels in the native AusE category /iː/, and the difficulty with /ɪ-ɛ/ is most likely due to the both vowels overlapping with the AusE vowel /e/. The variability between Dutch speakers in vowel productions may have contributed to listeners possibly perceiving the same Dutch vowel /ɪ/ in more than one AusE category, suggesting multiple category assimilation [e.g., 22]. Additionally, one speaker pair (speakers 2 and 3) helped listeners more than speaker pairs 1 and 2 and 1 and 3. Comparing the vowel spaces from Figure 1, it appears that speaker pair 2 (blue) and 3 (red) show the most distance between productions of /i/, even though the two speakers have significantly overlapping categories for /a/, /ɛ/, and /ɪ/. This overlap is most likely responsible for the interaction, wherein listeners show worse performance for specifically the /ɪ-ɛ/ and /i-ɛ/ contrasts with this pair.

This interaction between speakers and vowel contrasts tentatively suggests that listeners are not normalizing certain speaker characteristics during perception. Specifically, a speaker's vowel space, which can vary greatly in F1 and F2, may influence how non-native listeners discriminate their vowels and map the input onto the L1 vowel space. Here, listeners were particularly sensitive to one speaker pair, even though the speakers had relatively large overlap in their acoustic space. However, the vowel pairs with the lowest accuracy were those with vowels that showed the most overlap in that speaker pair: /i/, /ɪ/, and /ɛ/. It is not surprising that non-native listeners showed the most difficulty with those stimuli.

Post-hoc comparisons between the speaker vowel characteristics showed no significant differences in overall F1, F2 means or F2:F1 ratio across all vowels. Comparisons within speakers and across vowels were not possible due to the lack of power, but qualitatively, there do not appear to be large differences in the acoustic characteristics of the vowels for speakers 2 and 3 (see Table 1). This suggests that the voice quality of those speakers (or fundamental frequency, F0) may have influenced speech perception. Although many studies have shown that speakers are able to ignore this kind of speaker information during online speech processing, more recent studies have shown that listeners are not always able to filter out these cues with naturally-produced speech [14, 18]. In the future, examining if participants can differentiate between speakers using the same vowel identity (e.g., trials such as S1/a/-S2/a/:S1/a/:S2/a/) would elucidate whether listeners are able to perceive and use speaker information during discrimination. Indeed, using tasks that potentially elicit more 'acoustic' processing can provide information about how speaker variability may still affect fine-grained auditory processing of natural speech.

These results demonstrate that speakers can discriminate non-native vowel contrasts produced by different speakers above chance except when the vowels within the contrast are too acoustically similar to vowels in the native (L1) repertoire. This is expected, as a large body of work exists showing that similarity between phonetic contrasts influences how non-native listeners perceive and categorize those contrasts [2, 3]. However, we show this result elicited in a task setting where listeners are thought to bypass conscious linguistic knowledge and focus solely on acoustic processing. Indeed, using the same ISI as previous 4IAX tasks (as opposed to longer ones that may elicit phonological coding; see [8, 21]), prior language experience (in this case, AusE) influenced how listeners perceived and discriminated non-native vowel contrasts. This corroborates earlier work showing that native Japanese and native English speakers use their prior linguistic knowledge to discriminate between synthesized consonants [12]. The authors argue that low-level auditory processing can be influenced by linguistic experience and our results support this as well. Additionally, because we used natural speech tokens with varying speaker information, we also show that an interaction between non-linguistic (e.g., speaker) information and linguistic experience can hinder vowel discrimination to a certain extent. It remains to be seen if more exposure to the stimuli would result in better discrimination between the vowel contrasts. Because normalization to speaker variability has been shown to occur within minutes with access to lexical information [4, 9, 13], it may be the case that with isolated vowels, listeners do not have enough information or need more time to successfully normalize across all speakers, especially those who overlap on difficult (i.e., more similar to L1) vowel contrasts.

Overall, our results speak to models of L2 speech perception, showing that acoustic (as opposed to phonetic) processing can still occur with influence from linguistic experience, further supporting the idea that our perceptual systems are fundamentally shaped by our early perceptual experiences. However, testing these assumptions with natural compared to synthesized speech is important, as more studies show that there may be quantitative and qualitative differences in how humans perceive and process these. Future research should aim to examine within-participant differences to ensure that results are not due to the inherent complexity of natural speech, as evidenced by speaker variability and its effects.

5. REFERENCES

[1] Adank, P., Smits, R., van Hout, R. 2004. A comparison of vowel normalization procedures for language variation research. *J. Acoust. Soc. America*, 116, 3099–3107.

[2] Alispahic, S., Mulak, K., Escudero, P. 2017. Acoustic properties predict perception of unfamiliar Dutch vowels by adult Australia English and Peruvian Spanish Listeners. *Frontiers in Psychology*, 8, 52.

[3] Best, C., Tyler, M. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In: Munro, M.J., Bohn, O-S. (eds) *Second language speech learning: The role of language experience in speech perception and production.* Amsterdam: John Benjamins.

[4] Clarke, C.M., Garret, M.F. 2004. Rapid adaptation to foreign-accented English. *J. Acoust. Soc. America*, 116, 3647-3658.

[5] Elvin, J., Escudero, P., Vasiliev, P. 2014. Spanish is better than English for discriminating Portuguese vowels: acoustic similarity versus vowel inventory size. *Frontiers in Psychology*, 5, 1188.

[6] Elvin, J., Williams, D., Escudero, P. 2016. The relationship between perception and production of Brazilian Portuguese vowels in European Spanish monolinguals. *Loquens*, 3, e031.

[7] Escudero, P. 2009. The linguistic perception of similar L2 sounds. In: Boersma, P., Hamann, S. (eds) *Phonology in Perception.* Berlin-New York: Mouton de Gruyter, 152–190.

[8] Escudero, P., Benders, T., Lipski, S. 2009. Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, 37, 452-465.

[9] Evans, B., Iverson, P. 2004. Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *J. Acoust. Soc. America*, 115, 352-361.

[10] Iverson, P., Evans, B., 2007. Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *J. Acoust. Soc. America*, 122, 2842-2854.

[11] Jacobsen, T., Schröger, E., Alter, K. 2004. Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology*, 41, 654–659.

[12] Kataoka, R., Johnson, K. 2007. Frequency effects in cross-linguistic stop place perception: A case of /t/-/k/in Japanese and English. *UC Berkeley Phonology Lab Annual Report*, 273-301.

[13] Kralijc, T., Samuel, A. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory & Language*, 56, 1-15.

[14] Kriengwatana, B., Terry, J., Chládková, K., Escudero, P. 2016. Speaker and accent variation are handled differently: Evidence in native and non-native listeners. *PLoS ONE*, 11, e0156870.

[15] Mora, J. 2008. Methodological issues in assessing L2 perceptual phonological competence. *Proc. Phonetics in Teaching and Learning Conference*, London, 1-5.

[16] Pisoni, D., Lazarus, J. 1974. Categorical and noncategorical modes of speech perception along the voicing continuum. *J. Acoust. Soc. America*, 55, 328-333.

[17] Schouten, M.E.H., van Hessen, A.J. 1992. Modeling phoneme perception I: Categorical perception. *J. Acoust. Soc. America*, 92, 1841-1855.

[18] Tuninetti, A., Chládková, C., Peter, V., Schiller, N.O., & Escudero, P. 2017. When speaker identity is unavoidable: Neural processing of speaker identity cues in natural speech. *Brain & Language, 174*, 42-49.

[19] Tuninetti, A., Tokowicz, N. 2018. The influence of a first language: Training non-native listeners on voicing contrasts. *Language, Cognition, & Neuroscience*, 33, 750-768.

[20] van Leussen, J-W., Escudero P. 2015. Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, 6, 1000.

[21] Werker, J., Logan, J.S. 1985. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37, 35-44.

[22] Escudero, P., & Chladkova, K. 2010. Spanish listeners' perception of American and Southern British English vowels. *J. Acoust. Soc. America*,, 128, EL254-260.