Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 32–42
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Vowel perception: Effects of non-native language vs. non-native dialect

Anne Cutler *, Roel Smits, Nicole Cooper

Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Received 12 October 2004; received in revised form 2 February 2005; accepted 7 February 2005

Abstract

Three groups of listeners identified the vowel in CV and VC syllables produced by an American English talker. The listeners were (a) native speakers of American English, (b) native speakers of Australian English (different dialect), and (c) native speakers of Dutch (different language). The syllables were embedded in multispeaker babble at three signal-to-noise ratios (0 dB, 8 dB, and 16 dB). The identification performance of native listeners was significantly better than that of listeners with another language but did not significantly differ from the performance of listeners with another dialect. Dialect differences did however affect the type of perceptual confusions which listeners made; in particular, the Australian listeners' judgements of vowel tenseness were more variable than the American listeners' judgements, which may be ascribed to cross-dialectal differences in this vocalic feature. Although listening difficulty can result when speech input mismatches the native dialect in terms of the precise cues for and boundaries of phonetic categories, the difficulty is very much less than that which arises when speech input mismatches the native language in terms of the repertoire of phonemic categories available.

Keywords: Speech perception; Vowels; Perceptual confusion; Non-native language; Dialect

1. Introduction

Recognising spoken language entails correctly categorising the sounds of which speech signals are composed. If we hear the word *wrist* we have to perceive all four of its sounds [ɹɪst] correctly to be sure that we have not, for example, heard *list, rest, rift* or *risk*.[1] The identification of speech sounds is the first crucial stage in the listener's

* Corresponding author. Tel.: +31 24 352 1377; fax: +31 24 352 1213.
E-mail address: anne.cutler@mpi.nl (A. Cutler).

[1] On the other hand, in the Dutch word for *wrist*, namely *pols*, identifying the first three phonemes is enough because there is no other monosyllabic four-phoneme Dutch word beginning *pol-*.

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.001

conversion of an incoming speech signal to a meaningful representation of the speaker's intended message.

This stage of speech communication is notoriously more difficult when speaker and listener come from different language backgrounds. In particular, it is hard to make perceptual distinctions between phoneme categories of a non-native language when the native language requires no corresponding distinctions. Thus the Japanese consonant inventory contains only one category to which both the English phonemes /r/ and /l/ poorly map, and as a result distinguishing *wrist* from *list* is particularly hard for Japanese listeners to English.

The added difficulty of phoneme identification for non-native listeners appears to be as great in advantageous listening conditions as in difficult conditions. Cutler et al. (2004) presented CV and VC syllables spoken by a native speaker of American English to American listeners and to Dutch listeners proficient in English, comparing presentation under conditions of very little noise (16 dB SNR), mild noise (8 dB SNR) and moderate noise (0 dB SNR). The performance of all listeners deteriorated with increasing noise, but importantly, the effects of noise and listener background did not interact: The degree to which non-native identification fell short of native performance remained roughly constant across the three levels of noise masking compared in their study. In this case the stimulus materials, which were mostly meaningless syllables, offered no opportunity for listeners to recover from the effects of noise masking by exploiting contextual information; when such opportunities exist, native speech recognition proves more robust under noise masking than non-native recognition (Mayo et al., 1997).

In Cutler et al. (2004) study, both vowels and consonants were consistently identified less accurately by the non-native than by the native listeners, and for neither of these two subsets of the phonemic repertoire was there a differential effect of increasing noise for the non-native vs. native group. Dutch and American English have similar numbers of vowels and a similar distribution between monophthongs and diphthongs (Gussenhoven, 1999; Ladefoged, 1999); American English has rather more consonants than Dutch—about 25% again as many—but this mismatch did not seem to be associated with increased difficulty for the non-native listeners. For both vowels and consonants, of course, there are cases where American English contrasts are extremely difficult for Dutch listeners. But in general, having a native phoneme repertoire which differs from the repertoire of the presented non-native language seems to be the crucial factor in non-native phoneme recognition, and this factor can have equally deleterious effects for consonant and for vowel identification.

Repertoire mismatch can occur, however, not only across but also within languages, and it is not necessarily the case that vowels and consonants are equivalently affected. In some languages (e.g. Spanish), vowels remain largely constant while the repertoire of consonants can change across dialects (consider Castilian Spanish's [θ]). In other languages, vowel differences across dialects outnumber consonant differences. English is such a case. For instance, the three-way distinction between *look*, *luck* and *Luke* in most English varieties collapses to two in some varieties (e.g. *luck* vs. *look*/*Luke* in Scottish English, and *Luke* vs. *look*/*luck* in Yorkshire English). American English, though it in fact maintains this three-way distinction, has fewer vowels (16) than many other varieties of English (Wells, 1982). Some distinctions are further disappearing in some varieties of American English; Labov et al. (1991) showed that listeners from one American dialect background often fail to discriminate minimal pairs of words spoken by speakers from another area. The same result has been demonstrated for New Zealand English spoken by older speakers and perceived by younger compatriots (Warren et al., 2003). Such category mismatches and assimilations across dialects present listeners with much the same sort of categorisation problems as phonemic differences across different languages.

Note that vowels and consonants do not always pattern similarly in perceptual tasks. Listeners seem to be in general more cautious in vowel identification than in consonant identification. Thus response time to detect a vowel target in a phoneme-monitoring experiment is inversely correlated with vowel duration: the longer the vowel,

the faster listeners produce a detection response (Cutler et al., 1996). In word reconstruction, in which listeners are required to change non-words into the nearest available word, alterations of vowels are more readily and more rapidly produced than alterations of consonants (Cutler et al., 2000). These results have been explained as reflecting listener experience with vowel variability in context, and the consequent experience of often having had to alter initial hypotheses about vowel identity during listening.

Vowel variability is certainly well attested; perceptual confusion studies (e.g. Peterson and Barney, 1952; Hillenbrand et al., 1995) show that it can occur even in invariant context, and listeners often fail to agree on outlying tokens of vowel types. Even in a language with only five vowels, vowel types can exhibit considerable variability in natural speech (Keating and Huffman, 1984). Vowels excerpted from context are hard to identify (Koopmans van Beinum, 1980), especially transitions are variable (Schouten and Pols, 1979), and the more context—especially preceding context—is supplied, the better identification becomes (van Son and Pols, 1999).

This suggests that vowel mismatches across dialects may constitute a familiar perceptual problem, at least for listeners in languages such as English. The effects of repertoire mismatch between dialects may in such cases be fully analogous to the effects of repertoire mismatch across languages. The present study tests for such similarity, by directly comparing vowel identification under conditions of language vs. dialect mismatch. The materials are taken from the native vs. non-native listening study of Cutler et al. (2004) described above.

2. Method

2.1. Participants

Ten native listeners of Australian English, mostly students at the University of New South Wales, participated in the experiment. Sixteen native listeners of American English, students at the University of South Florida and at the City University of New York, and 16 Dutch-native listeners fluent in English, students at the University of Nijmegen, also participated. The American listeners had varying backgrounds and nearly all had lived in several different states. Some American listeners received course credit for participating; the rest, and all members of the other groups, received a small monetary compensation. The American and Dutch listeners participated as part of a larger eight-session experiment, and additionally received a monetary bonus upon completion.

2.2. Materials

Fifteen American English vowels (12 monophthongs: /i ɪ eɪ ɛ æ ɔ oʊ u ə˞ ɑ ʌ ʊ/ and three diphthongs: /aɪ ɔɪ aʊ/) were combined with the consonants /b/ and /v/ to form CV and VC sequences.

The resulting 60 syllables were included in a larger set of 645 syllables read from phonemic transcription by a phonetically trained female native speaker of American English (born and raised in the Mid-West). The recordings were made to Digital Audio Tape via a Sennheiser microphone in a quiet room, and stored to disc at 16 kHz. Each syllable was then centrally embedded in one second of multispeaker babble noise. The babble was constructed by adding together amplitude-equalized segments from individual speakers taken from a recording of conversation between six (three male, three female) English-speakers in a quiet room. The syllables were combined with the babble noise at three signal-to-noise ratios (SNRs: 0 dB, 8 dB, and 16 dB). These SNRs were chosen on the basis of a pretest to yield respectively difficult, intermediate, and easy phoneme perception for non-native listeners.

2.3. Procedure

The Australian listeners heard all 180 tokens (60 syllables at three SNRs) in a single session. The American and Dutch listeners heard the syllables as part of a larger experiment comprising eight testing sessions, including both consonant and vowel identification (for further details see Cutler et al., 2004). Every listener heard the items in a different pseudo-random order.

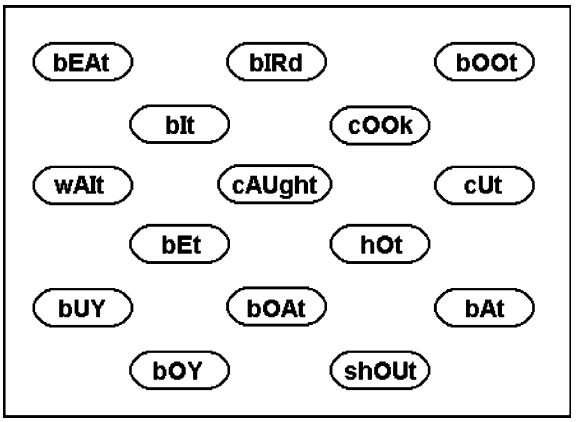

Fig. 1. Response display presented to participants in the vowel identification task.

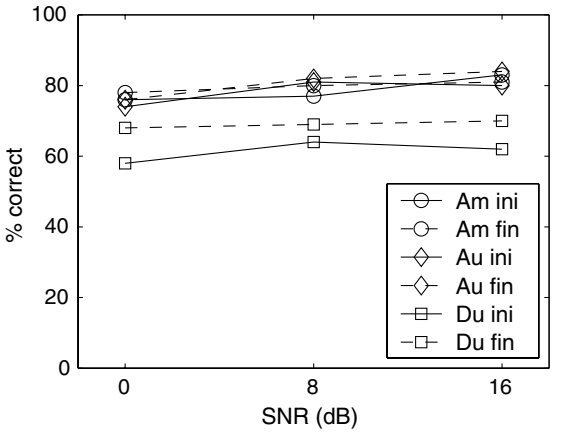

Fig. 2. Percentages of correctly recognised vowels, pooled across phonetic contexts and subjects, as a function of SNR, position ('ini' = initial, 'fin' = final), and language group ('Am' = American English, 'Au' = Australian English, 'Du' = Dutch).

Listeners signalled their response by clicking on a word exemplifying the appropriate vowel on a computer screen; they were familiarized with these words prior to the experiment. Fig. 1 shows the screen display. The presentation of items was self-paced. If the listener did not respond within 15 seconds after stimulus offset, the trial was recorded as a miss.

3. Results

No response ('miss') trials (<1% for each listener group) were discarded from the data set (i.e., were not counted as errors). Because the stimulus set presented to the Australian listeners included only bV, vV, Vb and Vv syllables, their performance was compared with the American and Dutch listeners' performance on the same subset, extracted from the full data set analyzed by Cutler et al. (2004).

Fig. 2 shows the overall percentages of correct responses for this subset, as a function of position and SNR, for each of the three listener groups. It can be seen that the identification performance of the American and Australian listener groups is highly similar, that neither group is affected by position of the vowel in the syllable, and that for both these groups SNR has small effects: averaged across position, the American listeners scored 77%, 79%, 82% correct at 0, 8, 16 dB SNR, and the Australians 75%, 82% and 82%. The Dutch listeners' performance, however, is significantly and consistently worse than the performance of both English-speaking groups; moreover, though their performance is also little affected by SNR, it is worse for vowels in initial (58%, 64%, 62% correct at 0, 8, 16 dB SNR) than in final position (68%, 69%, 70%).

An overall analysis of variance across subjects of all three groups confirmed that performance differed across listener groups ($F[2,39] = 8.13$, $p < .001$) and across SNR ($F[2,78] = 14.62$, $p < .001$). The listener group comparison did not interact with SNR, but did interact with position ($F[2,78] = 7.5$, $p < .001$). Full comparisons of the American and Dutch results are presented by Cutler et al. (2004), and the patterns observed in the present subset of those results exactly mimic the patterns in the whole set; accordingly we present here principally comparisons of the Australian listener group with the other two groups. These analyses showed that the Australian group did not differ in overall performance from the American group ($F < 1$), nor did this group comparison interact with any other factor. However, the Australian group performed significantly better than the Dutch group ($F[1,24] = 12.57$, $p < .005$), and an interaction of listener group with position was observed here ($F[1,24] = 6.14$, $p < .025$), reflecting

no significant effect of position for the Australian listeners, but significantly worse identification of initial than of final vowels for the Dutch listeners ($F[1, 15] = 32.74$, $p < .001$).

Fig. 3 presents the response patterns in terms of percentage of information transmitted for broad feature classes. Conversion to transmitted information from raw percent correct takes account of response biases, and gives a result of zero if subjects guess randomly (Miller and Nicely, 1955), irrespective of the number of response alternatives. Thus it enables comparisons between features with different numbers of values. As relevant features for vowels we used height (high /i ɪ ʊ u/ vs. mid /eɪ ɛ ʌ oʊ ɚ/ vs. low /æ ɑ ɔ/), backness (front /i ɪ eɪ ɛ æ/ vs. central /ɚ/ vs. back /ɑ ʌ ɪ oʊ ʊ u/) and tenseness (tense /i eɪ ɔ oʊ u ɚ/ vs. lax /ɪ ɛ æ ɑ ʌ ʊ/). The three diphthongs always change value on height and tenseness, and two also on backness, so we excluded them from Fig. 3 results; for the calculations of transmitted information we also excluded the few diphthong responses to monophthongs.

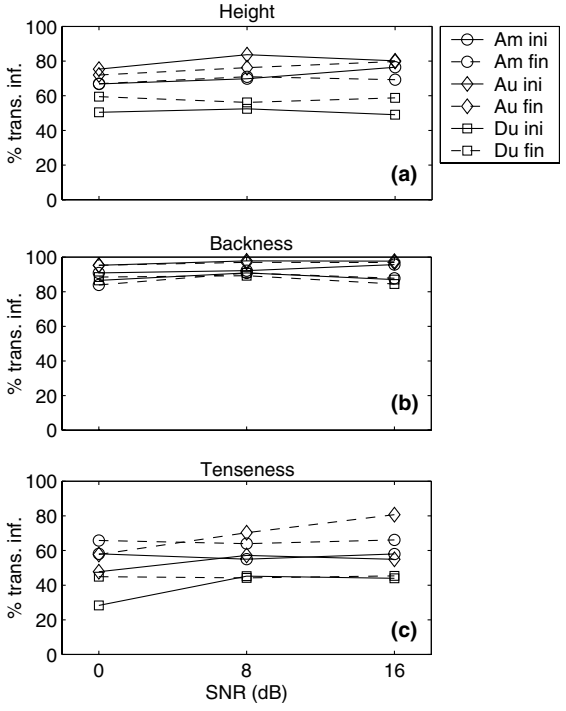

Fig. 3. Percentages of transmitted information, pooled across phonetic contexts and subjects, for three vocalic features as a function of SNR, position ('ini' = initial, 'fin' = final), and language group ('Am' = American English, 'Au' = Australian English, 'Du' = Dutch).

Statistical analyses of the comparisons in Fig. 3 showed that for all three listener groups, vowel backness information was transmitted most effectively and vowel tenseness information least effectively (all comparisons at least $p < .025$). There were no significant effects of position in the syllable. A comparison of the Australian vs. American groups for each feature revealed no significant main effects and only one significant interaction, of listener group with SNR for vowel tenseness ($p < .05$), due to the Australian group's performance on tenseness identification improving with increasing SNR (53% of the information was transmitted at 0 dB, 64% at 8 dB, 68% at 16 dB; $F[2, 18] = 4.8$, $p < .025$) while the American group's performance was unaffected by SNR (62%, 60%, 62%; $F < 1$). A comparison of the Australian vs. Dutch groups for each feature revealed that the Australian listeners were significantly better than the Dutch listeners at identifying each feature class (all comparisons at least $p < .025$); there were no interactions of listener group with SNR but one interaction with position, for vowel height ($p < .001$), due to an advantage for Dutch listeners for height judgements in final position (58% correct) over initial position (51%; $F[1, 15] = 16.3$, $p < .001$) but a (smaller) advantage of initial (80% correct) over final position (76%; $F[1, 9] = 6.98$, $p < .03$) for Australian listeners.

Tables 1–3 present confusion matrices of the responses of the American, Australian and Dutch listener groups respectively at 0, 8 and 16 dB SNR. It can clearly be seen that some vowels in this materials set were difficult for all listeners at all SNRs—for instance, the vowel /ɑ/ of *hot*, for which scores were never above 40%, even for American listeners at the clearest SNR. For all listener groups this vowel was frequently confused with the vowels /ʌ/ and /ɔ/ (of *cut* and *caught*). For American listeners, it was nevertheless always the case that the correct response received the highest score (though for the other two groups this was not always true). In an attempt to interpret the confusion patterns, we plotted the 12 monophthongal

Table 1
Confusion matrices for 15 American English vowels perceived in bV, vV, Vb and Vv context by American listeners, at 0, 8 and 16 dB SNR

		Response														
		i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	ɔ	oʊ	ʊ	u	aɪ	ɔɪ	aʊ	ɚ	miss
0																	
Stimulus	i	**92.2**	1.6		1.6				1.6							1.6	1.6
	ɪ		**84.4**		12.5	1.6							1.6				
	eɪ	1.6	1.6	**81.3**	3.1	6.3							1.6				4.7
	ɛ	1.6	14.1	3.1	**67.2**	4.7		1.6	1.6							3.1	3.1
	æ				7.8	**81.3**	1.6		3.1				1.6		3.1	1.6	
	ɑ			1.6		14.1	**31.3**	25.0	23.4	3.1							1.6
	ʌ			1.6		4.7	21.9	**62.5**	4.7	1.6	1.6						1.6
	ɔ					3.1	25.0	6.3	**59.4**	3.1	1.6						1.6
	oʊ						1.6			**87.5**		4.7		3.1			3.1
	ʊ						1.6	20.3	1.6	1.6	**65.6**	3.1		1.6	1.6		3.1
	u	1.6	1.6					1.6			15.6	**78.1**					1.6
	aɪ		9.4										**90.6**				
	ɔɪ								1.6					**95.3**	3.1		
	aʊ			1.6			1.6			9.4	3.1				**82.8**		1.6
	ɚ				1.6											**96.9**	1.6
8																	
Stimulus	i	**95.3**	1.6		3.1												
	ɪ	1.6	**85.9**		7.8							1.6					3.1
	eɪ	4.7	1.6	**89.1**		4.7											
	ɛ		12.5	3.1	**78.1**	6.3											
	æ			3.1	4.7	**82.8**			1.6					1.6	1.6		4.7
	ɑ					14.1	**40.6**	20.3	20.3	1.6							3.1
	ʌ			1.6	1.6	1.6	20.3	**57.8**	7.8	1.6	1.6	1.6	1.6		1.6		1.6
	ɔ					7.8	39.1	3.1	**46.9**			1.6					1.6
	oʊ									**89.1**		6.3		4.7			
	ʊ			1.6			4.7	17.2	4.7	1.6	**64.1**		1.6		3.1		1.6
	u									1.6	12.5	**79.7**		1.6	3.1		1.6
	aɪ		7.8	1.6									**90.6**				
	ɔɪ						1.6			1.6				**96.9**			
	aʊ								7.8	4.7	1.6	1.6		3.1	**81.3**		
	ɚ	1.6			1.6											**96.9**	
16																	
Stimulus	i	**96.9**			3.1												
	ɪ	3.1	**89.1**		4.7							1.6					1.6
	eɪ	3.1		**92.2**		4.7											
	ɛ	1.6	3.1	3.1	**81.3**	10.9											
	æ			1.6	3.1	**89.1**	1.6		1.6							1.6	1.6
	ɑ			3.1		14.1	**35.9**	21.9	20.3	1.6							3.1
	ʌ					6.3	14.1	**67.2**	7.8					1.6			3.1
	ɔ					1.6	29.7	4.7	**59.4**		1.6		1.6		1.6		
	oʊ									**95.3**	1.6	3.1					
	ʊ		1.6	1.6			4.7	14.1	4.7	1.6	**59.4**	4.7		1.6	1.6		4.7
	u										6.3	**89.1**			3.1		1.6
	aɪ		6.3		1.6								**92.2**				
	ɔɪ									1.6				**95.3**	3.1		
	aʊ						1.6		6.3	1.6					**89.1**		1.6
	ɚ	1.6														**98.4**	

Table 2
Confusion matrices for 15 American English vowels perceived in bV, vV, Vb and Vv context by Australian listeners, at 0, 8 and 16 dB SNR

		Response														
		i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	ɔ	oʊ	ʊ	u	aɪ	ɔɪ	aʊ	ɚ	miss
0																	
Stimulus	i	**97.5**	2.5														
	ɪ		**100.0**														
	eɪ	10.0		**90.0**													
	ɛ		10.0		**87.5**										2.5		
	æ				25.0	**67.5**									7.5		
	ɑ						**17.5**	52.5	22.5	2.5					5.0		
	ʌ						7.5	**75.0**	10.0						5.0		2.5
	ɔ						57.5		**42.5**								
	oʊ						2.5		2.5	**85.0**	10.0						
	ʊ						7.5		2.5		**77.5**	7.5			5.0		
	u	10.0									32.5	**50.0**	5.0	2.5			
	aɪ		5.0	7.5									**85.0**				2.5
	ɔɪ							2.5						**97.5**			
	aʊ								2.5	27.5					**70.0**		
	ɚ				2.5					10.0						**87.5**	
8																	
Stimulus	i	**97.5**	2.5														
	ɪ		**95.0**														5.0
	eɪ	2.5		**97.5**													
	ɛ		5.0		**90.0**					2.5							2.5
	æ			2.5	25.0	**67.5**									5.0		
	ɑ					2.5	**20.0**	57.5	12.5					2.5	2.5		2.5
	ʌ						7.5	**80.0**	2.5	2.5					7.5		
	ɔ						62.5		**32.5**		2.5			2.5			
	oʊ									**97.5**				2.5			
	ʊ							2.5	2.5		**85.0**	5.0					5.0
	u										5.0	**95.0**					
	aɪ			7.5									**90.0**			2.5	
	ɔɪ													**100.0**			
	aʊ								7.5	22.5					**70.0**		
	ɚ															**100.0**	
16																	
Stimulus	i	**95.0**	2.5		2.5												
	ɪ		**100.0**														
	eɪ	5.0		**95.0**													
	ɛ		5.0		**95.0**												
	æ	2.5			15.0	**80.0**									2.5		
	ɑ					5.0	**27.5**	57.5	10.0								
	ʌ					5.0	10.0	**80.0**		2.5					2.5		
	ɔ						52.5		**35.0**	10.0				2.5			
	oʊ						2.5			**97.5**							
	ʊ						2.5	2.5			**87.5**	5.0	2.5				
	u										15.0	**85.0**					
	aɪ			10.0									**85.0**			2.5	2.5
	ɔɪ													**100.0**			
	aʊ	2.5							2.5	22.5					**72.5**		
	ɚ				2.5											**97.5**	

Table 3
Confusion matrices for 15 American English vowels perceived in bV, vV, Vb and Vv context by Dutch listeners, at 0, 8 and 16 dB SNR

		Response														
		i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	ɔ	oʊ	ʊ	u	aɪ	ɔɪ	aʊ	ə	miss
0																	
Stimulus	i	**89.1**	7.8	1.6	1.6												
	ɪ		**96.9**									1.6	1.6				
	eɪ	25.0	4.7	**65.6**	1.6	1.6							1.6				
	ɛ		29.7	0.0	**51.6**	15.6									1.6	1.6	
	æ			1.6	35.9	**53.1**	1.6	1.6					1.6		4.7		
	ɑ					7.8	**31.3**	29.7	28.1		1.6		1.6				
	ʌ					1.6	34.4	**32.8**	18.8	7.8		1.6			3.1		
	ɔ					1.6	53.1	6.3	**34.4**	1.6	1.6			1.6			
	oʊ			1.6			7.8		3.1	**65.6**	6.3	14.1					1.6
	ʊ						1.6	4.7		9.4	**65.6**	15.6		1.6			1.6
	u	17.2	1.6				1.6			3.1	32.8	**43.8**					
	aɪ		1.6	18.8		1.6	1.6	1.6					**71.9**			3.1	
	ɔɪ			1.6			3.1		1.6	1.6				**92.2**			
	aʊ								1.6	29.7		1.6	1.6		**65.6**		
	ə							9.4	3.1						1.6	**85.9**	
8																	
Stimulus	i	**89.1**	6.3	3.1									1.6				
	ɪ		**93.8**	1.6			1.6				1.6					1.6	
	eɪ	17.2	3.1	**73.4**	3.1	3.1											
	ɛ		21.9		**60.9**	17.2											
	æ		1.6	4.7	35.9	**56.3**									1.6		
	ɑ				1.6	18.8	**25.0**	34.4	17.2	1.6					1.6		
	ʌ					9.4	29.7	**32.8**	15.6	3.1		1.6	3.1	1.6	3.1		
	ɔ					1.6	51.6	4.7	**40.6**	1.6							
	oʊ						6.3			**76.6**	6.3	9.4		1.6			
	ʊ						10.9	1.6	4.7	4.7	**59.4**	14.1		1.6	3.1		
	u						1.6			1.6	53.1	**42.2**		0.0	1.6		
	aɪ		3.1	17.2	1.6								**75.0**	1.6		1.6	
	ɔɪ						1.6						1.6	**96.9**			
	aʊ								3.1	15.6			3.1		**78.1**		
	ə	1.6			1.6			3.1								**92.2**	1.6
16																	
Stimulus	i	**87.5**	6.3	3.1	3.1												
	ɪ		**93.8**		1.6							1.6				3.1	
	eɪ	12.5		**78.1**	6.3	1.6			1.6								
	ɛ		15.6		**65.6**	18.8											
	æ			6.3	34.4	**51.6**	1.6	1.6		1.6			0.0		1.6	1.6	
	ɑ			1.6		25.0	**29.7**	20.3	17.2				3.1		1.6	1.6	
	ʌ		1.6			9.4	37.5	**46.9**	3.1						1.6		
	ɔ					4.7	56.3	3.1	**25.0**	1.6	3.1	1.6		4.7			
	oʊ						4.7			**84.4**	3.1	7.8					
	ʊ						39.1	3.1		4.7	**42.2**	9.4			1.6		
	u						1.6	1.6		1.6	40.6	**53.1**			1.6		
	aɪ		3.1	23.4									**73.4**				
	ɔɪ						1.6		1.6					**96.9**			
	aʊ				1.6				10.9	17.2			1.6		**67.2**		1.6
	ə	1.6						6.3								**92.2**	

vowels of the study in F1/F2 space (the three diphthongs moved consistently between their nearest component monophthongs). Fig. 4 shows the resulting plot.

Many of the confusions which occurred across all groups, and in particular the one confusion which was consistently made by American listeners but not by the other groups (/ʊ/ misreported as /ʌ/) represent errors principally along the F1 dimension. We do not have an explanation for this effect; it may be that our speaker's F1 was rather weak, or that the babble noise we used masked F1 variation to a rather high extent (note that the confusions increase with decreasing SNR). However, although these common errors were made far more by the non-American groups (especially the Dutch group) than by the American listeners, it is the errors made by the two non-American groups but *not at all* by the American listeners which are potentially most informative. There were two confusions which the Australian listeners made to a significant degree (15% or more erroneous identification) but the American listeners did not: /æ/ was misidentified as /ɛ/, and the diphthong /aʊ̃/ was misidentified as the monophthong plus glide /oʊ̃/; both of these confusions were also made by the Dutch listeners. As Fig. 4 shows, /æ/ and /ɛ/ were spectrally very similar; the American listeners presumably avoided confusing them because there was a marked difference in their durations (the former 37% longer than the latter). Clearly, the non-American listeners did not exploit this durational distinction. Because /aʊ̃/ and /oʊ̃/ were similar in duration, and ended at a similar point in F1/F2 space, the non-American listeners' confusions are likely to be due to confusion of the earlier portions. Besides these two patterns shared with the Australians, the Dutch group further consistently misidentified /ɛ/ as /æ/ and /aɪ/ as /eɪ/. The Dutch phoneme inventory contains only one vowel where English distinguishes two for the /æ/-/ɛ/contrast, and Dutch has a diphthong similar to English /eɪ/ but none similar to /aɪ/.

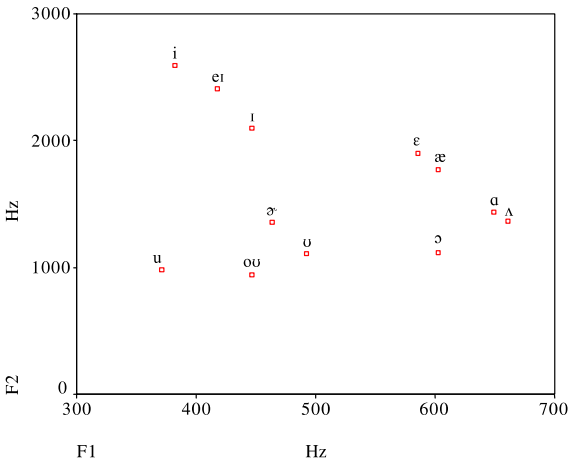

Fig. 4. Mean F1 and F2 values for 12 monophthongal American English vowels, averaged over the tokens used in the present study. N.B. although the vowels /eɪ/ and /oʊ̃/ (of *wait* and *boat*) include a terminal glide, they are generally considered monophthongal (Ladefoged, 1999), and have been treated as such in our study.

4. General discussion

The effect on vowel identification of having a different native language is disproportionately greater than the effect of having a different native dialect. The identification performance of Dutch listeners in our study fell clearly short of the performance of both the other groups. The Australian listeners' overall performance, however, was not significantly worse than that of the American native listeners.

Perceptual confusions which were made by all listener groups may reflect idiosyncrasies of our speaker's voice, or the masking characteristics of the multispeaker babble noise we used for the speech of the recorded speaker. Perceptual confusions made by only one group, however, are unlikely to be attributable to characteristics of the stimuli. There may be the possibility of some dialect mismatch even for our native listener group (who came from varying dialectal backgrounds, mostly different from the Mid-western background of our speaker), but this would be trivial compared to the mismatch for the other two groups, native speakers of Australian English and of Dutch. For these two groups, we can assume that any group-specific perceptual confusions arise from

the mismatch between the native system and American English.

For the Dutch listeners, there were many confusions. The vowels /æ/ and /ɛ/ were confused; although the former was misreported as the latter more often than vice versa, significant confusions occurred in both directions. Dutch contains only one vowel in this area of the vowel space, written in IPA as /ɛ/ but situated between the two English vowels. The diphthong /aɪ/ was often misidentified as /eɪ/; again, Dutch does not make this distinction, but in this case there was no significant pattern of confusion in the opposite direction, presumably reflecting the fact that Dutch has a diphthong /ɛi/ which is closer to English /eɪ/ than to /aɪ/. These effects suggest the well-known phenomenon of capture of non-native speech input by native category structure (Strange, 1995).

For the Australian listeners, we found only two significant confusions which the native listeners did not make. First, the Australian listeners made less use of duration to distinguish longer /æ/ from shorter /ɛ/. Note that some Australian vowel contrasts may be signalled in large part by durational distinctions (Fletcher et al., 1994; Harrington and Cassidy, 1994), so that Australian listeners should in principle be able to exploit durational cues to vowel identity; in this particular case, however, Australian English does not make as marked a durational distinction between the two vowels as American English does (Wells, 1982), and this presumably accounts for the Australian listeners' lesser reliance on the duration cue. Second, we found a difference between the American and the Australian groups in the analysis of information transmitted about vowel features. Specifically, increasing noise masked tenseness information for the Australian listeners to a greater degree than it did for the native listeners. The reason for this can be seen in Tables 1 and 2. Both listener groups make confusions among the back vowels, but misidentifications of the tense vowel /ɔ/ as the lax vowel /ɑ/ are, for the Australian listeners, actually more common than correct identifications of /ɔ/ at all SNRs, and other errors of tenseness (especially /u/ or /oʊ/ misidentified as /ʊ/, or /ʌ/ misidentified as /ɔ/) were more common for Australian listeners than for native listeners, especially at 0 dB and 8 dB. Although the American listeners also made all these kinds of errors, the Australian listeners made them far more frequently.

Australian English vowels are in general more fronted and tenser than the same vowels in other dialects of English (Wells, 1982), and it is thus not surprising that Table 2 reveals distinct asymmetry in the tenseness errors made by the Australian listeners: erroneous identifications of tense vowels as lax outnumber erroneous identifications of lax vowels as tense by more than two to one. (For the native listeners, tenseness errors in each direction were fewer, and more nearly equal in number.) Apparently, the American tense vowels were not tense enough for the Australian listeners. These results thus show a clear effect of native phonemic distribution on non-native listening, in this case listening to non-native dialect.

A dialect mismatch can thus cause perceptual confusions in the same way as a language mismatch can. But the overall levels of identification performance we observed attest that interference is far greater from a language than from a dialect mismatch. Two dialects often share a repertoire of phonemic categories, although there may be considerable difference in the boundaries of the categories. Across languages, however, the repertoire of categories itself usually varies. This latter mismatch, we assume, leads to far more serious perceptual problems. Recent research has suggested that the boundaries of native-language categories can be easily and rapidly adjusted to deal with variability across speakers (Norris et al., 2003), and, presumably, across dialects. Where a dialect actually collapses two categories which are distinct in the input (as Labov et al., 1991; Warren et al., 2003, have shown can occur), this will lead to perceptual confusions, but we would predict that recovery from such confusions would be easier for non-native-dialect listeners in natural listening situations, and further, that learning of the non-native dialect's different mapping should be easier than learning a non-native language's different repertoire. In short, the perceptual effects of having a mismatching dialect are very small compared to the effects of having a mismatching language.

Acknowledgments

This study was supported by a research stipend from the Max Planck Society to NC and by a SPINOZA award from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek to AC. We thank Andrea Weber for testing the American listeners, Winifred Strange for making the American listener population available, Sally Andrews for making the Australian listener population available, Natasha Warner for recording the speech materials, Marloes Weijers for technical assistance, Jonathan Harrington for advice on the Australian vowel system, and James McQueen and two anonymous reviewers for helpful comments on the text.

References

Cutler, A., Sebastián-Gallés, N., Soler, O., van Ooijen, B., 2000. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. Memory Cognition 28, 746–755.

Cutler, A., van Ooijen, B., Norris, D., Sánchez-Casas, R., 1996. Speeded detection of vowels: A cross-linguistic study. Percept. Psychophys. 58, 807–822.

Cutler, A., Weber, A., Smits, R., Cooper, N., 2004. Patterns of English phoneme confusions by native and non-native listeners. J. Acoust. Soc. Amer. 116, 3668–3678.

Fletcher, J., Harrington, J., Hajek, J., 1994. Phonemic vowel length and prosody in Australian English, In: Proc. 5th Australian Internat. Conf. on Speech Science and Technology, Vol. II, pp. 656–661.

Gussenhoven, C., 1999. Dutch. In: Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, UK, pp. 74–77.

Harrington, J., Cassidy, S., 1994. Dynamic and target theories of vowel classification: evidence from monophthongs and diphthongs in Australian English. Language Speech 37, 357–373.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Amer. 97, 3099–3111.

Keating, P.A., Huffman, M.K., 1984. Vowel variation in Japanese. Phonetica 41, 191–207.

Koopmans van Beinum, F.J., 1980. Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions. Unpublished Ph.D. thesis. University of Amsterdam.

Labov, W., Karen, M., Miller, C., 1991. Near-mergers and the suspension of phonemic contrast. Language Variation Change 3, 33–74.

Ladefoged, P., 1999. American English. In: Handbook of the International Phonetic Association. Cambridge University Press, Cambridge, UK, pp. 41–44.

Mayo, L.F.H., Florentine, M., Buus, S., 1997. Age of second-language acquisition and perception of speech in noise. J. Speech Hearing Res. 40, 686–693.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338–352.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychology 47, 204–238.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175–184.

Schouten, M.E.H., Pols, L.C.W., 1979. Vowel segments in consonantal contexts: A spectral study of coarticulation—Part I. J. Phonetics 7, 1–23.

Strange, W., 1995. Speech Perception and Linguistic Experience: Issues in Cross-language Speech Research. York Press, Timonium, MD.

van Son, R.J.J.H., Pols, L.C.W., 1999. Perisegmental speech improves consonant and vowel identification. Speech Comm. 29, 1–22.

Warren, P., Rae, M., Hay, J., 2003. Word recognition and sound merger: the case of the front-centering diphthongs in NZ English, In: Proc. 15th Internat. Congress of Phonetic Sciences, Barcelona, Spain.

Wells, J.C., 1982. Accents of English, 3 Vols. Cambridge University Press, Cambridge.