Consonant And Vowel Confusion Patterns By American English Listeners

Andrea Weber* and **Roel Smits**†

* Department of Psycholinguistics, University of the Saarland, Saarbrücken, Germany
aweber@coli.uni-sb.de

† Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
roel.smits@mpi.nl

ABSTRACT

This study investigated the perception of American English phonemes by native listeners. Listeners identified either the consonant or the vowel in all possible English CV and VC syllables. The syllables were embedded in multispeaker babble at three signal-to-noise ratios (0 dB, 8 dB, and 16 dB). Effects of syllable position, signal-to-noise ratio, and articulatory features on vowel and consonant identification are discussed. The results constitute the largest source of data that is currently available on phoneme confusion patterns of American English phonemes by native listeners.

1. INTRODUCTION

Under good listening conditions, humans can in general recognize the different sounds of their native language well. Nevertheless, depending on the details of the language's phoneme inventory, some phonemes are harder to recognize than others. For example, when two phonemes are acoustically highly similar, as is the case for /f/ and /θ/ in English (e.g., [1]), they may be confused even under optimal conditions. Even though previous research has given us insight into confusion patterns for restricted sets of phonemes (e.g., [2, 3, 4]), there is to date no dataset for confusion patterns, among all phonemes of a language, except [5], a gating study in Dutch. The present study compiled information about phoneme confusions for the entire phoneme inventory of American English, in all potential CV and VC contexts. The aim of the present study was to determine the accuracy with which native listeners can perceive vowels and consonants both in syllable-initial and syllable-final position in American English. Stimuli were embedded in multispeaker babble noise at three different signal-to-noise ratios (SNRs).

2. METHOD

2.1. Participants

Sixteen native listeners of American English, mostly students at the University of South Florida, participated in the experiment for either monetary compensation or course credit.

2.2. Material

22 consonants and 15 vowels were combined to form all possible English CV and VC syllables. Except for /h/, /j/, /w/, /ŋ/, and /ʒ/, all consonants and vowels occurred in both syllable-initial and syllable-final position. The syllables were transcribed phonemically. A phonetically trained female native speaker of American English was seated in a quiet room and read the transcriptions into a high-quality microphone. The sampling rate during digitization was 16 kHz. Each syllable was centrally embedded in 1 second of multispeaker babble at three different SNRs (0 dB, 8 dB, and 16 dB). These SNRs were chosen to yield easy, intermediate, and difficult phoneme perception for nonnative Dutch listeners, who also participated in the experiment, but whose data will not be discussed here.

2.3. Procedure

Over eight sessions, each listener heard all CV and VC syllables in three SNRs twice, once identifying the consonant and once identifying the vowel. 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. Each listener was presented with the items in a different pseudo-random order. In each session, listeners had to identify blocks of initial or final consonants and blocks of vowels. They responded by clicking the word that contained the appropriate sound on a computer screen. Different words were used for vowels, initial consonants, and final consonants. Participants were familiarized with the words prior to the experiment. In total, 3870 observations were collected for each participant.

3. RESULTS

Listeners' responses were summarized in confusion matrices, showing how often phoneme Y was perceived given phoneme X. Results for consonants and vowels at 0 dB SNR are presented in Tables I and II, respectively, pooling over syllable positions. The data are presented in percentages of correct responses.

ISBN 1-876346-48-5 © 2003 UAB

		response																								
		lip	hot	sick	off	path	pass	fish	such	hi	grab	odd	egg	love	smooth	buzz	beige	edge	yell	am	on	ring	ill	far	win	
		p	t	k	f	θ	s	ʃ	tʃ	h	b	d	g	v	ð	z	ʒ	dʒ	j	m	n	ŋ	l	r	w	miss
stimulus	p	**32.7**	9.6	9.4	8.1	4.6				19.6	3.8	.8	1.3	1.3	1.7			.4	.6	.8	.2		.2	.8	.4	3.8
	t	7.7	**48.3**	7.7	3.1	5.4	.6	.4	1.3	14.0	1.5	.8	1.7	.4	2.9	.8		.2	.4	.2	.4		.2		.4	1.5
	k	11.5	13.5	**44.6**	1.3	3.5		.2	1.5	14.0	1.3	.6	1.3	.6	1.5				.2	.2	1.0				.8	2.5
	f	16.5	6.0	5.2	**32.1**	10.2	.2	.4		7.1	4.8	.6	.4	2.5	5.6		.4	.2	.8	.2			.2	.8	1.9	3.8
	θ	10.8	11.7	4.0	22.1	**18.8**	.4	.2	.6	5.2	4.0	1.3	.8	3.1	11.0	.4	.2		.2	.4	.4	.4	.2	.4	.8	2.5
	s	.6	2.5	.6	11.0	9.4	**58.5**	2.5	.4	1.0	1.3		.2		5.6	4.8	.2										1.3
	ʃ	.2		.2		.2	1.0	**78.8**	16.9				.2	.2	.4		1.3	.4								.2
	tʃ	.2	4.4	.6	.8	.4		.6	**86.7**	.2			.2		.8		.8	3.5	.2		.2					.2
	h	14.6	5.0	4.6	9.6	4.6	.4		.4	**36.7**	7.1	.4	1.7	2.9	2.1			.4		1.7	.4		.4	.4	1.7	5.0
	b	1.7	.6	2.7	4.8	4.2	.2		.2	7.5	**27.3**	5.8	5.2	10.2	6.5	.4	1.3	1.0	1.9	6.5	1.0		2.3	1.5	1.9	5.4
	d		3.1	.2	2.5	5.2	1.0		.4	2.3	5.8	**28.8**	3.8	3.8	10.2	1.0	2.7	3.3	3.5	1.9	12.7	1.5	3.3	.2	.2	2.5
	g	1.0	2.1	1.3	2.9	4.0	.4		.4	5.2	4.6	5.8	**32.5**	9.2	3.8	.6	1.0	1.3	9.6	1.3	5.0	1.3	1.7	1.0	.6	3.5
	v	1.7	1.3	.8	7.5	3.3	.4		.4	4.4	10.6	2.5	5.6	**32.5**	10.2	.2	1.9	1.3	1.0	3.5	1.5	.4	.8	1.7	2.5	4.0
	ð		1.7		1.5	9.4	1.7	.2	.4	.8	6.3	13.3	4.2	11.7	**23.5**	5.0	2.7	4.8	.6	.6	2.9	.4	5.2	.2	.4	2.5
	z	.2	.8			5.8	5.0	.2	.2	.4	1.5	5.8	1.9	10.4	15.4	**34.2**	2.7	3.3	.2	.6	2.9		.8	1.3	3.8	2.5
	ʒ				.4		.8	2.5	2.1			2.5	1.3	4.2	4.6	3.8	**51.7**	23.3		.4	1.7		.4	.4		
	dʒ	.4	.6	.2	.2	1.5	.2	.2	4.0	.6	1.0	5.0	2.7	.6	4.4	.2	9.0	**66.7**	.4	.4	.2		.6	.4		.4
	j		.8			.4				2.9	3.3	5.4	3.3	1.3	1.3	1.7		2.9	**65.8**	2.5	2.5		1.7		2.1	2.1
	m	.2	.2	.2	1.5	.4				1.7	2.7	1.0	1.7	6.7	.4	.2		.2	.4	**60.0**	9.0	7.1	3.1	1.0	.8	1.5
	n					.6				.2	.4	2.9	.6	1.9	1.0	.8	.6	.4	.2	12.5	**68.8**	5.2	2.1	.6	.2	.8
	ŋ		.4	.4	.8				.4		.4	1.3	9.2	5.8	1.3		.4			15.4	25.4	**35.0**	.8	1.7		1.3
	l	.4	.4	.4	4.0	1.5			.2	.4	2.5	1.3	1.5	5.0	3.8	.2	.2	.2	1.0	6.9	2.7		**62.5**	1.5	.4	3.1
	r	.4	.4	.8	1.3	.4				3.5	2.9	.6	1.3	4.4	1.0			.6		1.9	.4		.4	**76.5**	1.5	1.7
	w	.8	.4							1.7	4.2		.8	2.9	.4				4.2	5.8			2.9	.4	**73.3**	2.1

Table I: Confusion matrix for consonants at 0 dB SNR. Percentages of correct responses were pooled over participants, vowel contexts and, syllable positions.

		response															
		beat	bit	wait	bet	bat	hot	cut	caught	boat	cook	boot	buy	boy	shout	bird	
		i	ɪ	eɪ	ɛ	æ	ɑ	ʌ	ɔ	oʊ	ʊ	u	aɪ	ɔɪ	aʊ	ɚ	miss
stimulus	i	**86.3**	4.2	.1	3.2				.3		.1	1.6	.6		.1	1.9	1.5
	ɪ	1.2	**83.1**	.6	9.3	.6		1.6	.1	.1	.1	.7	.4		.1	.9	1.0
	eɪ	3.1	3.6	**83.1**	3.2	4.2				.1	.1		.6		.1	.6	1.2
	ɛ	.6	5.2	2.3	**78.8**	5.7		1.7	1.2		.1		.3		.1	1.6	2.3
	æ		.9	3.8	8.1	**80.5**	.3		1.2			.1	.3	.1	2.3	1.2	1.2
	ɑ		.3	1.2	.6	9.0	**37.8**	18.6	26.9	.7	.1		.4		1.0	.9	2.5
	ʌ			.6	1.2	3.6	11.9	**65.1**	9.7	1.0	1.0	.1	.6	.1	2.2	1.5	1.3
	ɔ		.1	.6		1.6	29.9	4.1	**56.5**	2.3	.9		.3	1.0	.9		1.7
	oʊ	.1	.3		.1		3.2	.6	.7	**80.4**	4.5	3.6		2.3	2.2	.1	1.7
	ʊ	.1			.3		2.0	17.9	1.3	.7	**66.0**	4.7	.3	1.6	1.0	.4	3.6
	u	3.6	.7	.1	.4		.4	1.9	1.0	1.0	12.9	**72.4**	.1	.7	2.2	.6	1.7
	aɪ		7.1	1.5	.1		.1	.1	.1	.1			**89.4**	.3	.1	.3	.6
	ɔɪ	.1		.3	.1		.3	.3	.7	1.3	.7	.3	.1	**92.4**	2.6		.6
	aʊ	.3		.1	1.5	.6	2.6		4.4	5.4	.4	.1		1.3	**82.0**	.4	.9
	ɚ	.4	.3		1.3			.9	.1							**96.7**	.3

Table II: Confusion matrix for vowels at 0 dB SNR. Percentages of correct responses were pooled over participants, consonant contexts, and syllable positions.

ISBN 1-876346-48-5 © 2003 UAB

Participants' response types are listed in columns, stimulus types in rows. Null responses are listed in the "miss" column. At the top of each matrix, the words used as response buttons in the experiment are listed. In Table I, except for the initial-only consonants (/h/, /j/, /w/), words for consonant identification in VCs are given.

Visual inspection of the matrix reveals that, for consonants, the percentages of correct responses were lowest for stops and fricatives, higher for nasals, and highest for glides, liquids, and affricates. For stops, most errors were place errors. Voice errors were rare. Some manner errors were made for voiced stops. For voiced and voiceless fricatives, manner, place, and voice errors were found, especially for the dental and labiodental fricatives. /ʃ/ was recognized well. For nasals, place errors were most common. Glides and liquids were mainly confused with other manner categories. These confusion patterns match those reported in previous research (e.g., [2]).

For vowels, visual inspection of the matrix shows that the percentages of correct responses were lowest for back vowels, higher for front vowels, still higher for diphthongs, and highest for the central, tense, mid vowel /ɚ/. Whereas height was problematic for the identification of front vowels, tenseness was problematic for back vowels.

Figure 1 shows percentages of correct responses for consonants and vowels in initial versus final position, pooled over three SNRs. The figure suggests that vowels (78% correct) were better recognized than consonants (69% correct), and both phoneme types were on average better recognized in final (76% correct) than in initial position (71% correct). In order to statistically test these patterns we conducted an ANOVA on percentages of correct responses with phoneme class (consonant or vowel) and syllable position (initial or final) as fixed factors and participant as random factor. Both fixed factors proved to be significant (F[1,15] = 33.8, p<.0005, F[1,15] = 39.7, p<.0005, respectively), whereas their interaction was not (F < 1).

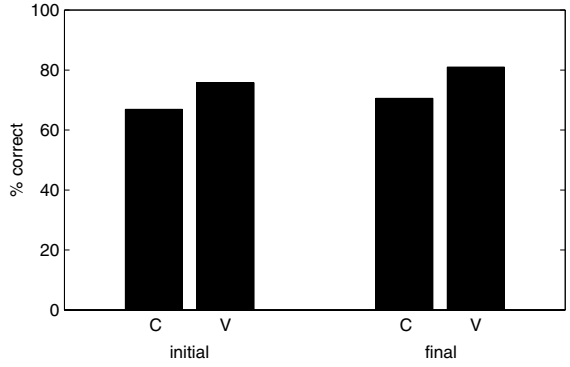

Figure 1: Mean percentages of correct responses for consonants and vowels in syllable-initial and syllable-final position.

The advantage of final over initial syllable position contrasts with the advantage of initial position reported by most - though not all - previous studies contrasting the two positions (e.g., [3]). An important difference between our method and the one used in [3] is that we presented our syllables in isolation, whereas in [3] the stimuli were recorded and presented in carrier sentences with words preceding and following the stimulus nonword. The uncertainty of the moment of stimulus onset in our experiment could have caused additional errors. Detection of stimulus onset may have been made still more difficult by our use of babble noise, which continuously varies in amplitude, as opposed to the stationary pink noise used in [3]. Finally, whereas our stimulus material employed all English phonemes, [3] tested only a limited set of consonants and vowels. Preliminary analyses of our data indicate that our advantage of final position vanishes when only the phonemes employed by [3] are considered.

For our next analysis, we split participants' responses into three SNRs, pooling over syllable position. As can be seen in Figure 2, consonant identification improved considerably across the SNRs (F[2,30] = 2520.1, p<.0005). Whereas at 0 dB only 49% of the consonants were identified correctly, at 8 dB 73%, and at 16 dB 85% were identified correctly. The differences between all three levels were significant in a post-hoc test. As suggested by Figure 2, identification of vowels also improved with increasing SNR (F[2,30] = 22.1, p<.001). A post-hoc test showed a significant improvement between 0 dB (77%) and 8 dB (79%), but no difference was found between 8 dB (79%) and 16 dB (79%). Furthermore, the significance of the interaction between phoneme type and SNR (F[2,30] = 1406.2, p<.0005) confirmed that consonant recognition improved more than vowel recognition. The relative flatness of the vowel curve in Figure 2 is probably due to a ceiling effect: At 0 dB, vowels were already identified correctly nearly 80% of the time. Note that 0 dB was selected as a difficult SNR for nonnative Dutch listeners listening to English. Native listeners are known to be more robust against noise interference in speech recognition, and indeed native listeners of American English did not seem to have difficulties identifying vowels at 0 dB.

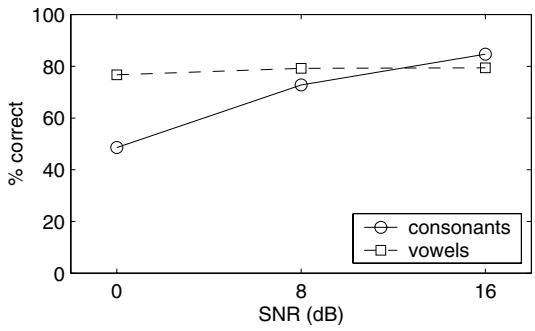

Figure 2: Mean percentages of correct responses for consonants and vowels at 0 dB, 8 dB, and 16 dB.

ISBN 1-876346-48-5 © 2003 UAB

Finally, phonemes were coded according to phonological features. Consonants were coded for manner (affricate, fricative, glide, liquid, nasal, stop), place (labial, dental, alveolar, palatal, velar, glottal), and voice (voiced, voiceless), vowels for height (low, mid, high), backness (back, central, front), and tenseness (lax, tense). Diphthongs were coded as a separate class.

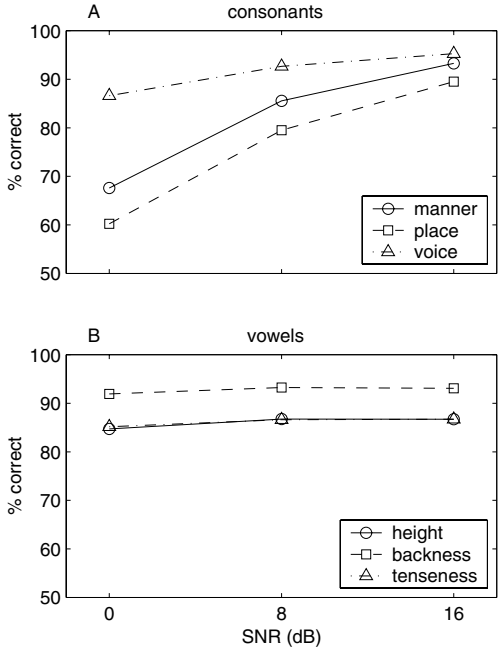

Figure 3: Mean percentages of correct responses for features of consonants (A) and vowels (B) at 0 dB, 8 dB, and 16 dB.

For consonants, the identification of manner as well as place and voice was significantly better at 8 dB than at 0 dB, and again significantly better at 16 dB than at 8 dB (Figure 3A). An ANOVA with feature (manner, place, or voice) and SNR (0 dB, 8 dB, or 16 dB) as fixed factors showed not only two significant main effects but also a significant interaction (F[4,60] = 383.9, p<.0005). Post-hoc tests revealed that recognition of manner, place, and voice benefited differently from each increase in SNR. A comparison of manner with place identification showed that for each SNR, manner was significantly easier to identify than place. The small number of levels for voice (2 levels) made a straight comparison with manner and place identification (6 levels each) invalid.

For vowels, the identification of height, backness, and tenseness was significantly better at 8 dB than at 0 dB (Figure 3B). For all three features, identification at 16 dB, however, did not differ from 8 dB. At 0 dB, all three vowel features were identified correctly well above 80%, thus there was again little room for improvement. In contrast to the consonant findings, an ANOVA with feature and SNR as fixed factors showed no significant interaction (F[4, 60] = 1.1, p>.3). This suggests that the observed effects of SNR were similar in size and direction for the three vowel features. Backness was at every level of SNR easier to identify than height. The difference in number of levels allowed again no direct comparison with tenseness.

4. CONCLUSIONS

This study was designed to assess the relative identifiability of American English phonemes. The data report phoneme confusion patterns of native listeners for the complete phoneme inventory of American English. Native listeners identified vowels and consonants embedded in multispeaker babble at three SNRs. As expected, vowels were better identified than consonants. Surprisingly, however, both vowels and consonants were better identified in syllable-final than in syllable-initial position. This contrasts with previous findings [3, 4]. The better identification of syllable-final phonemes in the present study might have been caused by presenting stimuli without preceding carrier words, by the different kind of noise used, or by testing the complete phoneme inventory of a language rather than a subset. Whereas correct consonant identification increased considerably when noise level decreased, vowel identification was less affected by a decrease in noise, presumably because it was already close to ceiling. In sum, the results represent a reliable database for phoneme confusion patterns of American English phonemes by native listeners. Future work will compare the present data to phoneme identification of the same material by nonnative Dutch listeners.

5. REFERENCES

[1] Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America*, **85**, 1718-1725.

[2] Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

[3] Redford, M. A., & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America*, **106**, 1555-1565.

[4] Benkí, J. R. (to appear). Analysis of English nonsense syllable recognition in noise. *Phonetica*.

[5] Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America*, **113**, 563-574.

ACKNOWLEDGMENTS

We would like to acknowledge the substantial contribution that Nicole Cooper has made to the project. We thank Winifred Strange and the Department of Communication Sciences & Disorders at the University of South Florida for their support with collecting data, and Anne Cutler for discussion and stimulation.