‘Many to One’ in the Articulation to Acoustics Map

Keith Johnson
Sarah Bakst

Abstract. The “many to one” problem arises when trying to map inversely from acoustic patterns to vocal tract configurations. In one famous demonstration, Atal et al. (1978) searched through the acoustic outputs of a synthetic vocal tract for vowels that matched each other exactly on the frequencies of the first three resonances (F1-3) and found that for each vowel tested [i], [a] and [u] there were several vocal tract configurations that gave the same formant frequencies. This result has been used to show that one (whether speech technologist or listener) cannot inversely map from acoustics to derive a unique possible vocal tract shape. We synthesized vowels using the formant frequencies reported by Atal et al. and show that listeners can detect differences between them even though the vowels are identical in the first three formants. Our conclusion is that listeners may not be as troubled by a many-to-one problem as has been assumed before.

Introduction

A framework for thinking about the inversion problem in speech acoustics was stated by Atal et al. (1978) in this way:

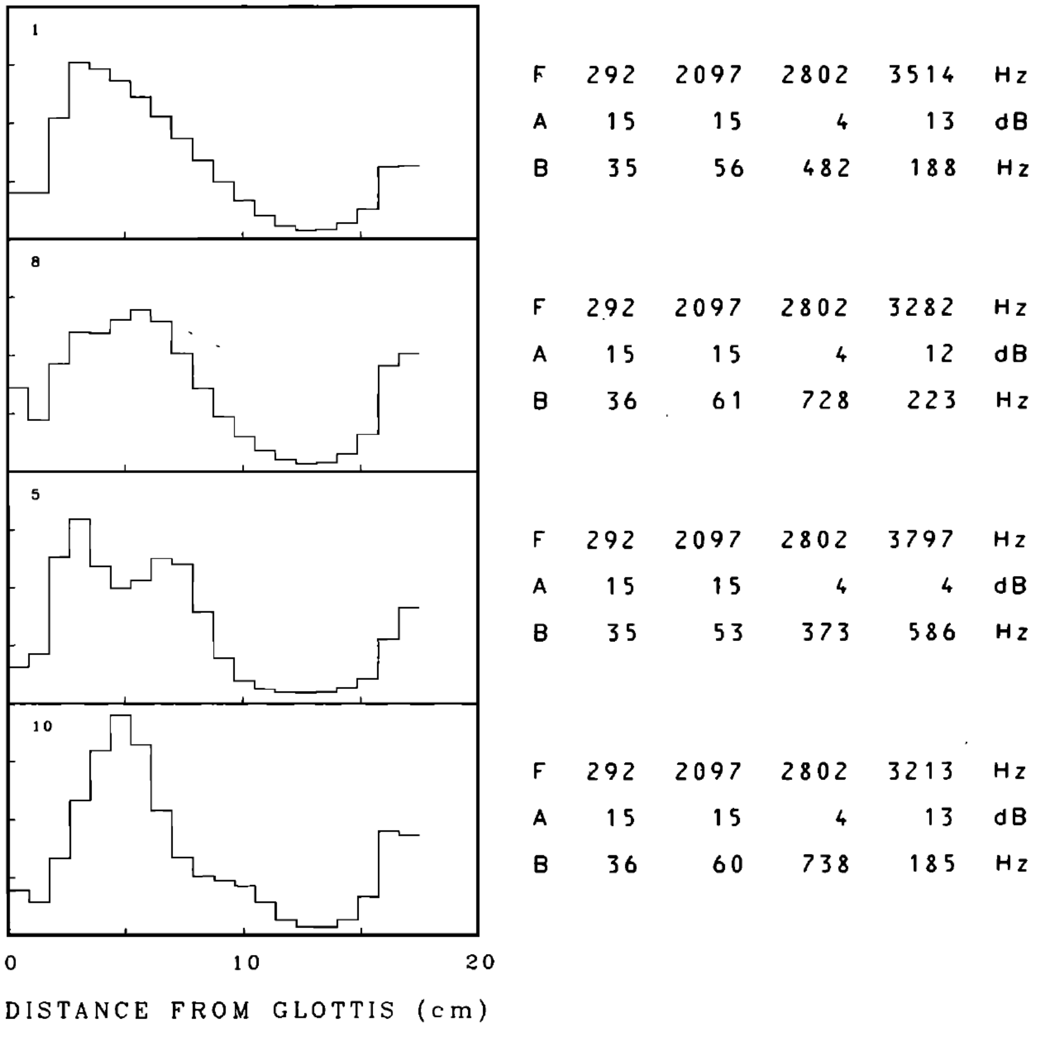
> “The relationship between the shape of the vocal tract and its acoustic output can be represented as a multidimensional function of a multidimensional argument
>
> $$y = f(x)$$
>
> where x is a vector describing the configuration of the articulators, y is a vector describing the resulting acoustics, and f is the function relating these vectors. We will assume that x has m dimensions and y has n dimensions.”

The inversion problem is that the mapping between the acoustic output of the vocal tract (*y*) and the articulatory configurations that gives rise to that acoustic pattern (*x*) is “multivalued for many situations of importance” (p. 1536). That is, there is a many-to-one mapping between articulation and acoustics .

The many-to-one problem was described by Schroeter & Sondhi (1994) as the non-uniqueness of the inverse mapping. They showed mathematically that there is no way to determine how many poles and zeros interact in producing a vocal tract transfer function, and that the inverse problem is compounded because acoustic characteristics of the voice source cannot be separated from filter characteristics. Research seeking to derive articulatory parameters from the speech acoustic signal – as a way to constrain speech recognition systems or as a way to train articulatory speech synthesizers – has generally assumed that the many-to-one problem is a major hurdle. And given the sparsity of representations (acoustic and/or articulatory) that is available in practical systems this has proven true (see for example, Papcun, et al., 1992; Neiberg, et al., 2008; and Demange & Ouni, 2013).

One the other hand, the direct realist view of speech perception (Fowler, 1986; 2006) holds that there is no inversion problem and that the mapping is not many-to-one. In this view of speech perception the acoustic signal “specifies” the vocal tract, so that the details of speech articulation are on full view for the listener based on information in the speech signal.

One difference in opinion between researchers who assume that there is a many-to-one mapping problem and those who assume that the acoustic signal fully specifies the vocal tract hinges on what Atal et al. (1978) called “situations of importance”, or more specifically in the number of dimensions that one would include in the acoustic vector *y*. Atal et al. operationalized their description of the acoustic space for vowels with information about the frequencies, bandwidths and amplitudes of the first three vocal tract resonances. Higher resonances were not considered to be important. The experiment that we report here tested the importance of F4 in vowel perception by synthesizing vowels that were reported by Atal et al. (1978) to illustrate the many-to-one mapping between articulation and acoustics. To preview our results, we find that listeners are well able to hear the differences between these “practically identical” vowels. This conclusion calls into question the suggestion that there is a many-to-one mapping problem for listeners.



F	292	2097	2802	3514	Hz
A	15	15	4	13	dB
B	35	56	482	188	Hz

F	292	2097	2802	3282	Hz
A	15	15	4	12	dB
B	36	61	728	223	Hz

F	292	2097	2802	3797	Hz
A	15	15	4	4	dB
B	35	53	373	586	Hz

F	292	2097	2802	3213	Hz
A	15	15	4	13	dB
B	36	60	738	185	Hz

Figure 1. Reproduced from Atal et al. (1978, figure 22). Four vocal tract area functions that result in /i/ vowels that have the same frequencies and amplitudes of the first three formants. Note that F4 frequency and bandwidth varies for these vowels.

Methods

Subjects. Fifteen listeners (7 men and 8 women) participated in the experiment. They were undergraduate and graduate students in linguistics at UC Berkeley. They were all native or near-native speakers of English. Three were bilingual speakers of English and Spanish (#103, 107, 109), two were bilingual in Mandarin and English (#104 & 105) and one was bilingual in English and Japanese (#101). All of the speakers had normal hearing ability (one subject, #114 had right ear hearing loss corrected to normal with tympanoplasty).

Stimuli. Four examples each of the vowels /u/, /i/, and /a/ were synthesized using the Klatt terminal analog speech synthesizer (Klatt & Klatt, 1990). The stimuli were 300 milliseconds long with a linearly falling pitch contour from 120 Hz to 90 Hz. The formant frequencies and formant bandwidths for these synthetic stimuli were taken from the results reported by Atal et al. (1978). For example, Figure 1 reprints Atal et al.'s figure 22. Here we see four versions of the vowel /i/ - four different vocal tract area functions and the formant frequencies (F), formant amplitudes (A), and formant bandwidths (B) for the first four vowel formants. The point of Atal et al.'s work is that very different vocal tract area functions can produce vowels that have identical values for F1-3. Our interest was in whether listeners would nonetheless be sensitive to the differences among these vowels on the basis of their F4 frequencies and amplitudes. Figure 2 shows spectra taken at vowel midpoint from the twelve steady-state vowel stimuli used in this experiment.

Procedure. Listeners heard pairs of vowel stimuli and judged whether they were the 'same' or 'different'. After eighteen practice trials with same or different pairs drawn from any of the three vowel qualities, the stimuli were presented blocked by vowel with 120 trials of /u/, then 120 trials of /i/, and then 120 trials of /a/. Each block was composed of five repetitions of 12 'different' pairs (all possible pairings of a different stimuli) and 12 'same' pairs (each of the four stimuli paired with itself, listed three times so there were 12 total). The list of 24 pairs was replicated 5 times for a total list of 120 trials and then presented in random order to listeners. On each trial, the screen was blank for one second, then the response options appeared on the screen (right button for 'different' responses, left button for 'same' responses) as

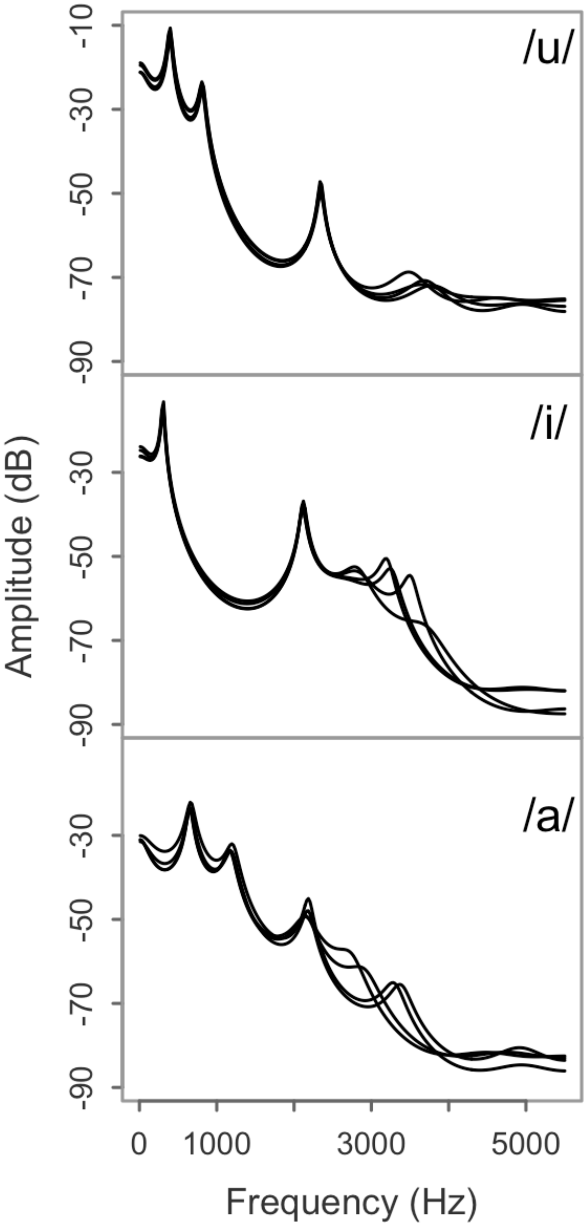


Figure 2. LPC Spectra of the 12 stimuli used in this experiment (10 kHz sampling rate, 12 LPC coefficients). For example, the /i/ stimuli were synthesized using the vowel formant frequencies and bandwidths shown in figure 1.

the pair of audio files was presented in sequence with a 250 msec inter-stimulus interval. The listener had 5 seconds to respond and after the response was registered a feedback message (“correct” or “incorrect”) was shown for 750 msec.

Results

Overall in this experiment, the hit rate (correct “different” responses) was 69%, while the false alarm rate (incorrect “different” responses) was only 45%. This results in a d’ value of 1.66 (Kaplan et al., 1978). Sensitivity in signal detection theory (d’) ranges from 0 when the hit rate is no different from the false alarm rate, to 6.9 when the hit rate 99% and the false alarm rate is 1%. So, a test of whether listeners are completely incapable of detecting a difference between stimuli in a discrimination task becomes a test of whether d’ is reliably different from zero.

To this end, we calculated d’ values using the R library “psyphy” (Knoblauch, 2014) separately for each listener’s responses to each vowel. The data set was thus made up of 45 d’ values (15 subjects * 3 vowels). We modeled these data using a linear mixed effects model with Vowel as a fixed effect (treatment coded) and with random intercepts for the subjects. Confidence intervals around the vowel coefficients were calculated by bootstrapped sampling the model 1000 times. The results (shown in Table 1) indicated that the d’ for the reference vowel /u/ was reliably greater than zero and that the coefficient for /i/ indicates that d’ for /i/ was slightly lower than for /u/ while the coefficienct for /a/ indicates that the d’ of /a/ was not reliably different from that for /u/.

Table 1. Model coefficients for the fixed effects in a linear mixed effects model [d_prime ~ vowel + (1|subj)]. Confidence intervals were calculated by bootstrapped sampling the model 1000 times.

	Estimate	95% CI	Std. Error	t value
(Intercept)	1.67	(1.23, 2.13)	0.22	7.45
voweli	-0.35	(-0.58, -0.11)	0.12	-2.88
vowela	0.02	(-0.21, 0.28)	0.12	0.23

An additional test of whether the d’ values for these vowels were different from zero was necessary because the mixed effects model does not directly ask this question. There was a reliably significant negative coefficient for the /i/ relative to the reference vowel [u], but we don’t know whether the magnitude of this coefficient was large enough to suggest that the d’ value for /i/ was not different from zero. So, we performed three additional t-tests, one for each vowel. These results (Table 2) confirm that the d’ value for all three vowels was reliably greater than zero. Figure 3 shows the d’ data in a box and whisker plot.

Table 2. T-tests showed that the observed d’ values were reliably different from zero. “CI” refers to the 95% confidence intervals, based on the *t* distribution.

Vowel	t-value	df	p-value	mean	95% CI
/u/	4.73	14	<0.001	1.64	(0.90, 2.39)
/i/	6.63	14	<0.001	1.32	(0.89, 1.75)
/a/	7.91	14	<0.001	1.72	(1.26, 2.19)

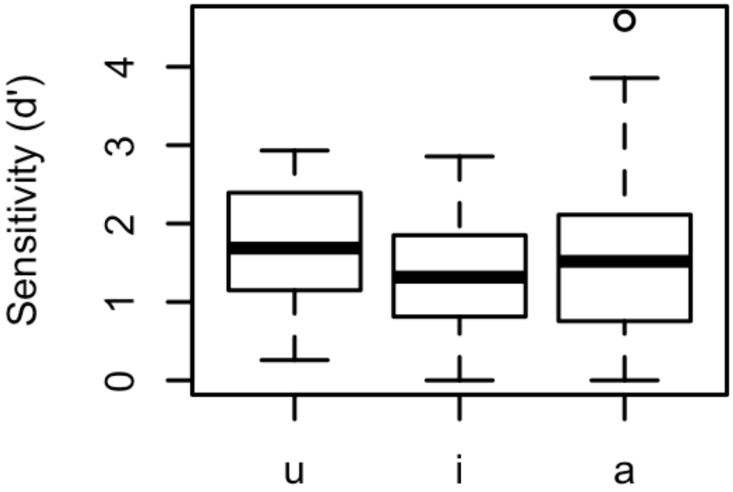


Figure 3. Box and whisker plots of d' values calculated separately for each test vowel. The vowels are shown in the order in which the blocks of trials were presented: /u/ trials in block 1, /i/ trials in block 2, and /a/ trials in block 3.

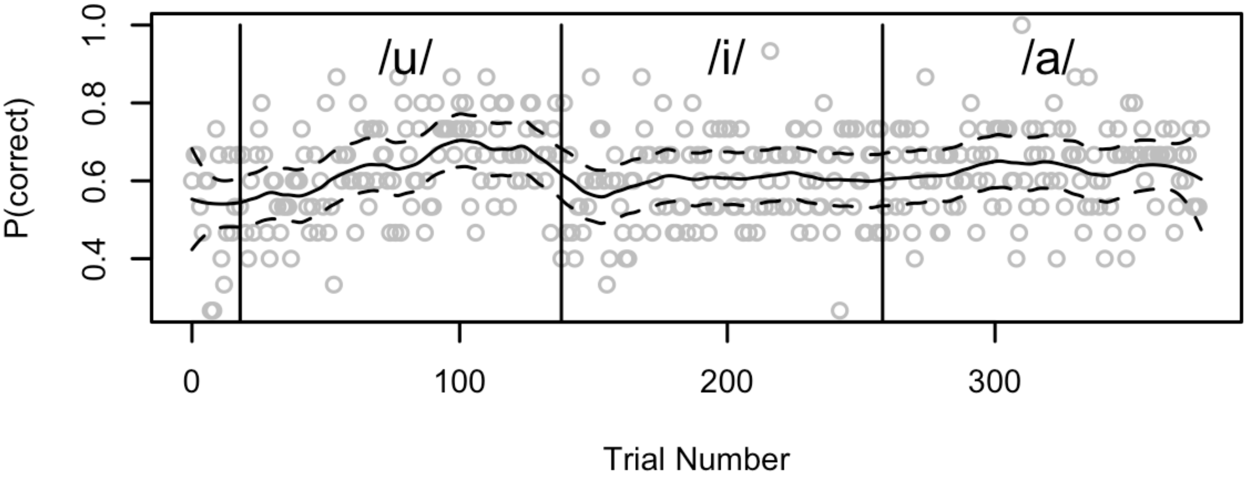


Figure 4. The total proportion of subjects (out of 15) who chose the correct response as a function of trial number. Because the order of stimuli was randomized separately for each subject these data are summed over trials that involved different stimuli and different correct response values ('same' or 'different' trials). The vowel blocks are indicated by the vertical black lines and vowel labels. For example, the /u/ block extended from trial 19 to trial 139. A loess curve (Cleveland et al., 1992) with 99% confidence interval is also plotted.

Finally, we examined the data for a training effect. The task used in this listening experiment provided feedback on every trial, and this presumably helped listeners achieve better performance than they would have reached without feedback. Figure 4 shows the proportion of correct responses as a function of trial number. Recall that in this experiment after 18 practice trials there were 120 trials with /u/ tokens, then 120 with /i/ tokens and then 120 with /a/ tokens. Because the order of the trials was randomized separately for each listener, the data in figure 4 are pooled over trials that had different stimuli and differed with regard to whether the correct response was 'same' or 'different'. The smoothed loess curve fit shows an increase of accuracy in responses to the /u/ stimuli, rising from 54% correct at the beginning of the /u/ block to a peak of 70% correct at trial 102. The range of performance (again taking the loess curve fit as our estimate of pooled performance) was smaller for the other two vowels in the experiment, with a low value of 56% and a peak of 62% for /i/, and a range from 60% to 65% for /a/. Overall, this pattern suggests that during the first block listeners were learning how to do the task, rather than learning specific acoustic values to listen for, because their experience with earlier blocks transferred to better performance on blocks of new stimuli.

Discussion

The results of this small study show that Atal et al. (1978) over-stated the extent of the many-to-one problem when it comes to human speech perception. Listeners can tell the difference pretty well between vocal tract shapes that Atal et al. considered to produce "the same" acoustic output. Our results here build on an earlier perceptual study (Johnson, 2011) which found that listeners are sensitive to F4 as a cue to the difference between retroflex and bunched /r/. Although listeners considered tokens with a raised F4 and tokens with a lowered F4 to be good examples of /r/, only those with the lowered F4 produced a compensation for coarticulation effect (compensation for tongue retraction) in perception. So, it is not completely surprising that listeners in the present study were sensitive to F4 differences.

Of course, these results do not suggest that there is no many-to-one problem in the inversion of the articulation to acoustics mapping, or that this problem is not a serious one in many practical applications where one would like to be able to perform such an inversion (see though Demange & Ouni, 2013 and references cited there). The results do suggest though that the direct realists' point on this question should be taken seriously when considering listeners' ability to perceive speech. The acoustic speech signal is rich with vocal tract information both in the dynamics of the signal and in spectral details that are often overlooked in simple models of vocal tract acoustics. Our data on listener sensitivity to F4 can't be taken as proof that listeners perceive vocal tracts, as claimed in direct realism, however these results do indicate that listeners are more sensitive to the rich acoustic speech signal than they are often given credit for.

References

Atal, B.S.; Chang, J.J.; Mathews, M.V. & Tukey, J.W. (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.* **63**(5), 1535-1555.

Cleveland, W.S.; Grosse E. and W. M. Shyu (1992) Local regression models. Chapter 8 of *Statistical Models in S.* Edited by Chambers, J.M. and Hastie, T.J. Wadsworth & Brooks/Cole.

Demange, S. & Ouni, S. (2013) An episodic memory-based solution for the acoustic-to-articulatory inversion problem. *J Acoust. Soc. Am.* **133**(5), 2921-2930.

Fowler, C. A. (1986) An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* **14**, 3-28.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics* **68**, 161-177.

Hogden, J.; Lofqvist, A.; Gracco, V.; Zlokarnik, I.; Rubin, P. and Saltzman, E. (1996) “Accurate recovery of articulator positions from acoustics: New conclusions based on human data, *J. Acoust. Soc. Am.* **100**, 1819–1834.

Johnson, Keith (2011) Retroflex versus bunched [r] in compensation for coarticulation. *UC Berkeley Phonology Lab Annual Report 2011*, 114-127.

Kaplan, H.L.; Macmillan, N.A. & Creelman, C.D. (1978) Tables of d’ for variable-standard discrimination paradigms. *Behavior Research Methods and Instrumentation* **10**(6), 796-813.

Klatt, D.H. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* **87**(2), 820-857.

Knoblauch, K. (2014) Functions for analyzing psychophysical functions. psyphy-package, Version 0.1-9.

MacMillan, N.A. & Creelman, C.D. (2005) *Detection Theory: A User’s Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Neiberg, D.; Ananthakrishnan, G. and Engwall, O. (2008) The acoustic to articulatory mapping: Non-linear or non-unique?, in *Proceedings of Interspeech*, Brisbane, Australia, pp. 1485–1488.

Papcun, G.; Hochberg, J.; Thomas, T.R.; Laroche, F.; Zacks, J. and Levy, S. (1992) Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, **92**(2), 688-700.

Schroeter, J. and Sondhi, M.M. (1994) Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trns. Speech and Audio Processing*, **2**(1), 133-150.