# Individual- and group-level associations between articulatory and acoustic variability for English consonants

Sarah Harper

University of California, San Francisco
sarah.harper@ucsf.edu

## ABSTRACT

This study extends previous research probing the relationship between articulatory and acoustic variability by examining the extent to which individual differences in articulatory variability are recoverable from acoustics, and vice versa, for a subset of American English consonants. Articulatory (constriction location, degree, and orientation) and acoustic (formant and spectral) measurements were extracted from tokens of /s/, /ʃ/, /l/, and /ɹ/ for 40 speakers in the Wisconsin X-Ray Microbeam Corpus. Analysis of goodness-of-fit statistics from linear mixed effects models fit to the data suggest that variability observed in articulation is generally recoverable in acoustics, and vice versa, with change in one domain consistently predicted by change in the other. Additionally, the comparison of dispersion metrics from raw and model-predicted values for each speaker indicate that interspeaker differences in variability are to some extent transmissible between articulation and acoustics, although this capacity varies substantially across segments and across measured dimensions.

**Keywords**: Articulatory-Acoustic Relations, Variability, Individual Differences

## 1. INTRODUCTION

A substantial body of research on articulatory-acoustic relations in speech production has examined the extent to which variation in articulation is encoded in acoustics (and vice versa). Early research in this area generally suggested that the non-linear and quantal mapping observed between articulation and acoustics [1] would lead to a dissociation of the variability observed in the realization of phonological segments across these modalities. Specifically, it was both predicted and observed that the articulatory patterns used to produce any phonological segment would be more variable than the resulting acoustic signal [2, 3].

Recent work on articulatory-acoustic relations in vowel production, however, has cast doubt upon the existence of a consistently asymmetric relationship between articulatory and acoustic variability, and upon the perspective that variation in articulation is obscured in the acoustic signal more generally. Examining individual differences in the realization of vowel height contrasts, Noiray et al. [4] found that speakers' idiosyncratic articulatory strategies for contrasting front vowel pairs were reflected in acoustics, with speakers who showed articulatory tongue height 'reversals' (i.e., producing a high vowel with a lower tongue position than a mid vowel) demonstrating a comparable reversal in first formant values. Whalen et al. [5] similarly found that articulatory variability was positively correlated with acoustic variability across vowels for speakers of American English, and that speakers differed from one another in whether they exhibited more variability in articulation or acoustics for the same vowel segments.

In this study, we extend previous research on articulatory-acoustic relations by examining whether patterns of variability in consonant production are recoverable between articulation and acoustics. Although studies support the idea that variability is, at least for vowels, relatively comparable in articulation and acoustics [4, 5], the articulatory constrictions used in consonant production could interact with the nonlinearities of vocal tract acoustics in a manner that minimizes acoustic variability (as suggested by [6]) and obscures the transmission of variability between these domains. As such, the extent to which a similarly comparable relationship between articulatory and acoustic variability exists for consonants remains unclear.

## 2. METHOD

### 2.1. Corpus and data selection

Articulatory and acoustic data for this study was taken from recordings of 40 American English speakers in the Wisconsin x-ray Microbeam (XRMB) database [7]. The kinematic articulatory data in the XRMB corpus consists of positional trajectories of the movement of small pellets (2.5 mm) attached to the tongue, jaw, and lip and to stable reference points in the mouth (Fig. 1). Acoustic data consists of audio recordings captured synchronously with the kinematic data.

A total of 7,298 tokens of word-initial and -final /s/, /ʃ/, /l/, and /ɹ/ were extracted from sentences read by participants in two of the corpus tasks, a Sentence Reading task and a Prose Passage Reading task. These consonants were chosen for analysis both because they are known to exhibit substantial variability in articulatory posture across and within speakers in American English ([8]-[12]) and because they are continuants with information about their identity transmitted acoustically throughout their production.

### 2.2. Articulatory measurement

The gestural movement extremum (**MAXC**) in each token was automatically identified using a modified version of the *findgest* algorithm (originally by Mark Tiede, Haskins Laboratories). Using an acoustic segmentation of the XRMB data created with the Penn Forced Aligner [13], tokens of the consonants of interest were automatically located in the articulatory data by finding articulatory frames corresponding to their acoustically defined segment start and end points. These frames defined a search window over which the algorithm looked for MAXC, defined as the velocity minimum closest to the token's acoustic midpoint. For tokens of /s/, /ʃ/, and /l/, MAXC was extracted using the 2D tangential velocity signal of the lingual pellet closest to the tongue tip (T1). For /ɹ/, MAXC was extracted for the lingual pellet calculated to have the smallest distance from the palate at its velocity minimum (either T1, T2 or T3).

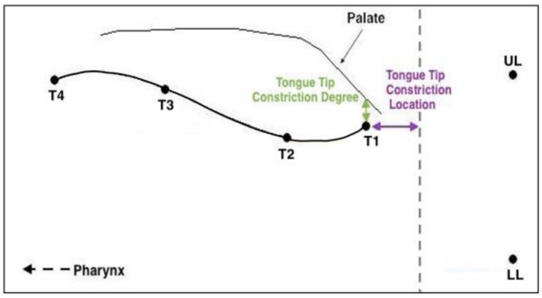
Palate
UL
Tongue Tip Constriction Degree
Tongue Tip Constriction Location
T4
T3
T2
T1
Pharynx
LL

**Figure 1**: Position of XRMB pellets with palate trace and schematic representation of constriction measurements. Dotted vertical line indicates the origin of the coordinate system used for location measurement.

Positional measurements of the upper and lower lips and the tongue were extracted for all lingual and labial pellets in the XRMB data at MAXC (Fig. 1). Constriction location (**CL**) and degree (**CD**) were logged for the pellet closest to the palate trace for each token (T1 for all tokens of /s/, /ʃ/, and /l/, and either T1, T2, or T3 for /ɹ/). CL was calculated as the x-axis distance of the pellet of interest from the coordinate system origin (at the tip of the maxillary incisors) and normalized by the length of the speaker's vocal tract. CD was calculated as the 2D Euclidean distance between the pellet's coordinate position and the closest point on the speaker's palate trace.

### 2.3. Acoustic measurement

The acoustic dimensions chosen for analysis differed between the examined liquids (/l/ and /ɹ/) and the fricatives (/s/ and /ʃ/). For the liquids, values of the first four formants (**F1-F4**) were automatically extracted at MAXC or, when MAXC occurred outside of the voicing interval for the target segment, at the closest voiced interval to this timepoint (following [14], [15]). Formant tracking was configured to find five formants below 5000 Hz for male speakers and 5500 Hz for female speakers. For fricatives, the first four spectral moments (**M1-M4**) were calculated using a DFT over a 50 ms Hamming window centered on MAXC. A high-pass filter with a cut-off at 500 Hz was applied to the fricative spectrum prior to spectral moment calculation to exclude spectral energy reflecting vocal fold vibration.

Tokens with one or more formant or spectral moment more than 2.5 standard deviations outside the speaker mean were visually inspected and manually corrected in Praat [16].

### 2.4. Statistical analysis of articulatory-acoustic relations

To directly examine relationships between articulatory and acoustic dimensions for each segment, a series of linear mixed effects models (LMMs) were fit to the data using the lmerTest package in R [17]. For each consonant of interest, two models were fit with an articulatory dependent variable (CD or CL) and the full set of formant or spectral moment measurements as fixed factors, and four models were fit with an acoustic dependent variable (F1-F4 or M1-M4) and the full set of XRMB pellet positions as fixed factors[1]. Random intercepts for Speaker and Phonetic Context[2] were included in all models. These models were used to evaluate (1) how well variation in articulation or acoustics was explained by the other domain across the group of speakers, and (2) the extent to which individual differences in variability are conveyed between articulation and acoustics.

Group-level encoding of variation between articulation and acoustics was evaluated using marginal and conditional $R^2$ values ($\mathbf{R^2_M}$ and $\mathbf{R^2_C}$) calculated for each LMM using the R package MuMIn [18]. $R^2_M$ represents the variance explained by the fixed effect structure of a model, while $R^2_C$ represents the variance explained by the fixed and random effect structures combined. An additional metric representing the variance explained by the random effects structure alone ($\mathbf{R^2_R}$) was calculated as the absolute difference between $R^2_M$ and $R^2_C$.

An analysis comparing patterns of variability in model-predicted values to those observed in raw data was conducted to evaluate how well interspeaker differences in variability are conveyed between articulation and acoustics. The predicted values for each datapoint in each LMM were calculated using the *predict* function in the lme4 package in R [19]. The interquartile range of each model's predicted values ($\mathbf{IQR_{PRED}}$) and the raw XRMB corpus data used as its dependent variable ($\mathbf{IQR_{REAL}}$) were calculated separately for each speaker and compared with Spearman's rank-order correlation.

## 3. RESULTS

### 3.1. Group-level encoding of variation

$R^2_M$ and $R^2_R$ for all LMMs fit to articulatory data are given in Table 1. $R^2_M$ values are highest for /l/ in the model fit to CL and for /l/ and /ɹ/ in the model fit to CD, suggesting that multi-dimensional acoustic space tends to explain a larger proportion of variance along the examined articulatory dimensions for the liquid consonants. The opposite pattern is observed for $R^2_R$, with higher values generally observed for the fricatives than for the liquids. This indicates that individual differences in the articulatory-acoustic mapping across speakers have greater explanatory power in the prediction of articulation for the examined fricatives than for the examined liquids.

**Table 1**: $R^2_M$ and $R^2_R$ values for all LMMs fit to an articulatory dependent variable.

| | | /s/ | /ʃ/ | /l/ | /ɹ/ |
|---|---|---|---|---|---|
| $R^2_M$ | CL | 2% | 10% | 12% | 8% |
| | CD | 3% | 10% | 16% | 22% |
| $R^2_R$ | CL | 85% | 75% | 44% | 32% |
| | CD | 56% | 59% | 15% | 44% |

$R^2_M$ and $R^2_R$ for all LMMs fit to acoustic data are given in Tables 2 and 3. Inspection of the $R^2_M$ values for these models suggest that multi-dimensional articulatory space explains a higher proportion of variance in acoustic dimensions for /ɹ/ than it does for the other consonants examined (with the high $R^2_M$ in the M1 model for /ʃ/ and F1 for /l/ notable exceptions). $R^2_M$ values are generally higher than they were for the models fit to articulatory dependent variables, indicating that the acoustic variation produced by different articulatory vectors is somewhat more unique than the set of possible articulatory vectors suggested by a particular acoustic signal. This is consistent with the view that similar acoustic signals can result from different vocal tract configurations [20]. $R^2_R$ values are also lower for the fricative consonants in this set of models than they were in the LMMs with articulatory dependent variables, potentially pointing towards lesser predictive power for individual differences in articulatory-to-acoustic mapping than in the acoustic-to-articulatory direction.

**Table 2**: $R^2_M$ values for all LMMs fit to an acoustic dependent variable.

| | | /s/ | /ʃ/ | | /l/ | /ɹ/ |
|---|---|---|---|---|---|---|
| $R^2_M$ | M1 | 17% | 31% | F1 | 31% | 32% |
| | M2 | 13% | 21% | F2 | 17% | 28% |
| | M3 | 8% | 17% | F3 | 3% | 24% |
| | M4 | 3% | 18% | F4 | 9% | 10% |

**Table 3**: $R^2_R$ values for all LMMs fit to an acoustic dependent variable.

| | | /s/ | /ʃ/ | | /l/ | /ɹ/ |
|---|---|---|---|---|---|---|
| $R^2_R$ | M1 | 48% | 48% | F1 | 32% | 28% |
| | M2 | 49% | 38% | F2 | 51% | 40% |
| | M3 | 36% | 48% | F3 | 54% | 42% |
| | M4 | 6% | 32% | F4 | 53% | 57% |

### 3.2. Encoding of individual differences in variability

Results of the correlation analyses of $IQR_{REAL}$ and $IQR_{PRED}$ for articulatory dimensions suggest $IQR_{PRED}$ is usually a significant predictor of $IQR_{REAL}$, with a positive relationship consistently observed such that speakers who are more variable in their actual production tend to also have more variability in their predicted data (Fig. 2).

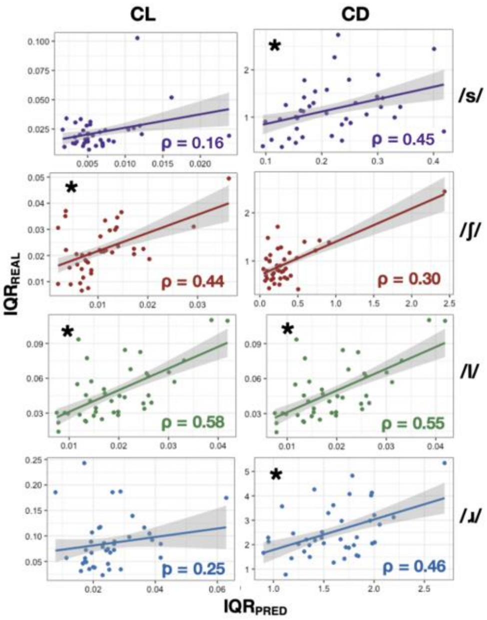

**Figure 2**: $IQR_{REAL}$ and $IQR_{PRED}$ comparisons across speakers for CD (left) and CL (right) in each consonant. Correlation coefficients are shown in the bottom right corner of each graph. Asterisk in the top left corner indicates significance.

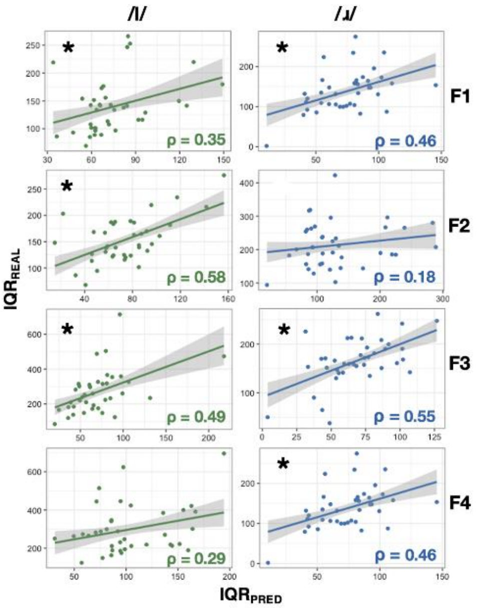

**Figure 3**: $\text{IQR}_{\text{REAL}}$ and $\text{IQR}_{\text{PRED}}$ comparisons across speakers for acoustic dimensions in liquid consonants.

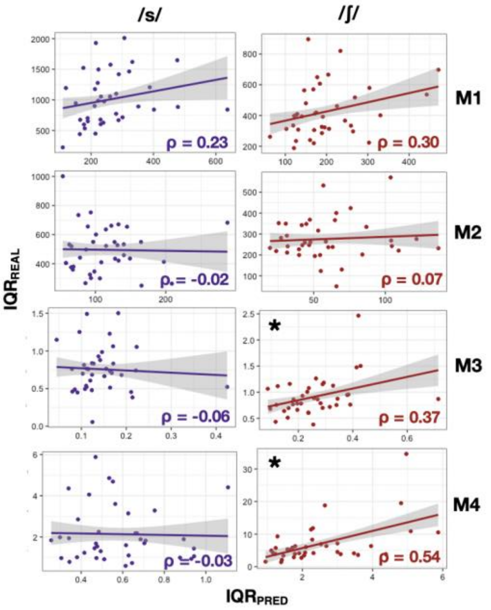

**Figure 4**: $\text{IQR}_{\text{REAL}}$ and $\text{IQR}_{\text{PRED}}$ comparisons across speakers for acoustic dimensions in fricative consonants.

The strength of the relationship between $\text{IQR}_{\text{PRED}}$ and $\text{IQR}_{\text{REAL}}$ varies substantially across segments and articulatory dimensions in the examined data, indicating that the recoverability of articulatory variability from the acoustic signal depends on which articulatory dimensions are being examined in which consonant segments. However, the observation that most individual comparisons (5 out of 8) are statistically significant suggests articulatory variability tends to be recovered from the acoustic signal across consonants and articulatory dimensions.

Correlation analyses of $\text{IQR}_{\text{REAL}}$ and $\text{IQR}_{\text{PRED}}$ for acoustic dimensions in /l/ and /ɹ/ similarly tend to exhibit a significant positive relationship (Fig. 3), with 6 out of 8 comparisons following this pattern. However, $\text{IQR}_{\text{PRED}}$ is not as consistently a significant predictor of $\text{IQR}_{\text{REAL}}$ for the fricative consonants (Fig. 4), with only 2 out of 8 comparisons significant across dimensions for /s/ and /ʃ/. These results indicate that the ability to recover acoustic variability from the articulatory signal may be quite sensitive to the identity of the consonant examined.

## 4. DISCUSSION

The results of this study confirm that both articulatory and acoustic variability are encoded in the other domain for consonants, although the strength of this relationship depends on the specific consonants and dimensions examined. Statistical models approximating articulatory-acoustic and acoustic-articulatory relationships in American English consonants were frequently able to use the variation observed in one domain to predict change in the other, as demonstrated by the prevalence of $R^2_M$ values accounting for at least 10% of the variation in predicted dimensions. These models were also largely successful at predicting individual differences in variability along specific articulatory or acoustic dimensions, suggesting a fine-tuned ability to predict how components in the acoustic signal reflect changes in articulation and to associate specific changes in the acoustic signal with underlying articulatory actions.

The capacity for variability in articulation to be recovered from acoustics, and vice versa, differs across articulatory and acoustic dimensions and across consonants. However, there is a general tendency for variability to be encoded across articulation and acoustics, in line with recent research on articulatory-acoustic relations in vowels. These results suggest that in natural speech there is less of a dissociation between articulatory and acoustic variability in the production of a single consonant than has been suggested by previous research. This may indicate that the functional consequences of vocal tract nonlinearities on acoustic-articulatory relationships may be smaller than traditionally thought (e.g., [6]), and that considerable information about inter-and intraspeaker patterns of articulatory variability may be accessible to listeners in the acoustic signal. Further research will be necessary to explore this possibility and its potential implications for the cognitive systems underlying production and perception.

## 5. REFERENCES

[1] Stevens, K. 1972. The quantal nature of speech: Evidence from articulatory-acoustic data. In: David, E., Denes, P.

(eds), *Human Communication: A Unified View*. McGraw Hill, 51-66.

[2] Johnson, K., Ladefoged, P., & Lindau, M. 1993. Individual differences in vowel production. *J. Acoust. Soc. Am.* 94(2), 701-714.

[3] Guenther, F., Husain, F., Cohen M., & Shinn-Cunningham, B. 1999. Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* 106, 2900-2912.

[4] Noiray, A., Iskarous, K., & Whalen, D. 2014. Variability in English vowels is comparable in articulation and acoustics. *Laboratory Phonology* 5(2), 271-288.

[5] Whalen, D., Chen, W., Tiede, M., & Nam, H. 2018. Variability of articulator positions and formants across nine English vowels. *J.Phon*. 68, 1-14.

[6] Nieto-Castanon, A., Guenther, F., Perkell, J., & Curtin, H. 2005. A modelling investigation of articulatory variability and acoustic stability during American English /r/ production. *J. Acoust. Soc. Am.* 117(5), 3196-3212.

[7] Westbury, J. 1994. *X-ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin, Madison, WI.

[8] Bladon, R., & Nolan, F. 1977. A video-fluorographic investigation of tip and blade alveolars in English. *J.Phon.* 5, 185-193.

[9] Dart, S. 1998. Comparing French and English coronal consonant articulation. *J.Phon.* 26, 71-94.

[10] Delattre, P., & Freeman, D. 1968. A dialect study of American r's by x-ray motion picture. *Linguistics* 44, 29-68.

[11] Lindau, M. 1985. The story of /r/. In: Fromkin, V. (ed), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*. Academic Press, 157-168.

[12] Mielke, J., Baker, A., & Archangeli, D. 2016. Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language* 92, 101-140.

[13] Yuan, J., & Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proc. Acoustics '08*, 5687-5690.

[14] Lin, S., Beddor, P., & Coetzee, A. 2014. Gestural reduction, lexical frequency, and sound change: A study of post-vocalic /l/. *Laboratory Phonology* 5(1), 9-36.

[15] Lawson, E., Stuart-Smith, J., & Scobbie, J. 2018. The role of gesture delaty in coda /r/ weakening: An articulatory, auditory and acoustic study. *J. Acoust. Soc. Am.*143, 1646-1657.

[16] Boersma, P., & Weenink, D. 2020. Praat: doing phonetics by computer (Version 6.1.16) [Computer software]. http://www.praat.org/.

[17] R Core Team. 2021. R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

[18] Barton, K. 2020. MuMIn: Multi-model inference. R package version 1.43.17. https://CRAN.Rproject.org/package=MuMIn.

[19] Bates, D., Maechler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software* 67(1), 1-48.

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

[1] The full set of pellet positions was used instead of derived articulatory positional measurements due to the potential for the entire vocal tract, not just the examined dimensions of interest, to affect the acoustic signal. Models using formant and spectral measurements as fixed effects were selected for analysis after observing that they outperformed models with MFCCs in the fixed effects structure.

[2] Phonetic Context defined as the segmental environment, word position (initial or final), and phrasal position (boundary-adjacent or phase-internal) of a token.