Eleanor Chodroff* and Colin Wilson

Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice

https://doi.org/10.1515/lingvan-2017-0047
Received October 15, 2017; accepted July 5, 2018

Abstract: The present study investigates patterns of covariation among acoustic properties of stop consonants in a large multi-talker corpus of American English connected speech. Relations among talker means for different stops on the same dimension (between-category covariation) were considerably stronger than those for different dimensions of the same stop (within-category covariation). The existence of between-category covariation supports a uniformity principle that restricts the mapping from phonological features to phonetic targets in the sound system of each speaker. This principle was formalized with factor analysis, in which observed covariation derives from a lower-dimensional space of talker variation. Knowledge of between-category phonetic covariation could facilitate perceptual adaptation to novel talkers by providing a rational basis for generalizing idiosyncratic properties to several sounds on the basis of limited exposure.

Keywords: stop consonants; talker variability; phonetic covariation; factor analysis; predictability.

1 Introduction

The phonetic realization of an individual sound category can vary substantially according to contextual, lexical, dialectal, and talker-specific influences. This variation is highly structured: previous research has documented strong dependencies among phonetic properties, as well as between phonetic properties and many sociolinguistic factors (e.g. Labov 1966; Foulkes et al. 2001; Foulkes and Docherty 2006; Guy and Hinskens 2016; Fruehwald 2017; Sonderegger et al. 2017). The present study focuses on two prominent types of linear dependency in phonetic variation. The first type of dependency holds among multiple categories along a single phonetic dimension ("between-category" covariation); the second holds among multiple phonetic dimensions within individual categories ("within-category" covariation).[1]

Instances of between-category phonetic dependencies have been observed among several speech sounds. Talker-specific vowel systems differ extensively in the log F1 × F2 formant plane, but the systems are highly parallel, suggesting covariation along these dimensions (e.g. Joos 1948; Nearey 1978; Nearey and Assmann 2007). Furthermore, largely constant spectral and temporal ratios are preserved among vowel categories across speaking rates and styles (Smiljanić and Bradlow 2008; DiCanio et al. 2015). Relations among vowels can also be preserved during diachronic sound change, as when multiple vowels undergo parallel shifts in their phonetic realization (e.g. Fruehwald 2013, Fruehwald 2017). Among fricatives, the spectral centers of gravity of [s] and [ʃ] vary substantially across talkers, yet

1 There may be other forms of statistical dependency, beyond linear relations, among phonetic variables. It is also conceivable that the values of one category on a given dimension could covary with those of another category on a different dimension. We considered this alternative notion of "between-category" covariation but found limited evidence for it in the present data.

***Corresponding author: Eleanor Chodroff,** Department of Linguistics, Northwestern University, Evanston, IL, USA, E-mail: eleanor.chodroff@northwestern.edu. http://orcid.org/0000-0003-4549-5919
Colin Wilson: Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA

within a talker, the mean COG of [s] is systematically higher than the corresponding mean COG of [ʃ] (Newman et al. 2001). Strong covariation of mean voice onset time (VOT) has also been observed among stop consonants across speakers of the same language (e.g. Zlatin 1974; Koenig 2000; Newman et al. 2001; Solé 2007; Theodore et al. 2009; Chodroff and Wilson 2017). Theodore et al. (2009) observed a similar difference between the mean VOT values of [p^h] and [k^h] across talkers. Chodroff and Wilson (2017) extended the study of between-category VOT covariation to all word-initial stop categories of American English (AE), in both isolated and connected speech, while controlling for many other sources of VOT variation (e.g. utterance position, following vowel, lexical properties). Correlations of talker VOT means were particularly strong among the voiceless aspirated stops, and moderate among the voiced stops and homorganic voiced pairs.

There is substantially less evidence for within-category phonetic dependencies across tokens and across talker means. For example, while vowel height (as indexed by F1) and vowel duration are known to covary across vowels in many languages (e.g. Lindblom 1967; Maddieson 1997), this relation does not appear to hold across individual tokens of the same vowel category (Toivonen et al. 2015).[2] Several studies have also examined the possibility that different cues to the voicing contrast, such as VOT and fundamental frequency (f0) at following vowel onset, covary within stop categories (e.g. Shultz et al. 2012; Dmitrieva et al. 2015; Kirby and Ladd 2015, Kirby and Ladd 2016; Clayards 2018). Positive correlations of the relevant cues would indicate enhancement of the contrast, whereas negative correlations would suggest cue-trading or compensation relations. These relationships could hold across tokens or across talker means, and would indicate systematic relations in the *use* of phonetic dimensions. The observed within-category dependencies tend to be weak and vary considerably by language and sample. Kirby and Ladd (2015) found a significant negative correlation between VOT and f0 across tokens of word-initial [p] in Italian, but this correlation did not reach significance in French. A weak negative correlation between VOT and f0 was observed across tokens of AE [p^h] by Dmitrieva et al. (2015), but another study of the same language yielded no significant linear relation between those dimensions for [p^h] or [b] across tokens or talker means (Clayards 2018).

Some factors involved in phonetic realization may induce both between- and within- category covariation. For example, Koenig (2000) found that median VOT and following vowel duration were highly correlated across talkers for both [p^h] ($r = 0.72$) and [t^h] ($r = 0.77$). This likely reflects talker specificity in global speaking rate, leading to the expectation that correlations would also be found between the stop categories and for other duration-sensitive cues. Covariation among cues that occurs both between and within categories may be reducible to global factors such as speaking rate or airflow rate; this point will be considered further in the discussion.

We examined covariation within and among the six AE word-initial stop consonants, focusing on three well-known cues to the place and voice contrasts: spectral center of gravity (COG; e.g. Forrest et al. 1988), positive VOT (e.g. Lisker and Abramson 1964), and f0 at vowel onset (e.g. Haggard et al. 1970; Ohde 1984). Correlation analyses (Section 2) revealed considerably stronger between-category covariation than within-category covariation. This accords with previous findings but is comprehensively demonstrated for the first time here, with all measurements performed on the same multi-talker data set. The between-category correlations can be accounted for with a principle of *uniformity* that constrains the mapping from phonological feature values to talker-specific phonetic targets (Chodroff 2017; Chodroff and Wilson 2017). This principle was quantitatively formalized and evaluated against the observed correlations within the dimensionality-reduction framework of factor analysis (Section 3). As we discuss in Section 4, covariation in phonetic realization across speakers implies predictability for listeners: listeners could use phonetic dependencies among stop categories to generalize from limited experience with a novel talker.

2 Strong pairwise correlations of talker mean log f0, F1, F2, and F3 have been observed when aggregated over all vowels (Nearey 1989; Assmann et al. 2008; see also Rose 2010 for F2 and F3, and Whalen and Levitt 1995 for f0 and F1). It remains unclear, however, which particular vowel categories exhibit these dependencies most strongly.

2 Correlation analysis

2.1 Methods

The data was extracted from an audited subset of the Mixer 6 corpus (Brandschain et al. 2010, Brandschain et al. 2013; Chodroff et al. 2016) containing approximately 45 minutes of read speech from 180 native AE speakers (102 female). Transcripts were aligned to the corresponding WAV files with the Penn Forced Aligner (Yuan and Liberman 2008), and all word-initial prevocalic stop consonants were further processed with AutoVOT (Keshet et al. 2014). AutoVOT automatically identifies the stop release and following vowel onset within a user-specified window of analysis. Further details about the talkers, read sentences, and boundary alignments can be found in Chodroff and Wilson (2017).[3]

COG, positive VOT, and onset f0 in the following vowel were measured for each stop. COG was calculated from a smoothed spectrum over the initial portion of the release burst. Each spectrum was computed by averaging FFTs from seven consecutive 3 ms windows, with the first window centered on the burst transient and a window shift of 1 ms (Hanson and Stevens 2003; Flemming 2007; Chodroff and Wilson 2014). Positive VOT was defined as the duration from stop release to the onset of periodicity in the vowel; this was automatically extracted from the AutoVOT boundaries or from manually-corrected boundaries when available. The f0 value was the first one measured by Praat (Boersma and Weenink 2016) within 50 ms after the following vowel onset.

For each stop category and cue separately, values 2.5 standard deviations above or below the talker-specific mean were excluded. Consequently, the total number of valid tokens varied somewhat by stop category and measurement (COG: 87,968 tokens; VOT: 96,357; onset f0: 76,144). Table 1 summarizes the data available per talker for each stop and cue combination.

2.2 Results

For each stop and cue combination, talker means were calculated from all available tokens (e.g. a talker's mean COG for [p^h] was calculated from all of his or her productions of that stop with non-outlier COG values). Pearson correlations were performed on the talker means between stop categories along a single dimension and, separately, between dimensions within each stop category. The bias-corrected and accelerated percentile (BCa) method was used to form 95% bootstrapped confidence intervals for the correlations (1000 replicates; Efron 1987). As shown in Table 2, the between-category correlations among stops were positive and

Table 1: For each stop category and measurement separately, the median number of tokens per talker (left column) and the range of tokens per talker (right column).

	COG		VOT		f0	
	Median	**Range**	**Median**	**Range**	**Median**	**Range**
p^h	75	45–98	77	44–100	59.5	6–96
b	86	57–127	98	64–138	80	9–127
t^h	45	16–75	46	17–77	37	4–67
d	115	53–170	140	64–192	113	12–173
k^h	91	48–112	93	50–114	77	4–110
g	82	52–116	91	54–122	78	5–111

3 The dataset here was somewhat larger than that analyzed in Chodroff and Wilson (2017), which included only stops at the beginning of *stressed* word-initial syllables. Because all talkers recorded the same set of sentences, any effects of stress (or other contextual factors) on VOT and other acoustic properties should be approximately consistent. We aimed to include as many tokens as possible to maximize the power available to identify phonetic dependencies.

Table 2: Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of stop-specific talker means for COG, VOT, and f0.

	COG		VOT		f0	
p^h–b	0.60	[0.49, 0.69]	0.17, *n.s.*	[0.00, 0.32]	0.98	[0.97, 0.98]
t^h–d	0.66	[0.57, 0.73]	0.53	[0.43, 0.63]	0.97	[0.94, 0.98]
k^h–g	0.69	[0.58, 0.76]	0.43	[0.32, 0.52]	0.97	[0.94, 0.98]
p^h–t^h	0.40	[0.26, 0.51]	0.83	[0.77, 0.88]	0.98	[0.97, 0.99]
t^h–k^h	0.47	[0.34, 0.58]	0.79	[0.73, 0.83]	0.98	[0.95, 0.99]
k^h–p^h	0.57	[0.45, 0.65]	0.83	[0.78, 0.87]	0.98	[0.98, 0.99]
b–d	0.55	[0.40, 0.66]	0.11, *n.s.*	[−0.03, 0.24]	0.98	[0.98, 0.99]
d–g	0.67	[0.58, 0.75]	0.37	[0.24, 0.50]	0.98	[0.94, 0.99]
g–b	0.63	[0.51, 0.72]	0.48	[0.37, 0.58]	0.98	[0.95, 0.98]

All *p*-values were less than 0.001 unless otherwise indicated.

Table 3: Pearson correlation coefficients and 95% BCa bootstrap confidence intervals of stop-specific talker means between COG, VOT, and f0.

	COG-VOT		COG-f0 (female)		COG-f0 (male)		VOT-f0 (female)		VOT-f0 (male)	
p^h	0.32*	[0.17, 0.44]	0.11	[−0.08, 0.29]	0.12	[−0.09, 0.37]	0.09	[−0.09, 0.28]	−0.03	[−0.22, 0.20]
b	0.37*	[0.25, 0.50]	0.03	[−0.17, 0.22]	−0.16	[−0.34, 0.05]	−0.11	[−0.33, 0.12]	−0.09	[−0.30, 0.13]
t^h	0.22	[0.07, 0.38]	0.10	[−0.08, 0.30]	0.11	[−0.13, 0.34]	0.03	[−0.19, 0.23]	−0.02	[−0.25, 0.21]
d	0.74*	[0.66, 0.80]	0.03	[−0.17, 0.24]	−0.19	[−0.39, 0.02]	0.00	[−0.20, 0.21]	0.00	[−0.24, 0.23]
k^h	0.23	[0.09, 0.35]	0.07	[−0.12, 0.25]	0.05	[−0.18, 0.28]	0.17	[−0.02, 0.36]	−0.08	[−0.30, 0.15]
g	0.54*	[0.43, 0.63]	0.12	[−0.09, 0.27]	−0.14	[−0.34, 0.09]	0.08	[−0.13, 0.26]	−0.06	[−0.26, 0.17]

For f0, the correlations are reported separately for female and male talkers. Correlations with *p*-values less than 0.001 are identified with an asterisk.

generally high.[4] For the COG dimension, moderate correlations were observed among the voiceless stops, and strong correlations were observed among the voiced stops and between homorganic stop pairs (e.g. [k^h]–[g]). For VOT, the pattern of correlations replicated that found in Chodroff and Wilson (2017): relations were very strong among the voiceless stops, and moderate to weak among the voiced stops and homorganic pairs. Finally, talker mean f0 was almost perfectly correlated for all stops.[5]

Within-category correlations of talker means are shown in Table 3. Note that for f0, the correlations were calculated separately for male and female talkers. COG and VOT means showed significant positive correlations within each of the stop categories, but the strength of the relation depended on the category. These two cues were weakly correlated within the voiceless stops and [b], but strongly correlated within [d] and moderately within [g]. Within-category correlations between COG and f0, and VOT and f0, were numerically quite weak and none reached significance. Additional within-category correlations calculated over tokens (instead of over talker means) are reported in the Appendix.

2.3 Discussion

We found strong between-category covariation on each dimension for the segments and phonetic cues investigated here. In part, this surely reflects anatomical differences among talkers: differences that affect the articulation and resulting acoustics of many sounds (e.g. the strong dependencies in onset f0 are partly due to cross-talker variation in vocal fold length and tissue density; Titze 2011). But anatomy does not wholly determine phonetic realization. Each talker could in principle have shown greater between-category differences in the phonetic targets that are indexed by COG (e.g. by specifying tongue tip position and contact width

4 The significance level was adjusted for multiple comparisons to the conservative value of $\alpha = 0.001$.

5 Correlations are described using modifiers based on recommendations in Evans (1996): a "strong" correlation is one above 0.59, a "moderate" correlation is between 0.40 and 0.59, and a "weak" correlation is below 0.40.

for coronal [t] differently than for [d]), VOT (e.g. by planning the duration or timing of glottal spreading for [p^h] differently than for [k^h]), and even f0 (e.g. by having a different pitch target for vowels following [d] than for those after [g]). Indeed, research on language- and dialect-specific phonetics (e.g. Lisker and Abramson 1964), as well as the dual phonetic systems of bilinguals (e.g. Flege 1991; Grosjean and Miller 1994; MacLeod and Stoel-Gammon 2005; Chang et al. 2011), has demonstrated that the phonetic targets associated with any given category, such as [p^h] or [b], can differ in ways that anatomy alone could never explain. Some additional principle must restrict the variation in phonetic targets for a given individual when speaking a given language.

One version of the principle would require speakers to have the same (or highly similar) *patterns* of phonetic targets for sounds that bear a given phonological feature value. For example, the principle could require each talker's laryngeal target for [p^h] to be systematically related to the same talker's laryngeal target for [k^h], on the grounds that both sounds are specified [–voice] (or [+spread glottis]). A more stringent version of the principle requires the talker-specific phonetic target corresponding to a given phonological feature value to be identical or *uniform* across all sounds bearing that specification.[6] In this version, a talker cannot independently specify the properties of the laryngeal spreading gesture and associated timing relations for [p^h], [t^h], and [k^h]: the mapping from phonological feature to phonetic target must be the same, in these respects, for all three voiceless aspirated stop categories. (Observed variation in cue values for identically specified sounds must then arise "automatically" from independent differences in other targets, as discussed in Section 4 below.)

Within-category covariation was generally quite weak and we do not propose a separate principle of phonetic implementation to explain the dependencies that were found. The positive COG-VOT correlation observed for several stops may be attributable to aerodynamics: a higher airflow rate may give rise to an upwards shift in the energy distribution across the spectrum (Zue 1976; Koenig et al. 2013; Chodroff and Wilson 2014) and to longer aspiration durations. Notably, we found that this dependency held not only within stop categories but also when comparing the same two cues across them (e.g. for mean COG of [d] and mean VOT of [p^h], $r = 0.46$, $p < 0.001$). Perhaps each talker has a relatively higher or lower airflow rate across all stops (plausibly due to uniform realization of the shared feature [–continuant]), and this affects all COG and VOT values accordingly. A remaining question is why the correlations are particularly high within [d] and [g]. We speculate that the presence of voicing during stop closure for some talkers may lower COG and VOT. The weaker within-category correlation for [b] is likely due to the fact that the (positive) VOT of this stop does not vary as much across talkers.

3 Factor analysis

The strong between-category correlations documented above indicate that population-level variation in the phonetic realization of AE stops can be accurately modeled with a relatively small number of *latent values* for each talker. This idea could be formalized with many dimensionality reduction methods, including traditional principle component analysis (PCA; e.g. Pols et al. 1973; van Nierop et al. 1973) and more recently proposed eigenvoice decompositions (Kuhn et al. 1998). The model developed in this section is an instance of *factor analysis* (FA; e.g. Harshman et al. 1977; Clopper and Paolillo 2006; Leinonen 2008), a formally simple method that can express easily interpretable hypotheses about the content of the latent values.

In FA generally, each *observed* vector ($\boldsymbol{x_i}$) is modeled as drawn from a multivariate normal distribution with mean $\boldsymbol{Wz_i} + \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Psi}$. The *factor loading matrix* $\boldsymbol{W}$ represents a linear map from a latent vector ($\boldsymbol{z_i}$) into the observation space, where the dimensionality of $\boldsymbol{z_i}$ is smaller than that of $\boldsymbol{x_i}$. The offset vector $\boldsymbol{\mu}$ represents aspects of the mean that are, according to the model, the same across all individuals. Two

6 Either version of the principle must apply separately to each language / dialect / register that is controlled by a speaker. More generally, the principle should allow wide latitude in the contextual factors that can affect phonetic realization (e.g. prosody and speaking rate alongside language and register), requiring only that such factors have uniform effects on all sounds that are identically specified with respect to the relevant phonological feature.

additional restrictions are enforced: (a) $\boldsymbol{\Psi}$ is required to be diagonal, so that the components of an observed vector $\boldsymbol{x_i}$ are independent conditional on the latent vector $\boldsymbol{z_i}$, and (b) the distribution over latent vectors is a multivariate normal with zero mean and unit covariance, so that the components of $\boldsymbol{z_i}$ are standardized and independent from one another. In summary, for each individual $i = 1, \ldots, n$

$$p(\boldsymbol{x_i}) = \mathcal{N}(\boldsymbol{W z_i} + \boldsymbol{\mu},\ \boldsymbol{\Psi}),\ \text{where } \boldsymbol{\Psi} \text{ is diagonal}$$
$$p(\boldsymbol{z_i}) = \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I})$$

It follows that two components of the observed vector are predicted to covary only if they "load on" (have non-zero influence from) one or more common latent variables. In this way, factor analysis represents correlated variables in a higher-dimensional space with uncorrelated variables in a lower-dimensional space through a simple (i.e. linear) transformation.

FA provides a method for formalizing the hypothesis that, within the set of AE stops, the talker-specific contribution to phonetic targets is uniform for each phonological feature specification (as suggested in Section 2). To evaluate this hypothesis, we take each observed $\boldsymbol{x_i}$ to be the vector of stop cue means for a particular talker (i.e. $\boldsymbol{x_i} = [\mathrm{p}_i^{\mathrm{COG}}, \mathrm{b}_i^{\mathrm{COG}}, \mathrm{t}_i^{\mathrm{COG}}, \ldots, \mathrm{p}_i^{\mathrm{VOT}}, \ldots, \mathrm{p}_i^{\text{Onset-f0}}, \ldots, \mathrm{g}_i^{\text{Onset-f0}}]^T$, where $\boldsymbol{x_i^c}$ is the sample mean for cue c of stop x computed from the productions of the ith talker). The latent vector $\boldsymbol{z_i}$ represents the talker-specific contributions to the phonetic targets for all six stops, as reflected in the acoustic cues. Under the idealization that each cue reflects the talker-specific target contribution for one phonological feature, the uniformity principle implies a factor loading matrix $\boldsymbol{W}$ with the following highly sparse structure (where zero entries are left blank and the matrix is transposed for display purposes):

$$\boldsymbol{W}^T = \begin{array}{c} \\ lab \\ cor \\ dor \\ vcl \\ vcd \\ pitch \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccc} p^{COG} & b^{COG} & t^{COG} & d^{COG} & k^{COG} & g^{COG} & p^{VOT} & b^{VOT} & t^{VOT} & d^{VOT} & k^{VOT} & g^{VOT} & p^{f0} & b^{f0} & t^{f0} & d^{f0} & k^{f0} & g^{f0} \end{array} \\ \begin{pmatrix} w_{lab} & w_{lab} & & & & & & & & & & & & & & & & \\ & & w_{cor} & w_{cor} & & & & & & & & & & & & & & \\ & & & & w_{dor} & w_{dor} & & & & & & & & & & & & \\ & & & & & & w_{vcl} & & w_{vcl} & & w_{vcl} & & & & & & & \\ & & & & & & & w_{vcd} & & w_{vcd} & & w_{vcd} & & & & & & \\ & & & & & & & & & & & & w_p & w_p & w_p & w_p & w_p & w_p \end{pmatrix} \end{array}$$

For example, according to $\boldsymbol{W}$ the latent factor identified with the feature *vcl* (i.e. [–voice] or [+spread glottis]) has the same influence (w_{vcl}) on the VOT of the three voiceless stops, and zero effect on all other stop-cue combinations. Similar logic applies to the other feature-cue combinations (e.g. given the idealization that COG reflects only place features); note that "pitch" is an ad-hoc feature that predicts covariation of talker-specific f0 means across all vowels. In our implementation of the model, we in fact multiplied each row of $\boldsymbol{W}$ by the empirical standard deviation of its label (e.g. the first row was multiplied by the standard deviation of the talker means for COG of [p^{h}]). This induces a marginal distribution over x_i in which each w_k^2 is interpretable as a positive and pooled correlation coefficient.

The free parameters of this version of the FA model (i.e. the coefficients of $\boldsymbol{W}$, the offset $\boldsymbol{\mu}$, and the diagonal elements of $\boldsymbol{\Psi}$) were fit to the talker mean vectors measured earlier. As expected, the factor-loading coefficients indicated strong correlations among the stop-cue combinations that reflect a common feature value ($w_{lab}^2 = 0.60$, $w_{cor}^2 = 0.67$, $w_{dor}^2 = 0.69$, $w_{vcl}^2 = 0.72$, $w_{vcd}^2 = 0.20$, $w_{pitch}^2 = 0.91$; these values should be compared to the correlations in Table 2). The values in the offset $\boldsymbol{\mu}$ account for talker-general differences in the values of stop-cue combinations that are otherwise unexpected given $\boldsymbol{W}$; for example, the offset for the VOT of [b] ($\boldsymbol{\mu} = 8.42$) is lower than that for the VOT of [g] ($\boldsymbol{\mu} = 16.86$). One interpretation of such offset differences is that they reflect "automatic" articulatory and acoustic effects – influences on the measurements that would be present even if the underlying phonetic targets studied here were exactly uniform within a talker. For example, effects of place on stop closure duration could contribute to differences in VOT values even if laryngeal targets and their timing with respect to supralaryngeal gestures are uniform (e.g. Weismer 1980; Maddieson 1997).

We compared the FA model above (the *target uniformity* model) with several alternatives that differed in the factor loading matrix: a *null covariation* model, in which $\boldsymbol{W}$ was the diagonal matrix; a sample of 500

Table 4: Marginal negative log-likelihood for observed talker mean vectors under different versions of the factor analysis model.

Model	Negative log-likelihood
Target uniformity	14,836.99
Null (diagonal)	16,694.72
Row permutation	16,438.98 (range: 15,696.90–16,583.83)
Exploratory	14,419.77

row permutation models derived by randomly exchanging the rows of the target uniformity model; and an *exploratory* model, in which the factor loading matrix had six columns with all cell values fit to the data. Table 4 reports the marginal negative log-likelihood values of the talker mean vectors for each model. The target uniformity model provided a significant improvement over the null model and all of the row-permutation variants, while the exploratory model was superior to target uniformity (similar results were obtained with other model comparison measures and with cross-validation). These results suggest that target uniformity is an important (but unlikely the only) principle of phonetic implementation that constrains the covariation of stop consonants within talkers.[7]

4 Covariation and predictability in perceptual adaptation

Patterns of phonetic covariation such as the one observed above have important implications not only for the mapping between phonology and phonetics, but also for their potential role in perceptual adaptation. On the basis of strong between-category covariation, listeners could reasonably predict a talker-specific target for one sound category after hearing productions of only one or more covarying categories. The findings of many studies of perceptual generalization are consistent with this idea. For instance, listeners generalize talker-specific spectral characteristics from exposure vowels to previously unheard vowels (e.g. Ladefoged and Broadbent 1957; Maye et al. 2008; Chládková et al. 2017) and have been shown to extrapolate a talker's characteristic VOT from [p^h] to [k^h] based on direct evidence about [p^h] only (e.g. Theodore and Miller 2010; Nielsen 2011). A listener may have low prior expectations for within-category covariation, but could infer talker-specific relations among cues through distributional learning (e.g. Clayards et al. 2008).

The FA model presented above encodes covariation with a lower-dimensional set of latent variables. If listeners attempted to infer the latent factor values of a novel talker, they would generalize across segments with shared phonological feature specifications in a way that is consistent with the population-level correlations. In this sense, the FA model could be interpreted as a cognitive model of adaptation. Between-category covariation has been incorporated to varying degrees in previous models of adaptation. Models that employ variants of mean subtraction or *z*-scoring within each phonetic dimension implicitly enforce covariation, provided the calculation incorporates values from several speech sounds (e.g. Sliding Template Model of Vowel Normalization: Nearey and Assmann 2007; c-CuRE: McMurray and Jongman 2011; VOT generalization: Nielsen and Wilson 2008). These models, however, have posited that each talker has a single "offset" value per cue and thus assume *perfect* covariation among *all* speech sounds represented on a given

7 Two columns of the loading matrix in the exploratory model were quite similar to columns in the theoretically-determined matrix $\boldsymbol{W}$. The first had large values only for the VOT means of the voiceless stops, closely emulating uniform realization of [–voice] or [+spread glottis]. The second had values near unity for all of the f0 means, and much smaller values elsewhere, in line with uniformity with respect to the "pitch" feature. A third column seemed to combine the place effects of $\boldsymbol{W}$, with particularly large values for the COG mean of [k^h] and [g], intermediate values for the other COG means, and smaller values for all other means. Two of the remaining columns appear to express correlations among COG and VOT for [d] and [g] separately; these within-category relations were found in our statistical analysis (see Section 2) but do not follow from the uniformity principle. The final column was generally difficult to interpret but assigned a particularly large value to the VOT mean of [b], the stop most likely to have closure voicing.

acoustic-phonetic dimension. To the extent that the empirical covariation is weak for some of the relevant sounds (e.g. as observed for VOT between stops contrasting in voice), this strong assumption could lead to suboptimal performance. Alternative models of talker adaptation, such as exemplar models or the ideal adaptor model, do not currently encode covariation of phonetic properties, and may therefore fail to model aspects of perceptual generalization across speech sounds (e.g. Johnson 1997; Kleinschmidt and Jaeger 2015). In comparison, FA can model selective, and ideally theoretically-grounded, patterns of between-category covariation, as opposed to assuming perfect covariation or omitting covariation altogether. Future research should be directed towards understanding the relation between measured phonetic covariation and patterns of perceptual generalization by human listeners.

5 Conclusion

The analyses of talker means for COG, VOT, and onset f0 within and among stop categories revealed greater between-category covariation in comparison to within-category covariation. As examined in Section 3, the observed covariation among phonetic categories may arise from a constraint of uniformity on the mapping from phonological features to phonetic targets underlying acoustic-phonetic properties. Further research is required to evaluate the predictions of uniformity as it applies to other segments and languages. In addition, perceptual knowledge of covariation could facilitate prediction in perceptual processing, and more generally, the measured covariation serves as a testable hypothesis of perceptual knowledge in generalized adaptation in speech perception.

Appendix

The analyses in Section 2 involved correlations of talker means; however, many previous studies have also examined correlations across individual tokens (e.g. Dmitrieva et al. 2015; Kirby and Ladd 2015, Kirby and Ladd 2016; Clayards 2018). For comparison with these studies, token-by-token correlations between phonetic cues were calculated for each stop category within and across talkers. Only stop consonants with non-outlier values for both cues were retained for these correlations. There were 71,852 stops for the COG-VOT analysis, 57,737 stops for the COG-f0 analysis, and 74,916 stops for the VOT-f0 analysis. The first correlation analysis, reported in Table A1, was conducted across all tokens (see also Dmitrieva et al. 2015; Clayards 2018). These correlations largely resembled the correlations of talker means in magnitude (especially between COG and VOT for [b], [d], and [g]); while many of these correlations reached significance, they were nevertheless quite weak. In the second analysis, correlations were limited to talkers with more than 20 tokens per stop category. The median number of talkers excluded from each analysis was four and the maximum was 55 talkers (between COG and f0 for [t^h]). Table A2 presents the median token-by-token correlation for each of the cue pairs and stop consonants, as well as the range across talkers. Consistent with findings in Kirby and Ladd (2016) for French and Italian intervocalic stops, the magnitude and direction of the by-speaker correlations varied substantially

Table A1: Token-by-token correlations for each cue pair and stop category aggregated over all talkers.

	COG-VOT	COG-f0 (female)	COG-f0 (male)	VOT-f0 (female)	VOT-f0 (male)
p^h	0.18*	−0.01	0.09*	−0.05*	−0.02
b	0.34*	0.00	−0.05*	−0.03	−0.01
t^h	0.09*	0.07*	0.12*	−0.11*	−0.06*
d	0.57*	−0.11*	−0.06*	−0.06*	−0.01
k^h	0.17*	0.10*	0.09*	−0.15*	−0.13*
g	0.52*	0.03	−0.01	−0.03	0.01

An asterisk reflects $p < 0.001$.

Table A2: For each stop category and cue pair separately, the median talker-specific token-by-token correlation (left column) and range of talker-specific token-by-token correlations (right column).

	COG-VOT		COG-f0		VOT-f0	
	Median	**Range**	**Median**	**Range**	**Median**	**Range**
p^h	0.17	−0.41 to 0.62	0.09	−0.41 to 0.44	−0.04	−0.47 to 0.54
b	0.33	−0.14 to 0.77	−0.01	−0.53 to 0.64	0.00	−0.37 to 0.41
t^h	0.06	−0.46 to 0.59	0.10	−0.34 to 0.51	−0.16	−0.59 to 0.41
d	0.54	−0.18 to 0.75	−0.06	−0.50 to 0.45	−0.03	−0.35 to 0.42
k^h	0.16	−0.31 to 0.57	0.10	−0.34 to 0.52	−0.20	−0.61 to 0.39
g	0.54	−0.04 to 0.79	0.01	−0.51 to 0.39	0.00	−0.38 to 0.33

across talkers. Together, these findings indicate that, while there may exist weak relationships across talker means, the token-by-token relationships within talker-specific productions are highly variable.

References

Assmann, P. F., T. M. Nearey & S. Bharadwaj. 2008. Analysis of a vowel database. *Canadian Acoustics* 36(3). 148–149.

Boersma, P. & D. Weenink. 2016. Praat: Doing phonetics by computer [Computer program]. Version 6.0.19, retrieved from http://www.praat.org/.

Brandschain, L., D. Graff, C. Cieri, K. Walker & C. Caruso. 2010. The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2441–2444. Malta: European Language Resources Association (ELRA).

Brandschain, L., D. Graff & K. Walker. 2013. *Mixer 6 Speech LDC2013S03*. Hard Drive. Philadelphia: Linguistic Data Consortium.

Chang, C. B., Y. Yao, E. F. Haynes & R. Rhodes. 2011. Production of phonetic and phonological contrast by heritage speakers of Mandarin. *The Journal of the Acoustical Society of America* 129(6). 3964–3980.

Chládková, K., V. J. Podlipský & A. Chionidou. 2017. Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance* 43(2). 414–427.

Chodroff, E. 2017. *Structured variation in obstruent production and perception*. Baltimore, MD: Johns Hopkins University dissertation.

Chodroff, E., M. Maciejewski, J. Trmal, S. Khudanpur & J. J. Godfrey. 2016. New release of Mixer-6: Improved validity for phonetic study of speaker variation and identification. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariana, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1323–1327. Portorož, Slovenia: European Language Resources Association (ELRA).

Chodroff, E. & C. Wilson. 2014. Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America* 136(5). 2762–2772.

Chodroff, E. & C. Wilson. 2017. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics* 61. 30–47.

Clayards, M. A. 2018. Individual talker and token covariation in production of multiple cues to stop voicing. *Phonetica* 75(1). 1–23.

Clayards, M. A., M. K. Tanenhaus, R. N. Aslin & R. A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108(3). 804–809.

Clopper, C. G. & J. C. Paolillo. 2006. North American English vowels: A factor-analytic perspective. *Literary and Linguistic Computing* 21(4). 445–462.

DiCanio, C. T., H. Nam, J. D. Amith, R. C. García & D. H. Whalen. 2015. Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. *Journal of Phonetics* 48. 45–59.

Dmitrieva, O., F. Llanos, A. A. Shultz & A. L. Francis. 2015. Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics* 49. 77–95.

Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82(397). 171–185.

Evans, J. W. 1996. *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Flege, J. E. 1991. Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *The Journal of the Acoustical Society of America* 89(1). 395–411.

Flemming, E. S. 2007. Stop place contrasts before liquids. In J. Trouvain & W. Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 233–236. Saarbrücken, Germany: Saarland University.

Forrest, K., G. Weismer, P. Milenkovic & R. N. Dougall. 1988. Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America* 84(1). 115–123.

Foulkes, P. & G. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34. 409–438.

Foulkes, P., G. Docherty & D. Watt. 2001. On the emergence of structured phonological variation. *University of Pennsylvania Working Papers in Linguistics* 7(3). 67–84.

Fruehwald, J. 2013. *The phonological influence on phonetic change*. Philadelphia, PA: University of Pennsylvania dissertation.

Fruehwald, J. 2017. The role of phonology in phonetic change. *Annual Review of Linguistics* 3. 25–42.

Grosjean, F. & J. L. Miller. 1994. Going in and out of languages: An example of bilingual flexibility. *Psychological Science* 5(4). 201–207.

Guy, G. R. & F. Hinskens. 2016. Linguistic coherence: Systems, repertoires and speech communities. *Lingua* 172–173. 1–9.

Haggard, M. P., S. Ambler & M. Callow. 1970. Pitch as a voicing cue. *The Journal of the Acoustical Society of America* 47(2, Part 2). 613–617.

Hanson, H. M. & K. N. Stevens. 2003. Models of aspirated stops in English. In M. Solé, D. Recasen & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*, 783–786. Barcelona, Spain: Universitat Autònoma de Barcelona.

Harshman, R., P. Ladefoged & L. M. Goldstein. 1977. Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America* 62(3). 693–707.

Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (eds.), *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Joos, M. 1948. Acoustic phonetics. *Language* 24(2). 5–136.

Keshet, J., M. Sonderegger & T. Knowles. 2014. AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]. Version 0.91, retrieved August 2016 from https://github.com/mlml/autovot/.

Kirby, J. P. & D. R. Ladd. 2015. Stop voicing and f0 perturbations: Evidence from French and Italian. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Paper number 0740. Glasgow, UK: University of Glasgow.

Kirby, J. P. & D. R. Ladd. 2016. Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *The Journal of Acoustical Society of America* 140(4). 2400–2411.

Kleinschmidt, D. F. & T. F. Jaeger. 2015. Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review* 122(2). 148–203.

Koenig, L. L. 2000. Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language, and Hearing Research* 43(5). 1211–1228.

Koenig, L. L., C. H. Shadle, J. L. Preston & C. R. Mooshammer. 2013. Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research* 56(4). 1175–1189.

Kuhn, R., P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, N. Field & M. Contolini. 1998. Eigenvoices for speaker adaptation. In R. H. Mannell & J. Robert-Ribes (eds.), *Proceedings of the 5th International Conference on Spoken Language Processing*, 1774–1777. Sydney, Australia: Australian Speech Science and Technology Association, Incorporated (ASSTA).

Labov, W. 1966. *The social stratification of English in New York City*, 2nd edn. New York: Cambridge University Press.

Ladefoged, P. & D. E. Broadbent. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29(1). 98–104.

Leinonen, T. 2008. Factor analysis of vowel pronunciation in Swedish dialects. *International Journal of Humanities and Arts Computing* 2(1–2). 189–204.

Lindblom, B. 1967. Vowel duration and a model of lip-mandible coordination. *Speech Transmission Laboratory – Quarterly Progress and Status Reports* 8(4). 1–29.

Lisker, L. & A. S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3). 384–422.

MacLeod, A. & C. Stoel-Gammon. 2005. Are bilinguals different? What VOT tells us about simultaneous bilinguals. *Journal of Multilingual Communication Disorders* 3(2). 118–127.

Maddieson, I. 1997. Phonetic universals. In J. Laver & W. J. Hardcastle (eds.), *Handbook of phonetic sciences*, 619–639. Oxford: Blackwells Publishers.

Maye, J., R. N. Aslin & M. K. Tanenhaus. 2008. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science* 32(3). 543–562.

McMurray, B. & A. Jongman. 2011. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review* 118(2). 219–246.

Nearey, T. M. 1978. *Phonetic feature system for vowels*. Edmonton, Alberta: University of Alberta dissertation.

Nearey, T. M. 1989. Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America* 85(5). 2088–2113.

Nearey, T. M. & P. F. Assmann. 2007. Probabilistic "sliding template" models for indirect vowel normalization. In M.-J. Solé, P. S. Beddor & M. Ohala (eds.), *Experimental approaches to phonology*, 246–269. New York: Oxford University Press.

Newman, R. S., S. A. Clouse & J. L. Burnham. 2001. The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America* 109(3). 1181–1196.
Nielsen, K. Y. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39. 132–142.
Nielsen, K. Y. & C. Wilson. 2008. A hierarchical Bayesian model of multi-level phonetic imitation. In N. Abner & J. Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 335–343. Los Angeles: Cascadilla Proceedings Project.
Ohde, R. N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America* 75(1). 224–230.
Pols, L. C. W., H. R. C. Tromp & R. Plomp. 1973. Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America* 53(4). 1093–1101.
Rose, P. 2010. The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics. *International Journal of Biometrics* 2(4). 316–329.
Shultz, A. A., A. L. Francis & F. Llanos. 2012. Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America* 132(2). EL95.
Smiljanić, R. & A. R. Bradlow. 2008. Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics* 36(1). 91–113.
Solé, M.-J. 2007. Controlled and mechanical properties in speech. In M.-J. Solé, P. S. Beddor & M. Ohala (eds.), *Experimental approaches to phonology*, 302–321. Oxford: Oxford University Press.
Sonderegger, M., M. Bane & P. Graff. 2017. The medium-term dynamics of accents on reality television. *Language* 93(3). 598–640.
Theodore, R. M. & J. L. Miller. 2010. Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America* 128(4). 2090–2099.
Theodore, R. M., J. L. Miller & D. DeSteno. 2009. Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America* 125(6). 3974–3982.
Titze, I. R. 2011. Vocal fold mass is not a useful quantity for describing F0 in vocalization. *Journal of Speech and Hearing Research* 54(2). 520–522.
Toivonen, I., L. Blumenfeld, A. Gormley, L. Hoiting, N. Ramlakhan & A. Stone. 2015. Vowel height and duration. In U. Steindl, T. Borer, H. Fang, A. Garcia Pardo, P. Guekguezian, B. Hsu, C. O'Hara & I. C. Ouyang (eds.), *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, 64–71. Somerville, MA: Cascadilla Proceedings Project.
van Nierop, D. J. P. J., L. C. W. Pols & R. Plomp. 1973. Frequency analysis of Dutch vowels from 25 female speakers. *Acustica* 29(2). 110–118.
Weismer, G. 1980. Control of the voicing distinction for intervocalic stops and fricatives: some data and theoretical considerations. *Journal of Phonetics* 8. 427–438.
Whalen, D. H. & A. G. Levitt. 1995. The universality of intrinsic f0 of vowels. *Journal of Phonetics* 23. 349–366.
Yuan, J. & M. Y. Liberman. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*. 5687–5790. Paris: Société Française d'Acoustique (SFA).
Zlatin, M. A. 1974. Voicing contrast: Perceptual and productive voice onset time characteristics of adults. *The Journal of the Acoustical Society of America* 56(3). 981–994.
Zue, V. W. 1976. *Acoustic characteristics of stop consonants: A controlled study*. Cambridge, MA: Massachusetts Institute of Technology dissertation.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2017-0047).