Detecting articulatory compensation in acoustic data through linear regression modeling

Alina Khasanova[1], *Jennifer Cole*[2], *Mark Hasegawa-Johnson*[3]

[1]unaffiliated
[2]Department of Linguistics, University of Illinois, Urbana-Champaign, U.S.A.
[3]Department of Computer and Electrical Engineering, University of Illinois, Urbana-Champaign, U.S.A.
akhasanova@nawaz.org, jscole@illinois.edu, jhasegaw@gmail.com

Abstract

Examining articulatory compensation has been important in understanding how the speech production system is organized, and how it relates to the acoustic and ultimately phonological levels. This paper offers a method that detects articulatory compensation in the acoustic signal, which is based on linear regression modeling of co-variation patterns between acoustic cues. We demonstrate the method on selected acoustic cues for spontaneously produced American English stop consonants. Compensatory patterns of cue variation were observed for voiced stops in some cue pairs, while uniform patterns of cue variation were found for stops as a function of place of articulation or position in the word. Overall, the results suggest that this method can be useful for observing articulatory strategies indirectly from acoustic data and testing hypotheses about the conditions under which articulatory compensation is most likely.

Index Terms: articulatory compensation, acoustic cues, linear regression modeling

1. Introduction

The articulatory processes underlying acoustic cues which are then interpreted to a linguistically meaningful unit (like a phoneme or a syllable), can be thought of as a series of interdependent events. For example, in the production of an English voiceless plosive, vocal folds are abducted, a constriction is formed, and then released, with coordination among the gestures in this sequence. Yet, this interdependence is not complete: there is some degree of freedom in the production of individual speech gestures, which speakers can employ to achieve particular linguistic goals. We describe two articulatory gestures as being in a trading relationship if impairment in the production of one gesture is compensated for in the production of the other gesture, with the overall result being that the intended phonological category is successfully cued. This process has been demonstrated in the production of English /u/ in both natural [1] and perturbed [2] speech, the production of /r/ in American English [3, 4], and the production of the German palatal sibilant using an artificial palate [5]. The projected result of compensatory articulation is generally understood as constraining acoustic variability in the cues to phonological categories, such as voicing or place of articulation.

An important question is under which circumstances speakers are likely to resort to articulatory compensation. A suggestion has been made that, in naturally produced speech, compensation is most likely to take place in the production of phones near a perceptual boundary [1]. Another possible reason for speakers to resort to compensation is the demanding nature of production of some phonemes, like fricatives and trills. In principle then, there can be multiple incentives for articulatory compensation, grounded in either production or perception.

Since almost all attention has concentrated on locating compensation in the articulatory domain, the possibility of observing it indirectly in the acoustic signal has been relatively unexplored. This paper offers a method for detecting articulatory compensation in acoustic data.

Our approach rests on the observation that acoustic cues to a phonological contrastive feature can enter into a specific relationship, based, presumably, on the interaction of production events of similar kind. Suppose that two acoustic measures are known to be cues to the same phonological feature. These cues may be either postively or negatively correlated with one another in their values. Positive correlation occurs when cues are subject to either concomitant weakening or strengthening; this kind of cue co-variation has been shown in clear speech experiments [6, 7], various reductive processes [8, 9], as well as prosodic studies [10, 11].

The other possibility is that, if one cue is weakened, such that it affords a smaller distinction between phonologically contrastive sounds, another cue to the same contrastive category may be strengthened to compensate for this. Such inverse correlation may correspond to the compensatory strategy demonstrated in the articulatory domain. Although it is of course not possible to establish a direct link between compensation in the articulatory and acoustic domain by relying on acoustic data alone, we hypothesize that compensation in acoustic cues is the result of trading relations in production. To our best knowledge, this relationship has not been addressed before.

We illustrate the method by considering co-variation patterns of a small set of acoustic cues characterizing the production of American English stop consonants. The goal is to demonstrate compensation in acoustic cues, as well as to shed light on which categories of stops are more likely to exhibit it. Specifically, we examine how voicing, place of articulation (POA), and position in the word modify co-variation patterns. Below we briefly consider which of these are likely to promote a compensatory strategy.

With respect to voicing, it has been argued that that production of voiced stops is more effortful than that of voiceless stops [12, 13]. Acoustically, voiced stops are characterized by cues that are diminished relative to those of voiceless counterparts: the duration of both closure and burst are shorter, which have been argued to be the result of reduced intra-oral pressure in voiced stops [14]. Although shorter duration of closure and release serve as secondary cues distinguishing voiced stops from voiceless ones, it may also jeopardize signaling the manner of the stops, as well as extraction of POA from the spectrum. Therefore there are potential benefits to compensating for shortened closure and burst cues in the production of voiced stops.

With respect to POA, it is difficult to make a cogent argument that one POA is more likely to induce compensatory effort than others. In the production of velars, the smallest volume of the intra-oral cavity and subsequently the quickest rise of pressure can increase the difficulty of maintaining a tight constriction. Yet acoustically, velar stops are not considered to have weak cues, with a possible exception of voicing: velars have the burst of longest duration and highest intensity [15], in addition to often featuring bursts with multiple transients. These characteristics likely contribute to velars being the most perceptually salient stops [16]. Labial stops are often mentioned as having weak acoustic cues. Despite having long closures, they feature short bursts of low amplitude, which is attributed to the location of the labial constriction. Due to the constriction being at the lips, there is effectively no post-constriction cavity and thus no obstacles that can increase the intensity of noise at the release or prolong its duration. Despite some definite features which suggest that labial stops possess weaker acoustic cues, their perceptual salience is relatively high (below velars but above coronals) [16]. As to coronals, from the production standpoint, they are intermediate between velars and labials in terms of both the pre- and post-constriction cavity sizes, so they avoid the problems that are connected therewith; on the other hand, in American English in certain phonetic contexts the coronal gesture is reduced [17]. Perceptually, coronal stops are characterized by intermediate values of both durational and amplitudinal cues, as compared to velars and labials, so this does not outright establish a basis for compensation. Nevertheless, they may be considered vulnerable given that they are the most confusable stop cross-linguistically [16]. Thus each POA has some advantages and disadvantages in its production and perception, making it unclear whether compensatory effort will be manifest for any of the stops by POA category.

As to position in the word, initial position is an environment that has been associated with segmental strengthening [18, 19] while word-final position is often observed to condition weakening effects, such as word-final devoicing in many Slavic languages or unreleased variants of stops in American English. The word-medial intervocalic position, having the strongest acoustic cues due to the availability of formant transitions in and out of the consonant, as well as the release burst of the target segment, has been implicated in gestural undershoot of the constriction in certain conditions (as in Spanish spirantization). So, final stops and, to a lesser extent, word-medial intervocalic stops do exhibit greater reduction than initial stops. However, it is not clear whether level of reduction is sufficient grounds for compensation, since they may in fact be useful in conveying prosodic information.

2. Methods

2.1. Database

The data are drawn from the Buckeye corpus of spontaneous American English [20], and consist of [p,t,k,b,d,g] in initial prevocalic, word-medial intervocalic, and post-vocalic final positions. Tokens identified in the corpus phonetic transcription as glottalized or flapped were excluded. Since the boundary between the closure and burst is not marked in the Buckeye corpus, it was automatically located by an algorithm we developed, which relies on finding the point of the largest energy increase corresponding to the initiation of the burst. Tokens where the automatic procedure failed to locate the burst were excluded from the analysis. The resulting number of tokens is 50335. The following acoustic cues were measured: closure duration, burst duration, c0ave of closure (the average of the zeroth cepstral coefficient over the closure interval, a measure of average spectral level), and c0ave of burst.

2.2. Statistical Analyses

To evaluate the relationship of a pair of acoustic cues, a linear mixed-effects multiple regression model is used, with one of the cues entered as the response variable while the other as the predictor variable. Additionally, categorical variables of voicing, POA, and position in the word were entered as well, in order to assess whether and how they modify the basic relationship among acoustic cues. The continuous and categorical variables are entered as the fixed effects, while subject is entered as a random effect. The process of model building starts with graphical data exploration, whereby raw data scatterplots are overlaid with linear regression lines (based on the subsets of data determined by categorical variables) and locally weighted scatterplot smoothers (loess), in order to assess the need for inclusion of categorical variables or higher order terms in the model. Following best practices developed in the statistical literature [21], model building includes determination of the optimal random and fixed components, as well as model fit evaluation through visual examination of residuals and normal probability plots. We select the model with the smallest Akaike infomation criterion (AIC) and Bayesian information criterion (BIC), and the largest log-likelihood statistics. Our data were often heteroscedastic, which was dealt with by specifying a form of within-subject variance other than a constant [22]. Even in those cases where inclusion of quadratic or cubic terms improved the fit of the models, first-order linear models were also fit in order to take into account the contributions of interaction terms and calculate the slope estimates for categorically-determined subsets of data (sometimes referred to as simple slope analysis), which are tested for being significantly different from zero.

2.3. Interpretation of regression slopes

To determine whether a pair of acoustic cues are in uniform or compensatory relationship, one has to establish what constitutes a strengthening or a weakening of a given cue. For acoustic cues used here, the duration intervals of closure and burst are considered more optimal when longer; the burst c0ave when it is higher; and the closure c0ave when it is lower. Thus the word ”optimal,” as used here, means ”optimally communicates manner of articulation.” The relationship of this form of optimality to the phonological features of voicing and POA is one of the objects of study.

The interpretation of the sign of regression slopes is done according to templates of cue co-variation, depending on their optimal ranges. In the pattern of uniform cue co-variation, two cues move jointly in the same direction of optimality of phonetic realization.

In the compensatory cue relationship, conversely, when one of the cues is sub-optimal, the other will be realized at an increased level of optimality. The templates in Fig. 1 represent two stylized examples of compensatory cue co-variation. The axis for a given cue is divided into two parts, one of which corresponds to its optimal end of the range and the other to the suboptimal one. The 2-dimensional space defined by C1 and

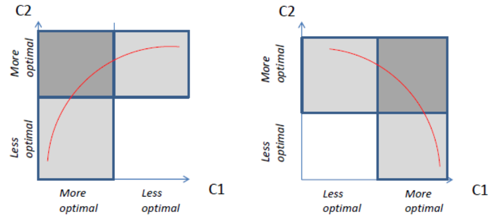

Figure 1: The compensatory relationships between cues C1 and C2 when C1 is more optimal and C2 is less optimal at the lower end of their ranges (left) or when both cues are less optimal at the lower end of their ranges (right).

C2 is divided into 4 quadrants, the optimality value of which is jointly determined by the values of optimality of C1 and C2. If both C1 and C2 are optimal, then this will define the maximally optimal quadrant (darker grey in the graphs), while if both are sub-optimal, then this quadrant is maximally sub-optimal (uncolored). The regions where one of the characteristics is sub-optimal and the other is optimal take on an intermediate value (lighter grey). If compensation takes place, one expects the data points to fall within the regions with maximal or intermediate optimality, and avoid the maximally sub-optimal quadrant. A stylized curve superimposed over the shaded regions represents a compensatory relationship between C1 and C2. Though the best-fit models are often nonlinear (as schematized in Fig. 1), tests for significance of correlation are based on a linear regression model.

3. Results

Regression modeling in six cue pairs (based on the four acoustic cues) was implemented. The majority of models with the best fit were second- or third-order due to some non-linearities in the data, most frequently exhibitted at the extreme values of cues; models almost always required non-constant variance functions which lessened the problem of heteroscedasticity in the data; the interaction of the continuous predictor variable with the categorical co-variates was frequently highly significant. One exception to these general trends was the relationship between closure c0ave and burst c0ave: the best fit model was first-order, the data were homoscedastic, and there were fewer categorical difference between slope magnitudes. It appears that the positive association between these two cues may not be reflective of the underlying articulatory events (of compensatory nature), but rather of the congruence in energy levels for adjacent speech signal portions. For this reason, this pair is not considered in the overall picture of results presented below.

3.1. Voiced Stops Demonstrate Compensatory Correlation

Closure duration and burst duration are correlated for [t,k,b,g] and for initial and final [p]. The slopes for voiceless stops are positive, with most being highly significant (the exception to this is medial [p]), i.e., longer closure durations are associated with longer burst durations. On the other hand, the slopes for voiced stops are negative and significant for [b] and [g], but not [d] (Table 1): longer closure durations are associated with shorter burst durations. Since both closure and burst are more perceptually salient when longer rather than shorter, the negative association may be hypothesized to be the result of compensatory coarticulation.

Table 1: Slope estimates for categorically conditioned data subsets: burst duration versus closure duration.

	Initial	Medial	Final
p	0.089***	0.03	0.052**
t	0.265***	0.206***	0.228***
k	0.189***	0.13***	0.152***
b	-0.145***	-0.204***	-0.182***
d	0.031*	-0.028	-0.006
g	-0.045**	-0.104***	-0.082***

*** sig. at p<0.001,** sig. at p<0.01,* sig. at p<0.05.

Burst c0ave and burst duration are correlated for all stops except medial [k] (Table 2). The slopes for voiced stops and [p] are negative and highly significant, suggesting a compensatory relationship between the cues. Slopes are significantly positive for [t], suggesting uniform cue association.

Table 2: Slope estimates for categorically conditioned data subsets: burst c0ave versus burst duration.

	Initial	Medial	Final
p	-21.29***	-14.3*	-22.22***
t	16.14***	23.13***	15.22***
k	-9.61*	-2.62	-10.54*
b	-136.72***	-129.73***	-137.65***
d	-99.29***	-92.3***	-100.21***
g	-125.04***	-118.05***	-125.97***

*** sig. at p<0.001,** sig. at p<0.01,* sig. at p<0.05.

3.2. Uniform Relationship: POA and Position

The pattern of uniform cue association for stops of all categorical attributes was prevalent in the remaining three pairs: closure duration vs burst c0ave, closure duration vs closure c0ave, and closure c0ave vs burst duration.

The results of the simple slope analysis for burst c0ave against closure duration are presented in Table 3. All values are positive and highly significant. Closure duration and burst c0ave are both optimal/sub-optimal at the same ends of their ranges, so the positive assiciation is interpreted as uniform cue co-variation.

Table 3: Slope estimates for categorically conditioned data subsets: burst c0ave versus closure duration.

	Initial	Medial	Final
p	35.81***	55.17***	57.23***
t	38.62***	57.98***	60.04***
k	32.71***	52.07***	54.13***
b	47.14***	66.5***	68.56***
d	49.96***	69.31***	71.37***
g	44.04***	63.4***	65.46***

*** sig. at p<0.001,** sig. at p<0.01,* sig. at p<0.05.

The regression slopes for burst duration against closure c0ave are given in Table 4. The majority of values are negative and highly significant, although very small in magnitude. The exceptions to this trend are the following: the final [b] and [d] are positive (statistically significant at 0.001 level), while the slopes of final [g] and initial [b] are not significantly different from zero. In the model of closure c0ave against closure duration, the interaction with voicing was found to be not significant. The values of the slope estimates for subsets conditioned by POA and position are given in Table 5. All the slope values are negative and highly significant. In both cases, the negative slope signifies uniform cue association, since manner of articulation is optimally communicated by small closure c0ave, large burst c0ave, and large burst duration.

Table 4: Slope estimates for categorically conditioned data subsets: burst duration versus closure c0ave.

	Initial	Medial	Final
p	-0.0005***	-0.0006***	-0.0001*
t	-0.0005***	-0.0007***	-0.0002***
k	-0.0008***	-0.0009***	-0.0004***
b	-7.3e-05	-0.0002**	0.0003***
d	-0.0001**	-0.0003***	0.0002***
g	-0.0004***	-0.0005***	-1.3e-05

*** sig. at p<0.001,** sig. at p<0.01,* sig. at p<0.05.

Table 5: Slope estimates for categorically conditioned data subsets: closure c0ave versus closure duration.

	Initial	Medial	Final
Labial	-27.12***	-32.09***	-39.24***
Coronal	-50.61***	-55.57***	-62.73***
Velar	-44.49***	-49.46***	-56.62***

*** sig. at p<0.001,** sig. at p<0.01,* sig. at p<0.05.

4. Discussion and Conclusions

Voiced stops exhibited the clearest cases of compensation in acoustic cues. This is consistent with the prediction that they are more likely to promote articulatory compensation, due to their more effortful articulation and less perceptually salient cues, and relative proximity to a perceptual boundary (voiced stops are more likely to be misperceived as approximants than are unvoiced stops). Among unvoiced stops, labials exhibited compensatory correlation of burst c0ave versus burst duration, but not of burst duration versus closure duration. Word position interacted significantly with correlation slope, but the direction of interaction was not consistent across cues or across phonemes.

Even in cases where no compensatory relationship between cues was found, there still may be some subtle articulatory adjustments which are essentially compensatory in nature. Two features in the present results suggest that this may be the case. First is the frequent difference in magnitude of regression slopes between categorical subsets, even when the slopes' direction is uniform. The question arises whether these differences are due to the presense of active articulatory adjustments by speakers that would modify the underlying interaction of articulatory events (a sort of weak compensatory articulations), or, on the other hand, lack thereof, in which case the slope magnitude differences reflect the natural effort required to produce these categories of phones. The second feature that is of additional interest is the non-linearities that were always observed in uniform relationships. Those occurred most frequently near the endpoints of acoustic cue ranges, and may be due to some compensatory articulation on smaller subsets of data, not encompassing the entire categories of phones.

This research has demonstrated that examination of acoustic data can be revealing as to which articulatory strategies are employed by speakers and under what circumstances. Whether speakers engage in compensatory or uniform articulation is apparently influenced by an exigent need to maintain a certain level of optimality of acoustic cues, specifically in cases when there are definite disadvantages at both production and perception levels. Articulatory adjustments of compensatory nature may be necessitated by a variety of reasons, for example, different speaking styles (careful versus casual), speaker-specific anatomic impediments, different degrees of difficulty in production of individual phonemes or of their perceptual salience. Although the example explored in this paper is of the latter kind, in principle the method can be extended to other sources of compensatory articulatory adjustments.

The attractiveness of the method lies in its relative ease and in the availability of a large amount of acoustic data (in comparison to articulatory data). Amongst its limitations is the difficulty of interpreting the results, which in large part depends on how well-grounded the hypotheses are about the relationship of production to acoustic cues and the optimal cue values. Despite these difficulties, the analyses presented in this work can be useful as either confirmatory or exploratory evidence for articulatory strategies in spontaneous speech acoustics.

5. References

[1] Perkell, J.S., Matthies, M., Svirsky, M.A., and Jordan, M.I., “Goal-based speech motor control: A theoretical framework and some preliminary data”, Journal of Phonetics, 23:23-35, 1995.

[2] Brunner, J., Hoole, P. and Perrier, P., “Motor equivalent strategies in the production of /u/ in perturbed speech”, The Journal of the Acoustical Society of America, 123(5):3076, 2008.

[3] Guenther, F.H., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L, Zandipour, M. and Perkell, J.S., “Articulatory tradeoffs reduce acoustic variability during American English /r/ production”, The Journal of the Acoustical Society of America, 105(5):2854-65, 1999.

[4] Nieto-Castanon, A., Guenther, F.H., Perkell, J.S. and Curtin, H.D., “A modeling investigation of articulatory variability and acoustic stability during American English /r/ production”, The Journal of the Acoustical Society of America, 117(5):3196, 2005.

[5] Brunner, J. and Hoole, P., “Motor equivalent strategies in the produciton of German /ʃ/ under perturbation”, Language and Speech, Feb. 2012.

[6] Ferguson, S. and Kewley-Port, D., “Talker differences in clear and conversational speech: Acoustic characteristics of vowels”, J. Speech, Language and Hearing Research, 50:1241-1255. 2007.

[7] Uchanski, R.M., “Clear speech”, in D.B. Pisoni and R.E. Remez, [Eds], Handbook of Speech Perception, 207-235, Blackwell Publishers, 2005.

[8] Mooshammer, C. and Geng, C., “Acoustic and articulatory manifestations of vowel reduction in German”, Journal of the International Phonetic Association, 38(02), 2008.

[9] Van Son, R.J.J.H. and Pols, L.C.W., “An acoustic description of consonant reduction”, Speech Communication, 28(2):125-140, 1999.

[10] Cho, T., “Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /I,i/ in English”, The Journal of the Acoustical Society of America, 117(6):3867, 2005.

[11] Cole, J.S., Kim, H., Choi, H. and Hasegawa-Johnson, M., “Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech”, Journal of Phonetics, 35(2):180-209, 2007.

[12] Ohala, J.J., “The origin of sound patterns in vocal tract constraints”, in P. MacNeilage, [Ed], The production of speech, 189-216, Springer-Verlag, 1983.

[13] Westbury, J.R. and Keating, P.A., “On the naturalness of stop consonant voicing”, Journal of Linguistics, 22(1):145-166, 1986.

[14] Livijn, P. and Engstrand, O., “Place of articulation for coronals in some Swedish dialects”, Lund Working Papers in Linguistics, 49:113-114, 2001.

[15] Stevens, K.N. Acoustic Phonetics, Current Studies in Linguistics, Vol. 24, MIT Press, 1998.

[16] Hume, E., Johnson, K., Seo, M. and Tserdanelis, G., “A cross-linguistic study of stop place perception”, Proceedings of the ICPhS99, 2069-2072, 1999.

[17] Browman, C.P. and Goldstein, L., “Tiers in articulatory phonology, with some implications for casual speech”, in J. Kingston and M.E. Beckman [Eds], Papers in laboratory phonology I: Between the grammar and the physics of speech, 341-476, Cambridge University Press, 1990.

[18] Fougeron, C. and Keating, P.A., “Articulatory strengthening at edges of prosodic domains”, The Journal of the Acoustical Society of America, 101(6):3728-40, 1997.

[19] Keating, P.A., Cho, T., Fougeron, C. and Hsu, C.-S., “Domain-initial articulatory strengthening in four languages”, in J. Local, R. Ogden, and R. Temple [Eds], Phonetic interpretation: Papers in Laboratory Phonology VI, 143-161, Cambridge University Press, 2004.

[20] Pitt, M.A., Johnson, K., Hume, E., Kiesling, S. and Raymond, W., “The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability”, Speech Communication, 45(1): 89-95, 2005.

[21] Zuur, A.F., Ieno, E.N. and Smith, G.M., Analysing Ecological Data, volume 36 of Statistics for Biology and Health, Springer, 2007.

[22] Pinheiro, J.C. and Bates, D.M. Mixed Effects Models in S and S-Plus. Springer, 2000.