BIAS AND CONSISTENCY OF INDIVIDUAL LINGUAL ARTICULATORY BEHAVIOR AND ITS RELATIONSHIP TO THE FIRST AND SECOND FORMANTS

Carolina Lins Machado[1], Lei He[1 2]

[1] Dept. Computational Linguistics, University of Zurich, Zurich, Switzerland
[2] Dept. Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland
cmachado@ifi.uzh.ch, lei.he@uzh.ch

ABSTRACT

Individual variation in articulatory behavior can be characterized by bias and consistency in movement outcome. Consistency can be indicated by *variable error* (VE) representing precision of individual performance and bias by *constant error* (CE) representing tendency in movement outcome. The present study employs CE and VE to characterize individual articulatory behavior, and assesses the relationship between consistency and bias in the articulatory and acoustic domains. We computed CE and VE of tongue blade and dorsum kinematic trajectories and the first two formants' curves in the production of /æ/ and /ɑ/ by 20 native U.S. English speakers. The relationship between acoustic and kinematic VE and CE were revealed using gradient boosting machines. Results indicate that individual CE and VE vary over the time course of a vowel and that movement outcome is affected by linguistic constraints.

Keywords: articulation, consistency, bias, individual variability, speech production

1. INTRODUCTION

Speakers differ in their vocal tract morphology and in the articulatory strategies they adopt to produce a highly similar acoustic output [2]. Besides phonemic information, an acoustic signal carries speaker-specific information. For instance, in vowels, formant contours have been shown to contain substantial individual information related to idiosyncratic articulatory behavior [23, 26, 16, 24, 18]. This behavior is far from constant, since speakers can produce the same sound employing different articulatory strategies [19].

In this study we seek to characterize speakers' articulatory behaviors using two measures employed in the investigation of human motor performance, the error scores *constant error* (CE) and *variable error* (VE) [22, 10, 9]. CE indicates the tendency, or bias, in individual motor performance. It represents the amount and direction of error, or deviation, from a target or criterion in a subject's movement [10]. In the context of this paper, *target* is defined as the mean of all speakers in a cohort, because in the motor control literature individual differences in movement outcome are described in relation to a predefined target or to other performers in the same environmental condition [22]. In speech, CE may indicate how much and in which direction a speaker deviates from a target. For instance, compared to others, a speaker could have the tendency to overshoot (direction) their tongue movement, fronting this articulator by 8 mm (amount) when producing an open vowel. Moreover, this speaker may be quite variable in their advancement. This inconsistency in movement outcome is indicated by the VE. Together, CE and VE scores can characterize the articulatory performance of individual speakers, highlighting their consistency and bias in movement outcome.

Articulatory movements modulate the acoustic signal; therefore, computing error scores of features in the acoustic domain (even though error scores are primarily used to describe movement performance) becomes a reasonable step to assess the relationship between the acoustic and articulatory dimensions. As previously stated, formant contours carry significant speaker-specific information. Thus, in this study, formant dynamics of the English vowels /æ/ and /ɑ/ were the acoustic feature selected for analysis. These vowels were chosen since they differ in inherent formant changes. While /æ/ requires formant movement to convey phonemic information [12], therefore requiring more precision in motor command, /ɑ/ is believed to have a relative stable acoustic outcome, less sensitive to small variations in articulatory movements [15]. Consequently, /ɑ/ may be more variable across speakers, since articulatory movements may be less stifled by linguistic constraints.

The present study is exploratory in nature, seeking to understand individual variation in the articulatory domain using error scores. More specifically, this study seeks to (i) investigate individual tendency and consistency of speech production in the articulatory dimension, and to (ii) examine the relationship between acoustic and articulatory error scores of the English vowels /æ/ and /ɑ/. We expect speaker's error scores to be less variable in the production of /æ/,

since linguistic constraints may affect the amount of individual information in formant contours.

2. METHOD

2.1. Material

Productions of the vowels /æ/ and /ɑ/ in single-word citation form by twenty native speakers of U.S. English (10 M, 10 F) with an upper Midwest American English dialect background were selected from the EMA-MAE corpus [1]. We excluded vowel tokens produced in the context of rhotic, lateral, nasal, and approximant syllable onset or codas, due to the coarticulatory effects on vowel formants related to these consonants [25].

Kinematic data (sampling rate = 400 Hz) were collected using the electromagnetic articulograph (NDI Wave) along with time-synchronized acoustic recordings (sampling rate = 22 kHz). Sensor data related to the position of the tongue dorsum (TD) and the tongue blade (TB) were analyzed in both anteroposterior (x) and vertical (y) directions. For each speaker the kinematic data were head-corrected and calibrated using a bite-plate. The tongue kinematic data in this analysis contains tongue-jaw compound movements corresponding to the tract variables of *tongue body constriction location/degree* and *tongue tip constriction location/degree* in Articulatory Phonology [7].

Acoustic and kinematic data were processed in Praat [20] following the same steps as [6]. The resulting first and second formant curves and kinematic trajectories, both comprising 5 analysis points, were subsequently processed in R [21]. After outlier removal (vowels with long silences in the middle or excessive creak leading to incorrect measures), subsets of the data were created per vowel containing two acoustic variables, F1 and F2, and four kinematic variables: TDx, TDy, TBx, TBy. Overall the datasets consisted of 1240 data points for /æ/ and 990 for /ɑ/.

2.2. Data analysis

2.2.1. Error scores

Speaker bias and consistency of articulatory and acoustic outcomes were determined by calculating error scores for every variable at each analysis point. CE, indicating bias, was calculated as shown in equation (1), where x_i is the value for repetition i, T is the target value (the average over all speakers), and k is the number of repetitions the participant performed. A CE score of zero indicates no bias in outcome. A value greater than zero indicates overshoot, and a score less than zero expresses movement undershoot.

$$\text{CE} = \Sigma(x_i - T)/k. \quad (1)$$

$$\text{VE} = \sqrt{\Sigma(x_i - M)^2/k}\,. \quad (2)$$

Equation (2) shows the calculation of VE, indicating consistency, where x_i and k are defined as in (1), and M is the speaker's mean. A VE score of 0 indicates that the speaker's productions were always exactly the same, and a value greater than zero indicates the extent of outcome inconsistency. The extent to which inconsistency is considered high or low is an empirical question not addressed in the present study. We believe the level of error to be a function of the linguistic environment, where some phonemes may allow more error with little consequences while others may not, thus affecting what is considered high or low inconsistency. Here, we used the mean VE over all speakers for each vowel as a threshold; a VE > 1.5 for /æ/ or a VE > 2 for /ɑ/ indicates high inconsistency.

The above computations yield CE and VE scores for the first two formants (CE_F1, CE_F2, VE_F1 and VE_F2) and the four articulatory variables (CE_TBx, CE_TBy, CE_TDx, CE_TDy, VE_TBx, VE_TBy, VE_TDx, and VE_TDy). Subsequently, to determine a possible effect of vowel on the error scores related to lingual movement, we employed the Levene's test for equality of variances [13]. CE or VE were the response variable grouped by vowel, articulatory variable and analysis point.

2.2.2. Gradient Boosting

To better understand the acoustic-articulatory relationship in each vowel, we used generalized boosted regression modeling from the R package gbm [4]. Our models may reveal an acoustic-kinematic relationship not despite the speaker effect, but by actually taking speaker characteristics into consideration, since the CE and VE reflect non-normalized speaker-specific characteristics.

For each vowel, gradient boosted models with gaussian loss function were built for CE and VE, with the formants error score as the predicted variables and the articulatory error scores as explanatory variables. Important hyper-parameters [3] were tuned over 5000 iterations using grid search with 10-fold cross-validation. For the hyper-parameter optimization and model performance metric, the root mean square error (RMSE) was used. Models were trained on 80% of the data and performance was evaluated on the remaining 20%.

3. RESULTS

3.1. Error Scores

Error scores indicated the consistency and bias of each speaker's motor performance. Here we only present the main differences between two randomly

selected speakers, due to the page constraint. Error scores of all speakers are available in the supplementary figures S1 and S2 via https://osf.io/xud7h.

Figure 1 shows the tendency (CE) and consistency (VE) in vowel productions of two male speakers. Regarding their articulatory tendencies for the vowel /ɑ/, speaker 15ENM's anteroposterior movements (TBx and TDx) tend to be directionally biased to fall short of the target, while speaker 35ENM tends to overshoot TBx. Further, speaker 35ENM tends to undershoot TBy and speaker 15ENM tends to overshoot this movement. In their production of /æ/, CE indicates that speaker 35ENM tends to undershoot his vertical tongue movements (TBy, TDy) as opposed to speaker 15ENM, whose vertical movements tend to be overshot.

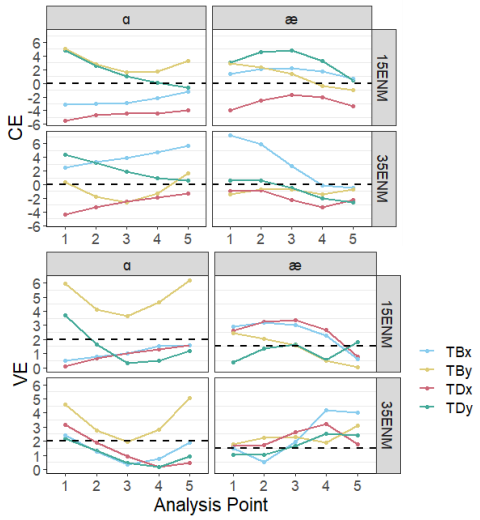

Figure 1: Constant Error (CE, top plot) and Variable Error (VE, bottom plot) scores of two male speakers calculated for four articulatory variables at five time points. In VE plots dashed lines correspond to the average inconsistency level.

In terms of consistency, VEs indicate that both speakers are most inconsistent in TBy of /ɑ/ production. In the anteroposterior movement (TBx and TDx), speaker 15ENM seems more consistent in the beginning of this vowel, while speaker 35ENM becomes more consistent towards the end. In the production of /æ/, speaker 15ENM becomes more consistent with time in his anteroposterior movement (TBx and TDx) and in TBy. Conversely, speaker 35ENM becomes more inconsistent with time. Together, CE and VE reveal that in the production of /ɑ/ 15ENM inconsistently overshoots TBy and 35ENM inconsistently undershoots it. Moreover, during the production of /æ/, although speaker 15ENM overshoots TBy, he tends to become more consistent with time. Contrariwise, speaker 35ENM tends to become less consistent with time undershooting TBy.

Due to differences in the magnitude and spread of VE and CE scores between /æ/ and /ɑ/ over all speakers, we decided to look into the effect of vowels in these scores. The results of the Levene's test indicated that the variance between vowels was not equal. Overall /ɑ/ had a significantly higher dispersion ($p < 0.001$) of CE and VE scores for all variables. The differences of variance between vowels for TBx were 1.54 (CE) and 0.75 (VE); for TBy 2.04 (CE) and 2 (VE); for TDx 4.28 (CE) and 1.12 (VE); and for TDy 0.99 (CE) and 3.06 (VE).

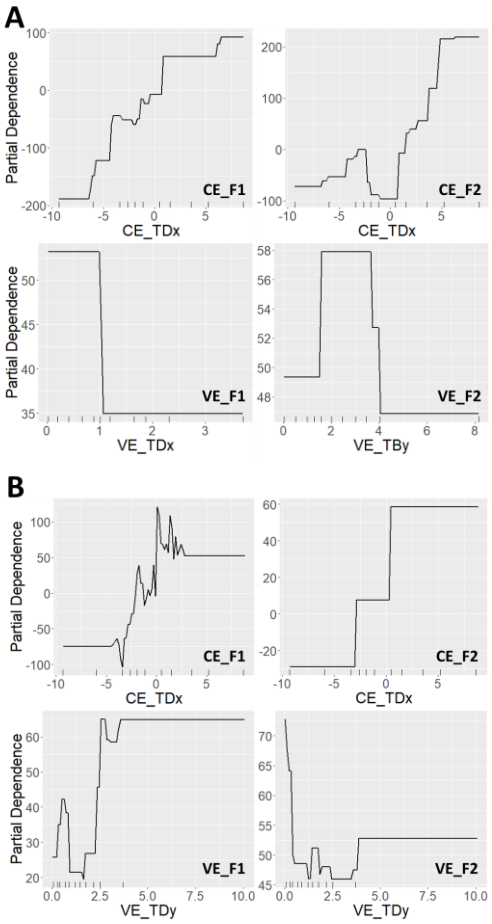

Figure 2: Partial dependence plots of the four models built for /æ/ (A), and /ɑ/ (B). The y-axis represents the marginal impact of the articulatory error scores to the acoustic error scores. Hash marks at the base of plots show distribution of error scores of each variable, in millimeters.

3.2. Gradient boosting

The normalized RMSE of all GBMs was equal or below 0.26, indicating good model fitting. The individual influence of the predictor variables on the response variable is indicated by their *relative importance*, a predictor's ranking based on their contributions to each model. For /æ/ the most important variables in predicting CE_F1, CE_F2 and VE_F1 were the median error scores of TDx, with a relative contribution of 45% (CE_F1), 59% (CE_F2) and 44% (VE_F1). In predicting VE_F2, TDx (30%)

was second to TBy (37%). For /ɑ/ the most important predictor of both formants CEs was TDx, with a contribution of 27% for F1 and 68% for F2. For the formants' VEs, TDy was the most important predictor to VE_F1 (39%) and VE_F2 (32%).

The partial dependence plots (Figure 2) illustrates the relationship between the error scores of the most relevant articulatory and acoustic variables. For both vowels, the partial dependence of CE_TDx on CE_F1 and CE_F2 is increasing over the main body of the data. Generally, a large CE_TDx means higher CEs for both formants. As for VE, in /æ/ VE_F1 seems higher with low VE_TDx, stepping downwards after a VE_TDx of 1, conversely, VE_F2 starts low, steeping upward after a VE_TBy of ca. 2. In /ɑ/, relationships between VE scores of F1–TDy and F2–TDy appear less straightforward. However, a general increase in the VE_TDy increases VE_F1 whereas a general increase in VE_TDy decreases VE_F2.

4. DISCUSSION

In this paper error scores were used to characterize individual articulatory performance. CE indicated bias, whether speakers in this dataset tend to undershoot or overshoot their tongue movements in the production of /æ/ and /ɑ/, and VE indicated how consistent their movements were.

Analyzing vowel dynamics revealed that bias in movement outcome is not a stationary characteristic. For instance, in speaker's 15ENM production of /ɑ/, although his TBx and TDx movements seem to show a quasi-constant CE score, other movements (e.g. TDy) seem to start with a great directional bias and approach the target over time. Similarly, changes in movement bias were observable in this vowel for speaker 35ENM, where TBx increased in bias with time, and TDy, showing the opposite progression, decreased in bias.

At an individual level, differences in the directional bias of the tongue articulatory variables could be an indication of quantal regions [17], i.e., a region of stability in the articulation-to-acoustics mapping. In this context, error scores may reveal which regions along the movement trajectory are more stable than others, therefore exhibiting a greater amount of movement bias. Our results seem to be in line with previous research, hypothesizing that formant trajectories may be more sensitive to continuous tongue movement at certain time points [5]. In addition, at a group level, differences in speakers' tendencies could be attributed to motor equivalence [19], which states that a similar acoustic outcome can be reached by employing different articulatory movements. Furthermore, we believe that aspects related to a speaker's physiology and motor preferences (e.g. [2]), along with linguistic constraints may determine the amount and direction of bias in tongue movement outcome.

Similarly, consistency in movement outcome seems to be restricted by linguistic constraints. Our results reflected the assumption that the production of /ɑ/ would be more variable than /æ/, because the latter relies on formant trajectories to denote phoneme identity, consequently requiring more controlled articulatory movements during its production, and thus constraining speaker-specific behavior. Moreover, our findings support previous studies demonstrating that some speech features are indeed less variable between speakers than others [24, 11, 8].

On the questions of the relationship between the error scores of articulatory and acoustic variables, we believe that interactions between the directions of tongue movement [14] may be expressed in the acoustic outcome. Regarding F2 bias, its dependence to CE_TDx for both vowels was unsurprising, since the relationship between this formant and the anteroposterior tongue movement follows a widely held view that F2 increases as the tongue advances. However, F2 consistency dependence to vertical tongue movements could be the result of an interaction between the vertical and horizontal directions of tongue movement. The current interpretation remains speculative since we did not test this assumption.

As to F1, we expected changes in tongue height to affect this formant, since vertical tongue movement alters the pharyngeal cavity where F1 resonates. Yet, with the exception of /ɑ/'s VE_F1, our results revealed a relationship between both vowels' F1 bias and /æ/'s consistency and the anteroposterior tongue dorsum movement (TDx). Ours is not the first study to find a relationship between F1 and the tongue anteroposterior direction (see [5] for an account on diphthongs). In the present study, we believe that tongue retraction may be the primary variable affecting the volume of the pharyngeal cavity [14] and consequently F1 CE and VE.

Ultimately, this study has shown that consistency and bias are non-stationary individual characteristics affected by linguistic constraints and that the acoustic-articulatory relationship still requires further investigation to understand its complexity.

6. ACKNOWLEDGEMENTS

This work was supported by the Forschungskredit of the University of Zurich (Grant No. FK-20-078), and the Swiss National Science Foundation (Grant #PZ00P1_193328) to LH.

7. REFERENCES

[1] A. Ji, J. J. Berry, and M. T. Johnson, "The Electromagnetic Articulography Mandarin Accented

English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 7719–7723. doi: 10.1109/ICASSP.2014.6855102.

[2] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker Variability in Hard Palate Morphology and Vowel Production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, Dec. 2013, doi: 10.1044/1092-4388(2013/12-0211).

[3] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[4] B. Greenwell, B. Boehmke, J. Cunningham, and G. B. M. Developers, "gbm: Generalized Boosted Regression Models." 2020. [Online]. Available: https://CRAN.R-project.org/package=gbm

[5] C. Dromey, G.-O. Jang, and K. Hollis, "Assessing correlations between lingual movements and formants," *Speech Communication*, vol. 55, no. 2, pp. 315–328, 2013, doi: https://doi.org/10.1016/j.specom.2012.09.001.

[6] C. Lins Machado, V. Dellwo, and L. He, "Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis," in *Proc. Interspeech 2022*, 2022, pp. 754–758. doi: 10.21437/Interspeech.2022-10397.

[7] C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989, doi: 10.1017/S0952675700001019.

[8] C. Schindler and C. Draxler, "Using spectral moments as a speaker specific feature in nasals and fricatives," in *Interspeech 2013*, Aug. 2013, pp. 2793–2796. doi: 10.21437/Interspeech.2013-639.

[9] D. Guth, "Space saving statistics: An introduction to constant error, variable error, and absolute error.," *Peabody Journal of Education*, vol. 67, no. 2, pp. 110–120, Jan. 1990, doi: 10.1080/01619569009538684.

[10] F. M. Henry, "Variable and Constant Performance Errors Within a Group of Individuals," *Journal of Motor Behavior*, vol. 6, no. 3, pp. 149–154, Sep. 1974, doi: 10.1080/00222895.1974.10734991.

[11] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, "A forensic phonetic study of dynamic sources of variability in speech: The dyvis project," in *Proceedings of the 11th Australasian international conference on speech science and technology*, 2006, pp. 13–18.

[12] G. S. Morrison and P. F. Assmann, Eds., *Vowel Inherent Spectral Change*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-14209-3.

[13] H. Levene, "Robust Tests for Equality of Variances," I. Olkin, Ed. Palo Alto, CA: Stanford University Press, 1960.

[14] J. H. Esling, "There Are No Back Vowels: The Laryngeal Articulator Model," *Canadian Journal of Linguistics/Revue canadienne de linguistiqu*e, vol. 50, no. 1–4, pp. 13–44, 2005, doi: 10.1017/S0008413100003650.

[15] J. S. Perkell *et al.*, "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *Journal of Phonetics*, vol. 28, no. 3, pp. 233–272, 2000, doi: https://doi.org/10.1006/jpho.2000.0116.

[16] K. McDougall, "Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies," *International Journal of Speech, Language and the Law*, vol. 13, no. 1, pp. 89–126, Jun. 2006, doi: 10.1558/sll.2006.13.1.89.

[17] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 17, no. 1–2, pp. 3–45, Jan. 1989, doi: 10.1016/S0095-4470(19)31520-7.

[18] L. He, Y. Zhang, and V. Dellwo, "Between-speaker variability and temporal organization of the first formant," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL209–EL214, 2019, doi: 10.1121/1.5093450.

[19] O. M. Hughes and J. H. Abbs, "Labial-Mandibular Coordination in the Production of Speech: Implications for the Operation of Motor Equivalence," *Phonetica*, vol. 33, no. 3, pp. 199–221, 1976, doi: doi:10.1159/000259722.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Jul. 22, 2021. [Online]. Available: http://www.praat.org/

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.Rproject.org/

[22] R. A. Schmidt, T. D. Lee, C. Winstein, G. Wulf, and H. N. Zelaznik, *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics, 2018.

[23] T. Kitamura and M. Akagi, "Speaker Individualities in Speech Spectral Envelopes and Fundamental Frequency Contours," in *Speaker Classification II*, vol. 4441, C. Müller, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 157–176. doi: 10.1007/978-3-540-74122-0_14.

[24] W. Heeren, "The contribution of dynamic versus static formant information in conversational speech," *International Journal of Speech Language and the Law*, vol. 27, no. 1, pp. 75–98, Aug. 2020, doi: 10.1558/ijsll.41058.

[25] W. Labov, S. Ash, and C. Boberg, *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter, 2006.

[26] X. Yang, J. B. Millar, and I. Macleod, "On the sources of inter- and intra- speaker variability in the acoustic dynamics of speech," *Proceedings of Fourth International Conference on Spoken Language Processing*. ICSLP '96, Oct. 1996, vol. 3, pp. 1792–1795 vol.3. doi: 10.1109/ICSLP.1996.607977.