BENIGN VS. HARMFUL VARIABILITY IN SECOND LANGUAGE VOWEL PRODUCTION

Jaekoo Kang[1,2], Hosung Nam[2,3], Wei-rong Chen[2] and D. H. Whalen[1,2,4]

[1]The Graduate Center, CUNY; [2]Haskins Laboratories; [3]Korea University; [4]Yale University
jkang@gradcenter.cuny.edu; dwhalen@gc.cuny.edu; chenw@haskins.yale.edu; hnam@korea.ac.kr

ABSTRACT

Variability is widespread in speech, but it is unlikely that all of it is harmful; variability in other domains has been shown to allow flexibility, within limits. Using one technique for separating the two, we applied Uncontrolled Manifold Analysis (UCM) to vowels in running speech. This results in two multidimensional manifolds, one the controlled manifold, where variation is harmful relative to the target, and a complementary uncontrolled manifold, where variation is benign. Utterances from 32 speakers of English were analysed acoustically and a UCM analysis (with intended vowel category as the target) was extracted. This was compared to the UCM analysis of 4 second-language (L2) speakers of English (Korean L1). Results indicate that L2 speakers had smaller uncontrolled variability, consistent with their lower accuracy. UCM suggests the possible direction of variability that L2 speakers need to master. Simulating along the controlled vs. uncontrolled manifolds further supports this suggestion.

Keywords: Uncontrolled Manifold; variability; vowel formants; second language; Korean-English.

1. INTRODUCTION

Variability in speech is both well-recognized and poorly understood, despite decades of research [7, 10]. Variability is extensive in all biological systems, of course, and recent advances in analysis have begun to separate the variations into two useful categories: harmful and benign [13, 14]. The technique of Uncontrolled Manifold Analysis (UCM) allows the exploration of different partitionings of variability into task-relevant (i.e., harmful-to-task) and task-irrelevant (i.e., benign-to-task) components without having to define these dimensions a priori [12]. The controlled manifold is that combination of dimensions in the data constrained to be within a minimal range, otherwise the target cannot be reached. It suggests the dimensions that are under neurological control [13]. The uncontrolled manifold, then, is that range of variability that can be tolerated while still reaching the target. The CM defines success, while the UCM defines flexibility.

Successful mastery of an action has been found to result in an overall decrease in motor variance while the UCM variability remained larger than the CM variability [8, 20]. Forcing the motor system to adapt to a broader range of conditions can also increase flexibility [3].

Very few UCM analyses of speech have been performed to date. Saltzman, Kubo and Tao [12] provide a theoretical basis for constructing the task space for the CM in terms of articulator constrictions; a corresponding CM based on acoustic targets would require a secondary hidden Markov model to provide the necessary input. Szabados and Perrier [16] addressed the issue of motor equivalence [11]. They found synergistic movements in the simulated 2D vocal tract to achieve formant targets for 10 French vowels. Given the variability of formants themselves [4] as well as the unnaturalness of the simulated speech, however, applicability to natural speech is unclear.

In addition, control and flexibility in second language (L2) speech remains to be explored. It is undetermined whether the L2 accented speech is comparable to L1 speech for the use of UCM and how that interacts with overall accuracy. Studies have demonstrated L2 speech deviates acoustically from L1 speech [2, 9]. Whether this deviation from L1 indicates harmful or benign to producing the target has not been explored.

The present project uses UCM analysis to address these issues through specific questions: Does the UCM differ for second language (L2) speakers compared to L1 speakers? Does this difference reveal harmful variability for the target achievement?

2. EXPERIMENTAL DATA

The basis for the UCM analysis of English L1 speakers was the X-ray Microbeam dataset (XRMB) from the University of Wisconsin [17]. 32 speakers who had been previously analysed by others [18] were selected. These speakers read a list of words and sentences as well as sequences of syllables while their productions were simultaneously recorded by the movements of gold pellets attached midsagittally on the tongue and the lips including other reference points. The pellet sensors were sampled at 145.54 Hz and the sound files were recorded at the sampling rate

of 21,739 Hz. These acoustic recordings were further re-sampled to 10 kHz to narrow the frequency range up to 5 kHz, which roughly corresponds to the linguistically meaningful range for American English vowels [4].

After selecting the nine monophthong vowels (/i, ɪ, ɛ, æ, ɑ, ʌ, ɔ, ʊ, u/) [18], the Short-Time Fourier Transform (STFT) was computed at each vowel mid-point based on the acoustic analysis window (45 msec) and the frequency bins (1024). Vowels smaller than this window were ignored. To reduce the redundancy in the STFT spectrum, 40 linear filter banks were computed. The extracted filter banks were z-score normalized by speaker for the cross-speaker comparison.

To compare with the second-language speakers, the electromagnetic articulography (EMA) recordings of the four L2 speakers (2 female) whose L1 is Korean were used. These L2 speakers read newspaper articles (one written in English and the other in Korean) at their normal speed in a laboratory environment. Only the recordings of the English article were analysed for the current project. Despite the L2 accentedness, speakers were relatively fluent in English. Their passage reading was recorded both articulatorily (sampled at 200 Hz) and acoustically (sampled at 16 kHz). Likewise, the acoustic recordings were re-sampled to 10 kHz. Speaker-wise normalization was applied by z-scoring.

10 principal components were calculated from the 40 normalized linear filter banks using Principal Component Analysis (PCA), combining both XRMB and L2 dataset to make shared PCs across speakers. Mel filter banks were also tested; however, the amount of the explained variability for the data was less (81.8%) than that of linear filter banks (88.75%), which led us to use linear filter banks instead.

3. UNCONTROLLED MANIFOLD METHOD (UCM)

3.1. Forward mapping computation

UCM analysis depends on both input and outcome data being multidimensional, allowing the method to apply to a great many types of data. The UCM method decomposes n-dimensional input data space into two subspaces ("manifolds"), one of which (the CM) will affect the outcome, and the other of which (the UCM) will not (meaningfully) affect the outcome. The CM and UCM manifolds are orthogonal to each other and their dimensional sum is n as defined within the n-dimensional input space. Once the two spaces, UCM and CM, are computed, we can project the input data on each of the spaces. If the variance on the UCM is higher than on the CM, there is synergy and the UCM represents benign variation; if not, there is no benign variability (UCM = CM) or even destructive compensation (UCM < CM). The larger the UCM/CM ratio, the greater the synergy [6].

UCM analysis requires a known forward mapping from the elemental variables to the task variables. The mapping from the acoustics to the perceptual sound categories, however, are not directly estimable. Using linear regression [1, 5] can possibly approximate this relationship; however, the linear methods are often error-prone where the underlying data includes more complex relationship rather than being linear. This led us to the use of deep neural networks. The hyper parameters of our neural networks are: three hidden layers, 500 nodes, sigmoid activation, dropout with 0.5 [15], 20% held-off for test set. The vowel categories were determined as the forced aligned segmental labels [21] because both dataset were elicited speech. That is, there were explicit target for the speech. Whether the segmental labels match with the human perceptual identification of vowels is not pursued in the current project. To account for the cross-speaker differences, the neural-network forward mapping function was trained for individual XRMB speakers, resulting in 32 forward mapping models.

3.2. Jacobian matrix and null space

After establishing the forward mapping function from the acoustics (10 dimensional; i.e., 10 PCs) to vowel targets (9 dimensional; i.e., 9 vowels), the Jacobian matrix is defined as the partial derivatives of the vowel target variables with respect to the n acoustic variables, as follows:

$$Y = f(X) \quad (1)$$

$$J(X) = \partial Y / \partial X, \quad (2)$$

where X is 10-d acoustic variables, Y is 9-d vowel target variables and f is a neural-net forward mapping function. The function $J(.)$ is the Jacobian matrix, and ∂X and ∂Y indicates small displacements in acoustic spaces and task (vowel target) spaces, respectively.

The Jacobian matrix (J), defined in (2), captures how much changes in the elemental variables (i.e., 10 PCs from linear filter banks) result in the changes in the task variables (i.e., vowel probabilities). For the current project, the calculations of partial differentiation were approximated numerically. Once the Jacobian matrix was calculated, the UCM space is defined as the null space (spanned by the basis vector ε^{10d}) of the Jacobian, calculated by Singular Value Decomposition, denoted as ε as in (3):

(3) $J \cdot \varepsilon = 0$, where $\varepsilon \neq 0$.

Varying along this UCM space indicates benign-to-task variability (flexibility). The space which is orthogonal to the UCM space is defined as the CM space, denoted as $\varepsilon_{\perp}$. Movements along the CM space are harmful to the task (i.e., vowel probabilities). The amount of benign-to-task and harmful-to-task variabilities were calculated as the standard deviations of the projection of 10-d acoustic features on to the UCM space and the CM space respectively, normalized by the degree of freedom of each space. The formulation is described in (4) and (5):

(4) $$UCM_{projected} = (X \cdot \varepsilon) \cdot \varepsilon^T$$
$$CM_{projected} = (X \cdot \varepsilon_{\perp}) \cdot \varepsilon_{\perp}^T,$$

(5) $$UCM_{score} = \frac{\sigma UCM_{projected}}{DF_{UCM}}$$
$$CM_{score} = \frac{\sigma CM_{projected}}{DF_{CM}},$$

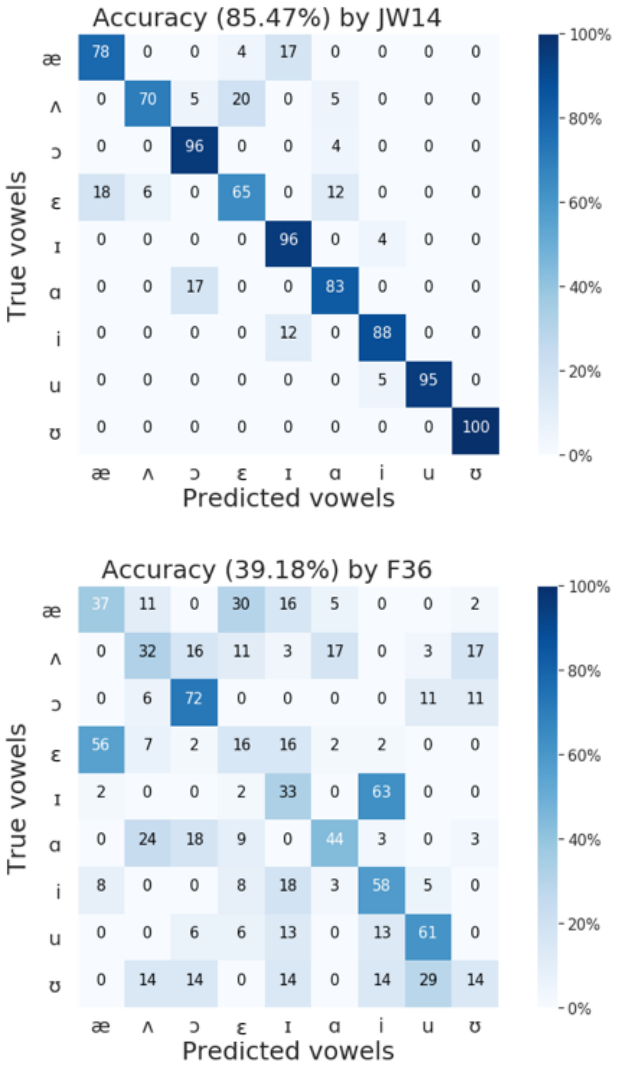
where the standard deviation of $UCM_{projected}$ and $CM_{projected}$ is divided by the degrees of freedom, DF_{UCM} (i.e. 1) and DF_{CM} (i.e. 9), respectively. UCM_{score} denotes the amount of benign-to-task variability and CM_{score} that of harmful-to-task variability. The computations of UCM and CM spaces for native speakers were based on each speaker's own forward mapping model as described in Sec. 3.1. For L2 speakers, the UCM and CM spaces were calculated based on a selected single native English speaker's forward mapping (JW14). Instead of training data from all native speakers, only a single speaker was chosen as a reference because accounting for individual vocal tract differences is beyond the scope of the current study. This was left for the future work. For the same reason, only one L2 speaker (F36) was chosen for the comparison.

4. RESULTS

4.1. Accuracy of the forward mapping

The mean accuracy of our neural-net forward mapping models for the native speakers was 78% (SD=5%), tested on the held-off test set (20% of the total data). Figure 1 demonstrates the confusion matrices for a female native speaker (JW14) and a female L2 speaker (F36), both of whom were in their mid-thirties. As expected, the mean accuracy for F36 (L2) (39.18%) is lower than that for JW14 (L1) (85.47%). In particular, the English vowels pairs /i, ɪ/ and /ɛ, æ/ each map on to a single vowel in Korean [19] and were more susceptible to be confused than others for Korean speakers as shown in Figure 1 (58% for /i/, 33% for /ɪ/, 16% for /ɛ/ and 37% for /æ/).

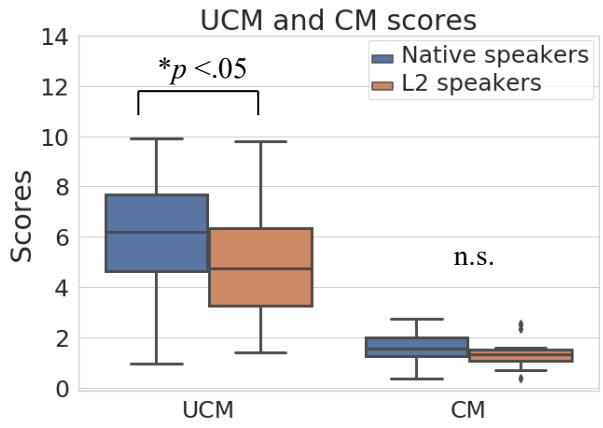
Figure 1. The confusion matrices for 9 vowel categories from the forward mapping model. Accuracies are annotated in each cell. **Top**: result from a female native speaker (JW14). **Bottom**: result from a female L2 speaker (F36).



4.2. UCM and CM scores

The overall distribution of UCM and CM scores is shown in Figure 2. The results of two one-way ANOVA models, for UCM and CM separately, with Nativity ('Native speakers' vs. 'L2 speakers') as the main effect, indicated a significant effect of Nativity on the UCM scores ($F(1, 203) = 6.53$, $p < 0.05$), but not on the CM scores ($F(1, 203) = 3.26$, $p = 0.07$).

Figure 2. The overall UCM and CM scores separated by native English speakers (blue) and L2 speakers (red). n.s. indicates no significant difference.



4.3. A comparison of L1 vs. L2

The UCM scores for the native English speakers were significantly higher than those of L2 speakers as shown in 4.2., indicating lesser benign variability in L2 speech than that of the native speakers. Whether this difference in the amount of benign variability, or UCM score, is related to the acoustical differences were further investigated by reconstructing the possible range (−1.5 to +1.5 SD) of benign variability based on the native English speaker's vowel production. Figure 3 and Figure 4 illustrate the vowel spectra of English /ɪ/ and /ɛ/, respectively, by the same female native speaker (JW14) simulating changes in the UCM weighting from −1.5 to +1.5. For comparison, two vowels from a female L2 speaker (F36) are shown in the lower panel of Figure 3: an L2 /ɪ/ and an L1 (Korean) /i/, '이'. There is very little difference between these two, consistent with a lack of distinction in L2 production. As can be expected, the formants for L2 /ɪ/ do not match any of the candidates from the L1 English speaker's UCM range in the upper panel of Figure 3. For English /ɛ/ in Figure 4, L1 /ɛ/ and L2 /ɛ/ are separable under 2236 Hz (lower panel). The alignment of F2 is also fairly close between the native /ɛ/ and L2 /ɛ/, '애'; however, it has relatively lower amplitude than the native speaker's. The spectral envelop for L2 /ɛ/ still does not align with the native speaker's UCM range, indicating the sizable deviation from the benign variability.

Figure 3. **Top**: a native English speaker's range of benign acoustic variability from −1.5 (red) to +1.5 (blue) weighting of the UCM for the vowel /ɪ/. **Bottom**: an L2 speaker's mean spectrum of /ɪ/ (black line) overlaid with their mean L1 Korean vowel /i/, '이' (dashed line).

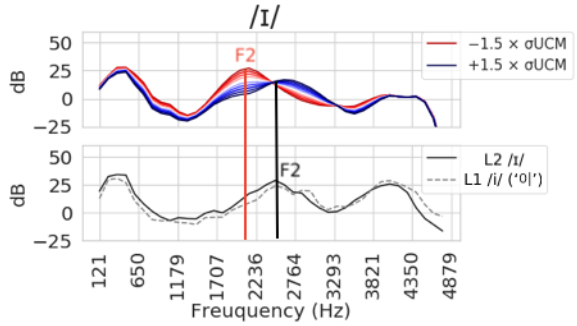
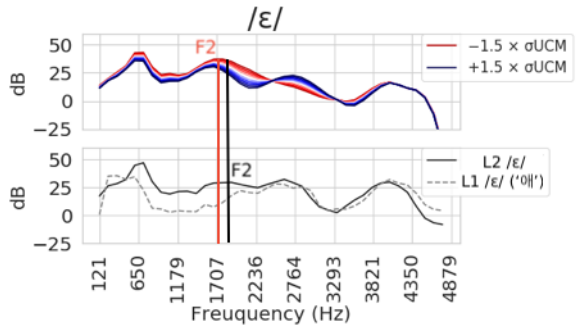


Figure 4. Vowel spectra of English /ɛ/. **Top**: a native English speaker's range of benign variability from −1.5 (red) to +1.5 (blue) weighting of the UCM for the vowel /ɛ/. **Bottom**: a L2 speaker's mean spectra of each vowel (black line) overlaid with their mean L1 Korean vowel /ɛ/, '애', (dashed line) for the comparison.



5. DISCUSSION

Taking acoustics as a means of examining variability in L1 and L2 vowel production allowed us to separate benign and destructive variability via the uncontrolled manifold (UCM) analysis. L2 speakers were found to have consistently smaller UCM scores, indicating a less flexible control of the English vowel space. A simulation in the UCM space further demonstrated how the L1-L2 difference in UCM scores is reflected in the spectrum, possibly as an indicator of L2 accentedness.

Redundancy is common in language, and here it can be seen as a part of flexibility: When multiple means of conveying a category exist, speakers are able to adapt to changing circumstances. While there are individual differences in such variability [18], there are presumably language-specific differences as well. Mastering that variability is therefore presumably part of competence and, ultimately, the reduction of L2 accent.

6. ACKNOWLEDGMENTS

Work supported by (US) NIH grant DC-002717 (Haskins Laboratories).

7. REFERENCES

[1] Ferreira de Freitas, S. M. S., & Peter Scholz, J. (2010). A comparison of methods for identifying the Jacobian for uncontrolled manifold variance analysis. *Journal of Biomechanics*, *43*(4), 775–777.

[2] Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470.

[3] Greve, C., Hortobàgyi, T., & Bongers, R. M. (2015). Physical demand but not dexterity is associated with motor flexibility during rapid reaching in healthy young adults. *PLoS ONE*, *10*(5), 1–21.

[4] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

[5] Krishnamoorthy, V., Goodman, S., Zatsiorsky, V., & Latash, M. L. (2003). Muscle synergies during shifts of the center of pressure by standing persons: Identification of muscle modes. *Biological Cybernetics*, *89*(2), 152–161.

[6] Latash, M. L., Scholz, J. P., & Schöner, G. (2002). Motor control strategies revealed in the structure of motor variability. *Exercise and Sport Sciences Reviews*, *30*(1), 26–31.

[7] Lindblom. (1990). Explaining phonetic variation; A sketch of H & H theory,in speech production and speech modeling. *Kluwer Academy, Dordrecht, Netherland*, 403–439.

[8] Morrison, A., Mcgrath, D., & Wallace, E. S. (2016). Motor abundance and control structure in the golf swing. *Human Movement Science*, *46*, 129–147.

[9] Munro, M. J. (1993). Productions of English Vowels By Native Speakers of Arabic: Acoustic Measurements and Accentedness Ratings. *Language and Speech*, *36*(1), 39–66.

[10] Perkell, J. S., & Klatt, D. H. (Eds.). (2014). *Invariance and variability in speech processes*. *Psychology Press*.

[11] Perkell, J. S., Matthies, M. L., Svirsky, M. a, & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study. *The Journal of the Acoustical Society of America*, *93*, 2948–2961.

[12] Saltzman, E. L., Kubo, M., & Tsao, C.-C. (2006). Controlled variables, the Uncontrolled Manifold, and the Task-dynamic model of speech production. *Dynamics in Speech Production and Perception*, 21–31.

[13] Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research*, *126*(3), 289–306.

[14] Scholz, J. P., & Schöner, G. (2014). Use of the uncontrolled manifold (UCM) approach to understand motor variability, motor equivalence, and self-motion. *Advances in Experimental Medicine and Biology*, *826*, 91–100.

[15] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, (15), 1929–1958.

[16] Szabados, A., & Perrier, P. (2016). Uncontrolled Manifolds in vowel production: Assessment with a biomechanical model of the tongue. In *Proceedings of Interspeech* (pp. 3579–3583).

[17] Westbury, J. (1994). *X-ray microbeam speech production database*. Madison, WI: Waisman Center, University of Wisconsin.

[18] Whalen, D. H., Chen, W. R., Tiede, M. K., & Nam, H. (2018). Variability of articulator positions and formants across nine English vowels. *Journal of Phonetics*, *68*, 1–14.

[19] Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics*, *24*(2), 245–261.

[20] Yang, J.-F., & Scholz, J. (2005). Learning a throwing task is associated with differential changes in the use of motor abundance. *Experimental Brain Research*, *163*(2), 137–158.

[21] Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, *123*, 3878.