Comparing Vowel Category Response Surfaces over Age-varying Maximal Vowel Spaces within and across Language Communities

Andrew R. Plummer[1], *Lucie Ménard*[2], *Benjamin Munson*[3], *Mary E. Beckman*[1]

[1]Dept. of Linguistics, The Ohio State University, Columbus, OH, USA
[2]département de linguistique, Université du Québec à Montréal, Montréal, Canada
[3]Dept. of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN, USA
{plummer,mbeckman}@ling.ohio-state.edu, menard.lucie@uqam.ca, munso005@umn.edu

Abstract

We investigate vowel category perception within and across languages by proffering a statistical methodology for creating vowel category response surfaces over maximal vowel spaces based on the responses of subjects from five different language communities to vowel stimuli generated by an age-varying articulatory synthesizer. The methodology is based on an additive modeling approach to surface regression within the general smoothing spline approach to statistical modeling. We also put forward a simple method for the comparison of surfaces and demonstrate its basic utility by comparing response surfaces derived from Greek and Japanese subjects. We discuss the results of the comparison with attention to the potential of the approach to reveal meaningful differences between and within the vowel systems of different language communities.

Index Terms: vowel categorization, cross-language perception, response surface, regression spline, additive model

1. Introduction

In this paper, we present a statistical modeling methodology based on analysis of a set of cross-language vowel categorization experiments [1]. Seven sets of vowel stimuli were generated by the *Variable Linear Articulatory Model* (VLAM) [2], one for each of the seven ages: 6 months, 2, 4, 5, 10, 16, and 21 years. Each set of stimuli was situated within a maximal vowel space [3, 4] producible by the model at the corresponding age. The stimuli were categorized by members of 5 different language communities: Cantonese (n=15), English (n=21), Greek (n=21), Japanese (n=21), and Korean (n=20). Each listener assigned each stimulus a vowel category from the listener's native language, along with a "goodness rating" [5, 6] indicating how good the listener felt that stimulus was as an example of the assigned category. The experimental components are further described in Section 2.

The statistical methodology (Sections 3 and 4), based on a smoothing spline approach [7, 8] to additive modeling [9, 10, 11, 12], provides a process for estimating a set of "vowel category response surfaces" over the maximal vowel space for each age, based on a listener's identification responses and associated goodness ratings for the 38 stimuli for that age. We present a process for comparing surfaces, along with an application of the comparative method, focusing on the response surfaces yielded by the Japanese and Greek listeners (Section 5). Specifically, we explore the similarities and differences between their five-vowel systems brought to light by the comparative method. We conclude with discussion of the application of the method, interpretation of its results, and directions for further study (Section 6).

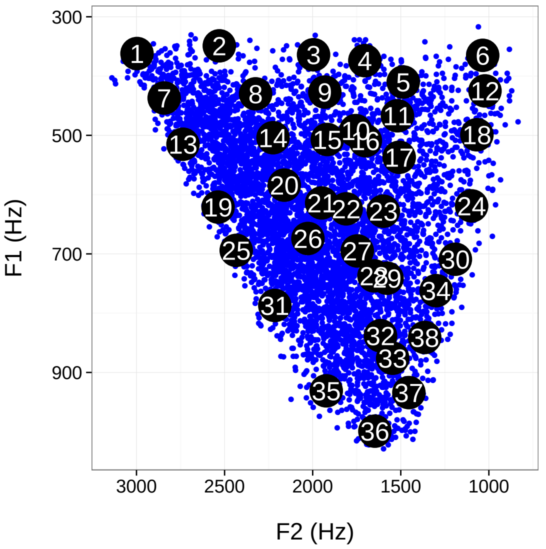

Figure 1: *Age 10 vowel prototypes* P(10) *within the age 10 maximal vowel space* MVS(10).

2. Vowel perception experiments

The VLAM [2] is a computational model of the articulatory system and its speech production capacities. Midsagittal representations are wrought by configuring "articulatory blocks" [13, 14, 15] corresponding to jaw height, tongue body position, tongue dorsum position, tongue apex position, lip protrusion, lip height, and larynx height. The VLAM is age-varying and capable of representing vocal tract lengths ranging from those of infants to young adults, calibrated in accordance with age-related "organic variation" [16, 17].

Given an age in years, the set of all articulatory configurations of the VLAM at that age that do not result in occlusion of the oral cavity yield a corresponding *maximal vowel space* (MVS) [3, 4] for that age. We take each MVS to be characterized by a set of formant patterns. Each formant pattern within an MVS is identified with a *formant vector* whose components are the first three formant frequencies of that formant pattern. We fix the following age index AGES $= \{0.5, 2, 4, 5, 10, 16, 21\}$ for indexing the MVSs discussed below. For each $a \in$ AGES, let $\text{MVS}(a)$ be a dense sampling of the MVS for age a, subject to the following constraints: minimal constriction area of each articulatory configuration was fixed at 0.3 cm^2 for ages 2 and over, and 0.15 cm^2 for the 6 m.-o., in accordance with previous modeling [18], and lip area was constrained from 0.1 cm^2 to 8 cm^2 for all ages. Each $\text{MVS}(a)$ contains approximately 5,000 formant vectors.

The stimuli used in the perceptual experiments [1] were *vowel prototypes* [19], or simply *prototypes*, selected from the MVSs for each $a \in \text{AGES}$. The selection process [20, 18] is meant to yield a set of cross-linguistically relevant prototypes in accordance with the dispersion-focalization theory [21, 22]. For each $a \in \text{AGES}$, a set of 38 prototypes, denoted by $\text{P}(a)$, were selected from $\text{MVS}(a)$. Prototypes in $\text{P}(a)$ are indexed $\mathbf{p}^i$, $1 \leq i \leq 38$, though the superscript is often dropped. The set $\text{P}(10)$ of prototypes for age 10 is depicted within $\text{MVS}(10)$ in Figure 1.

The perceptual experiments [1] elicited responses from native speakers of Cantonese (n=15), American English (n=21), Greek (n=20), Japanese (n=21), and Korean (n=20) for all 38 prototypes in $\text{P}(a)$ for all 7 ages. Cantonese- and American English-speaking listeners categorized each prototype by clicking on any of 11 keywords representing the monophthongal vowels in each language. Listeners for the other languages categorized by clicking on a symbol or symbol string that unambiguously represented a (short monophthongal) vowel in isolation, choosing among 7 vowels (Korean-speaking listeners) or among 5 vowels (Greek- and Japanese-speaking listeners).

In addition to assigning a category to each prototype, each listener provided a visual analog scale (VAS) [23, 5, 6] value indicating the "goodness" of that prototype as a representative of the assigned category. The VAS values ranged from 90-535, (90 best, 535 worst), though it is convenient to range normalize (we use min-max range normalization, though others are viable), and order reverse the scale, which hereafter ranges from 0-1 (1 best, 0 worst).

3. Formalizing subject responses

In this section, we put forward a formal method for generalizing over a subject's response to prototypes in $\text{P}(a)$. Consider, say, subject 12 from the Greek language community (G), which we denote s_{12}^G. Let $C_G = \{\mathsf{i}, \mathsf{e}, \mathsf{a}, \mathsf{o}, \mathsf{u}\}$ be the set of vowel categories for language community G, and let VAS denote the interval $[0, 1]$ of possible VAS values for vowel prototype category ratings. Let $a \in \text{AGES}$. For each $\mathbf{p} \in \text{P}(a)$, let c and γ, respectively, be the category and VAS goodness rating assigned to $\mathbf{p}$ by s_{12}^G. We can then define a function $R(s_{12}^G, a) : \text{P}(a) \rightarrow C_G \times \text{VAS}$ such that $\mathbf{p} \mapsto (\mathsf{c}, \gamma)$. That is, the codomain of this function is a set of ordered pairs called *responses* whose first component is a category in C_G, and second component a VAS "goodness" value. In the case of a "no-response" from s_{12}^G we may augment C_G and VAS with an element NA.

We can extend $R(s_{12}^G, a)$ to reflect implicit judgments about how well each prototype represents categories not assigned to that prototype. For each $\mathsf{c} \in C_G$, let $\gamma_{\mathsf{c}} : C_G \times \text{VAS} \rightarrow C_G \times \text{VAS}$ such that $(\mathsf{c}', \gamma) \mapsto (\mathsf{c}, \gamma)$ if $\mathsf{c}' = \mathsf{c}$, and $(\mathsf{c}, \alpha(1-\gamma))$ otherwise, where $0 \leq \alpha \leq 1$. The *response category extension parameter* α allows us to make a more general model than in previous work [24] (which corresponds to a choice of alpha = 0), to be more compatible with general Signal Detection Theory approaches [25] (which might be approximated by choosing alpha = 1). To exemplify, suppose s_{12}^G gives response $r = (\mathsf{i}, 0.9)$ and $\alpha = 0.5$. Then $\gamma_{\mathsf{i}}(r) = (\mathsf{i}, 0.9)$, while $\gamma_{\mathsf{o}}(r) = (\mathsf{o}, 0.05)$. We can then compose each γ_{c} with $R(s_{12}^G, a)$ to obtain functions

$$R_{\mathsf{c}}(s_{12}^G, a) =_{def} \gamma_{\mathsf{c}} \circ R(s_{12}^G, a) : \text{P}(a) \rightarrow C_G \times \text{VAS}$$

such that, for each $\mathsf{c} \in C_G$, each $\mathbf{p} \in \text{P}(a)$ is mapped to a response reflecting its goodness as an example of c.

Let $\text{LANG} = \{\text{C}, \text{E}, \text{G}, \text{J}, \text{K}\}$ be an index set over denotations of the five language communities. Given a language community $\ell \in \text{LANG}$, let C_ℓ denote the set of vowel categories for ℓ. Let n_ℓ denote the number of subjects for ℓ. Given an age $a \in \text{AGES}$, a subject s_τ^ℓ, where $1 \leq \tau \leq n_\ell$, and a category $\mathsf{c}^\ell \in C_\ell$, the function $R_{\mathsf{c}}(s_\tau^\ell, a)$ is called an *individual category response function for* s_τ^ℓ *at age a*, or simply an *individual category response function* (ICRF). We construct ICRFs for each subject and category from each language community in LANG, for each age $a \in \text{AGES}$.

4. Vowel Category Response Surfaces

In this section, we put forward a formal method for extending the domain of an ICRF $R_{\mathsf{c}}(s_\tau^\ell, a)$ from $\text{P}(a)$ to all of $\text{MVS}(a)$. The basic idea is to use a regression technique to construct a "surface" of responses over $\text{MVS}(a)$ using $R_{\mathsf{c}}(s_\tau^\ell, a)$. The response surface value for a formant vector $\mathbf{f} \in \text{MVS}(a)$ is meant to approximate s_τ^ℓ's VAS goodness rating of $\mathbf{f}$ as an example of vowel category c for age a.

The regression is carried out using smoothing spline-based additive models [10, 11, 26, 12]. In the statistical formulation, we begin with a *response variable* Y and observed *response vector* $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, and *design variables* X_j with observed *design vectors* $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n,j})^T$, where $1 \leq j \leq p$. Design vectors are arranged in a matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, whose rows are denoted $\mathbf{x}^i$, $1 \leq i \leq n$. We are interested in deriving a *fit* $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$ such that $\mathbf{y} = \hat{\mathbf{y}} + \epsilon$, for residuals $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$, that bears a smooth relationship to the $\mathbf{x}_j$, though not necessarily that of a least-squares line. One of the simplest ways to obtain such a fit is through the use of smooth functions $g_j(X_j)$, and an *additive predictor* $\beta + \sum_{j=1}^p g_j(X_j)$.

To illustrate, consider the univariate case of estimating a smooth function g from the n observations (y_i, x_i) such that $y_i = g(x_i) + \epsilon_i$, where ϵ_i is a random error term. We can estimate g using a "thin-plate regression spline" method [26, 12], which involves finding a function $\hat{g}$ minimizing

$$\sum_{i=1}^{n} [y_i - h(x_i)]^2 + \lambda J_{md}(h).$$

The term on the left determines closeness of fit, while the term on the right controls the smoothness of the fit. The *smoothing parameter* λ, which can be estimated along with g, controls the trade off between these terms: as $\lambda \rightarrow \infty$ the fit approaches a straight line, while $\lambda = 0$ yields an unpenalized regression spline estimate. The operator J_{md} has the general form:

$$\int \cdots \int_{\mathbb{R}^d} \sum_{v_1 + \cdots + v_d = m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial {x_1}^{v_1} \ldots \partial {x_d}^{v_d}} \right)^2 dx_1 \ldots dx_d.$$

where d is the number of arguments to h, and m is the desired order of partial derivative. The univariate case $d = 1$ considered above generalizes easily to additive predictors involving more than one design variable, but also to smooth functions over more

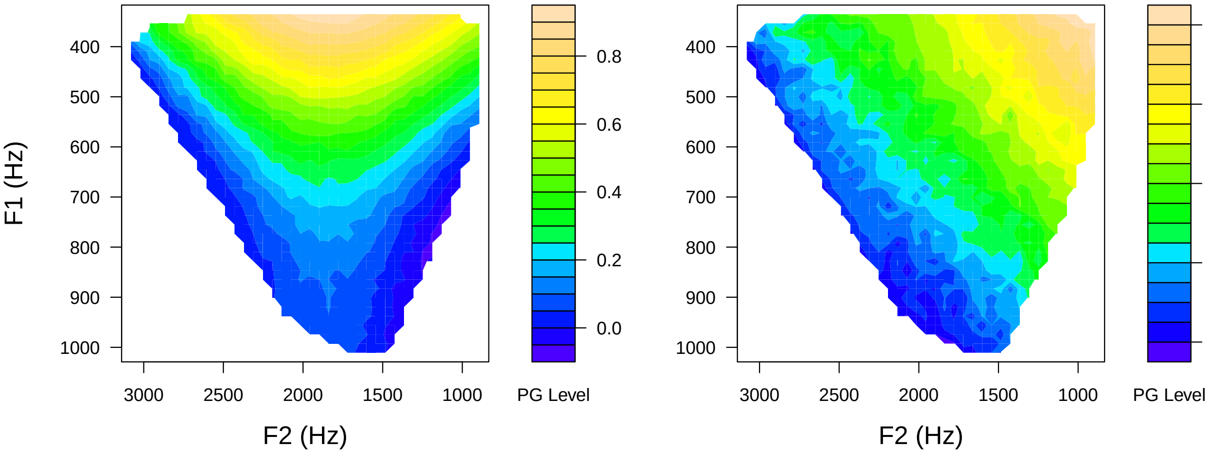

Figure 2: *Vowel category response surfaces for $R'_u(s^J_{20}, 10)$ (left) and $R'_u(s^G_{12}, 10)$ (right) where the response category extension parameter $\alpha = 0.5$. The "PG Level" is the predicted goodness level, approximating a subject's goodness rating of a formant vector as an example of the category* u.

than one variable. We use the computationally tractable "thin-plate regression spline" method for response surface regression available in the mgcv R package [26, 12].

Given an $\textsc{mvs}(a)$, let F^a_1, F^a_2, and F^a_3 be variables whose observed values $f^a_{k,1}$, $f^a_{k,2}$, and $f^a_{k,3}$ constitute the first, second, and third components, respectively, of the formant vectors in $\textsc{mvs}(a)$. Let $\mathbf{f}^a_1$, $\mathbf{f}^a_2$, $\mathbf{f}^a_3$ be the vectors of observed values for F^a_1, F^a_2, and F^a_3, respectively. Form the data matrix $\mathbf{F}^a = (\mathbf{f}^a_1\ \mathbf{f}^a_2\ \mathbf{f}^a_3)$. Similarly, form a data matrix $\mathbf{P}^a$ from the rows of $\mathbf{F}^a$ that correspond to the prototypes in $\textsc{p}(a)$. Thus each formant vector in $\textsc{mvs}(a)$ is indexed by its row position in $\mathbf{F}^a$, and each prototype in $\textsc{p}(a)$ by its row position in $\mathbf{P}^a$. Let P^a_1, P^a_2, and P^a_3 be variables whose observed values are the first, second, and third columns of $\mathbf{P}^a$.

Given an ICRF $R_{\mathsf{c}}(s^\ell_\tau, a)$, for each $\mathbf{p}^i \in \textsc{p}(a)$, let γ_i be the second component of the response $R_{\mathsf{c}}(s^\ell_\tau, a)(\mathbf{p}^i) = (\mathsf{c}, \gamma)$, i.e., $\gamma_i = \gamma$, and let $\mathbf{y}_{\mathsf{c}} = (\gamma_1, \ldots, \gamma_{38})^T$. We can now estimate smooth functions $g_1(P^a_1)$, $g_2(P^a_2)$, and $g_3(P^a_3)$ for an additive predictor $\sum_{j=1}^3 g_j(P^a_j)$ using $\textsc{p}(a)$ as a data matrix and $\mathbf{y}_{\mathsf{c}}$ as response vector, yielding an additive model $\mathbf{y}_c = \sum_{j=1}^3 g_j(P^a_j) + \epsilon$. More importantly, we can now predict category goodness ratings for each $\mathbf{f} \in \textsc{mvs}(a)$ by applying the additive model to $\mathbf{F}^a$. Given $\mathbf{f}^k \in \textsc{mvs}(a)$, let γ_k be its predicted goodness value under the additive model derived from $R_{\mathsf{c}}(s^\ell_\tau, a)$ and $\mathbf{y}_{\mathsf{c}}$, and let $r^k = (\mathsf{c}, \gamma_k)$. Pairing each $\mathbf{f}^k$ with r^k, we can define

$$R'_{\mathsf{c}}(s^\ell_\tau, a) : \textsc{mvs}(a) \rightarrow C_\ell \times \textsc{vas},$$

which approximates an extension of $R_{\mathsf{c}}(s^\ell_\tau, a)$ from $\textsc{p}(a)$ to all of $\textsc{mvs}(a)$.

Given an age $a \in \textsc{ages}$, a category $\mathsf{c}^\ell \in C_\ell$, and a subject s^ℓ_τ, where $1 \leq \tau \leq n_\ell$, the function $R'_{\mathsf{c}}(s^\ell_\tau, a)$ is called a *vowel category response surface for* s^ℓ_τ *at age* a, or simply a *vowel category response surface* (VCRS). Figure 2 depicts the VCRSs $R'_{\mathsf{u}}(s^G_{12}, 10)$ and $R'_{\mathsf{u}}(s^J_{20}, 10)$ where the response category extension parameter $\alpha = 0.5$.

5. Comparing surfaces

We now define a method for comparing VCRS patterns within and across language communities. Let $\ell_1, \ell_2 \in \textsc{lang}$. Given subjects $s^{\ell_1}_\tau$ and $s^{\ell_2}_\pi$, and categories $\mathsf{c}_1 \in C_{\ell_1}$ and $\mathsf{c}_2 \in C_{\ell_2}$, consider the corresponding VCRSs $R'_{\mathsf{c}_1}(s^{\ell_1}_\tau, a)$ and $R'_{\mathsf{c}_2}(s^{\ell_2}_\pi, a)$. We begin by defining similarity between $R'_{\mathsf{c}_1}(s^{\ell_1}_\tau, a)$ and $R'_{\mathsf{c}_2}(s^{\ell_2}_\pi, a)$ in a simple manner. Let $\mathbf{z}^{\mathsf{c}_1}$ be the vector whose kth component is the predicted VAS value from the response $R'_{\mathsf{c}_1}(s^{\ell_1}_\tau, a)(\mathbf{f}^k) = (\mathsf{c}_1, \gamma_k)$, i.e., γ_k. Construct $\mathbf{z}^{\mathsf{c}_2}$ in similar fashion. The *distance between* $R'_{\mathsf{c}_1}(s^{\ell_1}_\tau, a)$ and $R'_{\mathsf{c}_2}(s^{\ell_2}_\pi, a)$ is

$$||\mathbf{z}^{\mathsf{c}_1} - \mathbf{z}^{\mathsf{c}_2}||_1 =_{\text{def}} \sum_{k=1}^{|\textsc{mvs}(a)|} abs(z^{\mathsf{c}_1}_k - z^{\mathsf{c}_2}_k) \qquad (1)$$

where $z^{\mathsf{c}_1}_k$ and $z^{\mathsf{c}_2}_k$ are the k components of $\mathbf{z}^{\mathsf{c}_1}$ and $\mathbf{z}^{\mathsf{c}_2}$, respectively, and $|\textsc{mvs}(a)|$ the cardinality of $\textsc{mvs}(a)$.

Distance between VCRSs can now be used to reason about differences in vowel category perception within a given language community (the case where $\ell_1 = \ell_2$ and $\mathsf{c}_1 = \mathsf{c}_2$), as well as differences across communities (the case where $\ell_1 \neq \ell_2$). We exemplify with an application involving a comparison of VCRS patterns concerning the point vowels [i], [a], and [u] within and across the Greek (G) and Japanese (J) language communities.

Recall the number of subjects $n_G = n_J = 21$, and let $S = \{s^\ell_\tau \mid \ell = G, J;\ 1 \leq \tau \leq 21\}$ denote the set of subjects from G and J. Assuming that $C_G = C_J = \{\mathsf{i}, \mathsf{e}, \mathsf{a}, \mathsf{o}, \mathsf{u}\}$, let $C = \{\mathsf{i}, \mathsf{u}, \mathsf{a}\}$. For each $a \in \textsc{ages}$, $\mathsf{c} \in C$, and $s^\ell_\tau \in S$, we construct VCRSs $R'_{\mathsf{c}}(s^{\ell_1}_\tau, a)$. The regression spline basis cardinality for each smooth function in each additive model was set to 3, and the response category extension parameter α was set to 0.5. Ordered pairs $(s^{\ell_1}_\tau, s^{\ell_2}_\pi)$, where $s^{\ell_1}_\tau, s^{\ell_2}_\pi \in S$, are called *cross-language pairs*, if $\ell_1 \neq \ell_2$, and *within-language pairs*, otherwise. Within-language pairs are called *Greek pairs*, if $\ell_1 = \ell_2 = G$, and *Japanese pairs*, if $\ell_1 = \ell_2 = J$. For each $a \in \textsc{ages}$, $\mathsf{c} \in C$, and let $d^a_{\mathsf{c}}(s^{\ell_1}_\tau, s^{\ell_2}_\pi)$ denote the distance between $R'_{\mathsf{c}}(s^{\ell_1}_\tau, a)$ and $R'_{\mathsf{c}}(s^{\ell_2}_\pi, a)$.

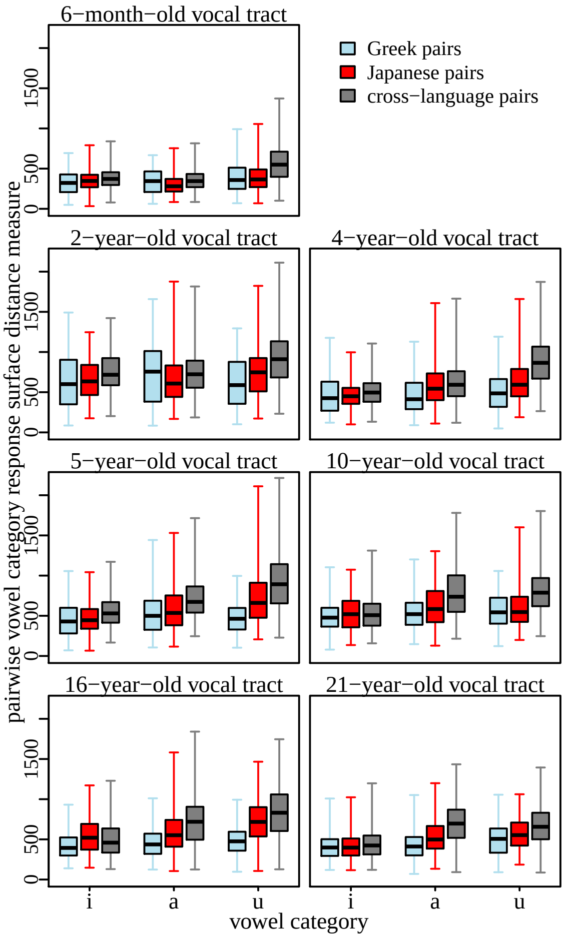

Figure 3: *Boxplots (median, interquartile range, and full range of values) summarizing distances between VCRSs over within- and cross-language pairs.*

Given what we know about the phonetic values of the point vowels [i], [a], and [u] in the two languages [27, 28, 29, 30, 31, 32], we expect the distances to be largest for d^a_{u}, since the Japanese high back vowel, unlike its Greek counterpart, is generally unrounded, lowest for d^a_{i}, since the high front vowel in both languages has a very peripheral cardinal vowel quality, and intermediate for d^a_{a} since the Greek low vowel is on average more front than the Japanese low vowel. We also expect distances over within-language pairs to be smaller in each case. For each $a \in \text{AGES}$, and $\text{c} \in C$ we computed the distances d^a_{c} for all cross-language pairs over S, and distances for all within-language pairs, excluding identity pairs $(s^\ell_\tau, s^\ell_\tau)$, whence $d^a_{\text{c}} = 0$, trivially. Boxplots (median, interquartile range, and full range of values) over distances are shown in Figure 3.

6. Discussion

Our predictions concerning patterns of distances within and across languages were mostly borne out. Generally, distances over cross-language pairs were highest for u, lowest for i, with distances for a falling in between. Moreover, except for the Greek pairs for a at ages 6 months and 2 years, median distances for within-language pairs were always smaller than median distances for cross-language pairs for the same age and vowel category.

A third, unanticipated finding was that median differences for within-Japanese pairs tended to be larger than median differences for within-Greek pairs, particularly for the a and (especially) the u categories. This tendency was more pronounced for older vocal tracts. We speculate that this difference between the two languages is related to a difference in the orthographic representation of the five vowel categories for the response in interaction with the length of the stimuli. That is, the five vowel categories for Greek were labeled with a letter or letter combination that can denote either unstressed vowels or (the somewhat longer) stressed vowels of the language, and the five categories for Japanese were labeled with the kana symbol for the short vowel, but the stimuli were all about 590 ms, which is much longer than the typical duration of vowels in both languages. This duration is especially long by comparison to the typical length of a short vowel of Japanese or an unstressed vowel of Greek [27, 28, 29, 30, 31]. Additionally, [u:] and [u] in Japanese can be rounded in clear speech [32], so the greater distance for the category label u in particular could be interpreted in terms of further variability in the interpretation of the stimuli relative to this stylistic variation.

The statistical methodology appears to be useful in revealing differences between the vowel systems of different language communities. Importantly, it was sensitive enough to pick up on the vowel category differences between two vowel systems that have roughly the same set of categories, and quantify the corresponding differences in vowel category perception. The instantiation of the methodology presented in this paper represents a "simplest possible" approach in that a very basic smoothing-spline method was used for the additive model surface regression. Moreover, the splines used in additive predictors were all univariate and had a minimum basis cardinality. Finally, the distance computation over response surfaces is simply the L_1 norm over surface values. It may be worth investigating whether complicating any of these components may be needed in the study of more complex vowel system phenomena.

7. Acknowledgements

Work supported by NSF grants BCS-0729306, BCS-0729277, and BCS-0729140. Thanks to Asimina Syrika and Kiyoko Yoneyama for recruiting and testing the Greek- and Japanese-speaking participants, to Hyunju Chung and Chanelle Mays for recruiting and testing the Korean- and the English-speaking participants, and to Catherine McBride-Chang for providing the resources to recruit the Cantonese-speaking subjects.

8. References

[1] B. Munson, L. Ménard, M. E. Beckman, J. Edwards, and H. Chung, "Sensorimotor maps and vowel development in English, Greek, and Korean: A cross-linguistic perceptual categorizaton study (A)," *Journal of the Acoustical Society of America*, vol. 127, p. 2018, 2010.

[2] L.-J. Boë and S. Maeda, "Modélization de la croissance du conduit vocal. Éspace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse," in *Journée d'Études Linguistiques: "La Voyelle dans Tous ces États"*, Nantes, France, 1997, pp. 98–105.

[3] L.-J. Boë, P. Perrier, B. Guérin, and J.-L. Schwartz, "Maximal vowel space," in *EUROSPEECH 09*, Paris, France, 1989, pp. 281–284.

[4] J.-L. Schwartz, L.-J. Boë, and C. Abry, "Linking dispersion-focalization theory and the maximum utilization of the available distinctive features principle in a perception-for-action-control

theory," in *Experimental Approaches to Phonology*, . M. O. M.-J. Sole, P. S. Beddor, Ed. Oxford University Press, 2007, pp. 104–124.

[5] J. L. Miller, "On the internal structure of phonetic categories: A progress report." *Cognition*, vol. 50, 1994.

[6] J. L. Miller, "Internal structure of phonetic categories," *Language and cognitive processes*, vol. 12, 1997.

[7] G. Wahba, *Spline models for observational data*. Society for Industrial Mathematics, 1990, vol. 59.

[8] C. Gu, *Smoothing spline ANOVA models*. Springer, 2002.

[9] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.

[10] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models," *The Annals of Statistics*, pp. 453–510, 1989.

[11] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. New York: Chapman and Hall, 1990.

[12] S. Wood, *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2006, vol. 66.

[13] J. Lindblom and J. E. F. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement," *Journal of the Acoustical Society of America*, vol. 50, pp. 1166–179, 1971.

[14] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal, Eds. The Netherlands: Kluwer Academic Publishers, 1990, pp. 131–149.

[15] S. Maeda, "On articulatory and acoustic variabilities," *Journal of Phonetics*, vol. 19, pp. 321–331, 1991.

[16] J. M. Beck, "Organic variation of the vocal apparatus," in *Handbook of Phonetic Sciences*. Cambridge, England: Blackwell, 1996, pp. 256–297.

[17] U. G. Goldstein, "An articulatory model of the vocal tract of the growing child," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[18] L. Ménard, J.-L. Schwartz, and L.-J. Boë, "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1892–1905, 2002.

[19] N. Vallée, L.-J. Boë, and Y. Payan, "Vowel prototypes for UPSID's 33 phonemes," in *Proceedings of ICPhS 2*, Stockholm, 1995, pp. 424–427.

[20] L. Ménard and L.-J. Boë, "Exploring vowel production strategies from infant to adult by means of articulatory inversion of formant data," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 2000, pp. 465–468.

[21] J.-L. Schwartz, L.-J. Boë, N. Vallée, and C. Abry, "The dispersion-focalization theory of vowel systems," *Journal of Phonetics*, vol. 25, pp. 255–286, 1997.

[22] J.-L. Schwartz, L.-J. Boë, N. Vallée, and C. Abry, "Major trends in vowel system inventories," *Journal of Phonetics*, vol. 25, pp. 233–253, 1997.

[23] D. W. Massaro and M. M. Cohen, "Categorical or continuous speech perception: A new test," *Speech Communication*, vol. 2, pp. 15–35, 1983.

[24] A. R. Plummer, "Manifold alignment, vocal imitation, and the perceptual magnet effect," in *The Annual International Child Phonology Conference (ICPC 2012)*, Minneapolis, MN, 2012.

[25] T. D. Wickens, *Elementary Signal Detection Theory*. Oxford University Press, USA, 2002.

[26] S. N. Wood, "Thin-plate regression splines," *Journal of the Royal Statistical Society (B)*, vol. 65, no. 1, pp. 95–114, 2003.

[27] P. A. Keating and M. K. Huffman, "Vowel variation in Japanese," *Phonetica*, vol. 41, no. 4, pp. 191–207, 1984.

[28] Y. Hirata and K. Tsukada, "Effects of speaking rate and vowel length on formant frequency displacement in Japanese," *Phonetica*, vol. 66, no. 3, pp. 129–149, 2009.

[29] M. Fourakis, A. Botinis, and M. Katsaiti, "Acoustic characteristics of Greek vowels," *Phonetica*, vol. 56, no. 1-2, pp. 28–43, 1999.

[30] J. W. Hawks and M. S. Fourakis, "The perceptual vowel spaces of American English and Modern Greek: A comparison," *Language and speech*, vol. 38, no. 3, pp. 237–252, 1995.

[31] A. Jongman, M. Fourakis, and J. A. Sereno, "The acoustic vowel space of Modern Greek and German," *Language and Speech*, vol. 32, no. 3, pp. 221–248, 1989.

[32] H. Okada, "Japanese," in *Handbook of the International phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge, U.K.; New York, New York: Cambridge University Press, 1999, pp. 117–119.