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).

2. Vowel perception experiments

The VLAM [2] is a computational model of the articulatory system and its speech production capacities. Mid sagittal 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, 5, 10, 16, 21\} for indexing the MVSs discussed below. For each $a \in$ AGES,
let 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 cm² for ages 2 and over, and 0.15 cm² for the 6 m.-o., in accordance with previous modeling [18], and lip area was constrained from 0.1 cm² to 8 cm² for all ages. Each 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 ∈ 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 ∈ AGES, a set of 38 prototypes, denoted by P(a), were selected from MVS(a). Prototypes in P(a) are indexed p, 1 ≤ i ≤ 38, though the superscript is often dropped. The set P(10) of prototypes for age 10 is depicted within 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), and Korean (n=20) for all 38 prototypes in 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 P(a). Consider, say, subject 12 from the Greek language community (G), which 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).

and α = 0.5. Then γ_0(r) = (1, 0.9), while γ_0(r) = (0, 0.05). We can then compose each γ_0 with R(s_{12}^a, a) to obtain functions

\[ R_\ell(s_{12}^a, a) = \text{def} \gamma_0 \circ R(s_{12}^a, a) : P(a) \rightarrow C_{\ell} \times \text{VAS} \]

such that, for each c ∈ C_{\ell}, each p ∈ P(a) is mapped to a response reflecting its goodness as an example of c.

Let LANGLE = \{C, E, G, J, K\} be an index set over denotations of the five language communities. Given a language community \ell \in LANGLE, let \ell denote the set of vowel categories for \ell. Let \ell denote the number of subjects for \ell. Given an age a ∈ AGES, a subject s_{12}^a, where 1 ≤ r ≤ \ell, and a category c ∈ C_{\ell}, the function \ell(s_{12}^a) is called an individual category response function for s_{12}^a at age a, or simply an individual category response function (ICRF). We construct ICRFs for each subject and category from each language community in LANGLE, for each age a ∈ AGES.

4. Vowel Category Response Surfaces

In this section, we put forward a formal method for extending the domain of an ICRF \ell(s_{12}^a) from P(a) to all of MVS(a). The basic idea is to use a regression technique to construct a “surface” of responses over MVS(a) using \ell(s_{12}^a). The response surface value for a formant vector f ∈ MVS(a) is meant to approximate s_{12}^a’s VAS goodness rating of 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 y = (y_1, y_2, ..., y_n)^T, and design variables X with observed design vectors: X_i = ([x_{i1}, x_{i2}, ..., x_{im}])^T, where 1 ≤ j ≤ p. Design vectors are arranged in a matrix X = (x_1, x_2, ..., x_n), whose rows are denoted x_i, 1 ≤ i ≤ n. We are interested in deriving a fit \hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T such that y = \hat{y} + \epsilon, for residuals \epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T, that bears a smooth relationship to the x_i, 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 \hat{y} = \sum_{j=1}^p g_j(X_j).

To illustrate, consider the univariate case of estimating a smooth function \hat{g} from the n observations (y_i, x_i) such that \hat{g} = g(x_i) + \epsilon_i, where \epsilon_i is a random error term. We can estimate \hat{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 \lambda, which can be estimated along with \hat{g}, controls the trade off between these terms: as \lambda \rightarrow 0 the fit approaches a straight line, while \lambda = 0 yields an unpenalized regression spline estimate. The operator J_{md} has the general form:

\[ \int \sum_{x_{md} = d} \frac{m!}{m_1! \cdots m_d!} \left( \frac{\partial^m f}{\partial x_{1}^{m_1} \cdots \partial x_{d}^{m_d}} \right)^2 dx_1 \cdots dx_d, \]

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
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 MVS(a), let \( F^n_1, F^n_2, \) and \( F^n_3 \) be variables whose observed values \( f^n_1, f^n_2, \) and \( f^n_3 \) constitute the first, second, and third components, respectively, of the formant vectors in MVS(a). Let \( f^1, f^2, f^3 \) be the vectors of observed values for \( F^n_1, F^n_2, \) and \( F^n_3 \), respectively. Form the data matrix \( F^n = (f^1 f^2 f^3) \). Similarly, form a data matrix \( P^n \) from the rows of \( F^n \) that correspond to the prototypes in \( p(a) \). Thus each formant vector in MVS(a) is indexed by its row position in \( F^n \), and each prototype in \( P^n \) by its row position in \( P^n \). Let \( P^n_1, P^n_2, \) and \( P^n_3 \) be variables whose observed values are the first, second, and third columns of \( P^n \).

Given an ICRF \( R_c(s_1^a, a) \), for each \( p^1 \in p(a) \), let \( \gamma_i \) be the second component of the response \( R_c(s_1^a, a)(p^1) = (c, \gamma) \), i.e., \( \gamma = \gamma_i \), and let \( y_k = (\gamma_1, \ldots, \gamma_{88})^T \). We can now estimate smooth functions \( g_1(P^n_1), g_2(P^n_2), \) and \( g_3(P^n_3) \) for an additive predictor \( \sum_{j=1}^{3} g_j(P^n_j) \) using \( p(a) \) as a data matrix and \( y_k \) as response vector, yielding an additive model \( y_k = \sum_{j=1}^{3} g_j(P^n_j) + \epsilon \). More importantly, we can now predict category goodness ratings for each \( f \in MVS(a) \) by applying the additive model to \( F^n \). Given \( F^n \) in \( MVS(a) \), let \( \gamma_k \) be its predicted goodness value under the additive model derived from \( R_c(s_1^a, a) \) and \( y_k \), and let \( r^k = (c, \gamma_k) \). Pairing each \( F^n \) with \( r^k \), we can define

\[
R_c(s_1^a, a) : MVS(a) \rightarrow C_x \times VAS,
\]

which approximates an extension of \( R_c(s_1^a, a) \) from \( p(a) \) to all of \( MVS(a) \).

Given an age \( a \in AGES \), a category \( c \in C \), and a subject \( s_1^a \), where \( 1 \leq \tau \leq n_a \), the function \( R_c(s_1^a, a) \) is called a vowel category response surface for \( s_1^a \) at age \( a \), or simply a vowel category response surface (VCRS). Figure 2 depicts the VCRSSs \( R_c(s_1^a, 10) \) and \( R_c(s_2^a, 10) \) where the response category extension parameter \( \alpha = 0.5 \).

![Figure 2: Vowel category response surfaces for \( R_c(s_2^a, 10) \) (left) and \( R_c(s_2^a, 10) \) (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 \( a \).](image-url)

5. Comparing surfaces

We now define a method for comparing VCRS patterns within and across language communities. Let \( \ell_1, \ell_2 \in \text{LANG} \). Given subjects \( s_1^{\ell_1} \) and \( s_1^{\ell_2} \), and categories \( c_1 \in C_{\ell_1} \) and \( c_2 \in C_{\ell_2} \), consider the corresponding VCRSSs \( R_{c_1}(s_1^{\ell_1}, a) \) and \( R_{c_2}(s_1^{\ell_2}, a) \). We begin by defining similarity between \( R_{c_1}(s_1^{\ell_1}, a) \) and \( R_{c_2}(s_1^{\ell_2}, a) \) in a simple manner. Let \( \mathbf{z}^1 \) be the vector whose \( k \)th component is the predicted VAS value from the response \( R_{c_1}(s_1^{\ell_1}, a) (\mathbf{F}) = (c_1, \gamma_k) \), i.e., \( \gamma_k \). Construct \( \mathbf{z}^2 \) in similar fashion. The distance between \( R_{c_1}(s_1^{\ell_1}, a) \) and \( R_{c_2}(s_1^{\ell_2}, a) \) is

\[
||\mathbf{z}^1 - \mathbf{z}^2||_1 = \alpha_{k} \sum_{k=1}^{|\text{MVS}(a)|} \text{abs}(z_1^k - z_2^k) \quad (1)
\]

where \( z_1^k \) and \( z_2^k \) are the \( k \) components of \( \mathbf{z}^1 \) and \( \mathbf{z}^2 \), respectively, and \( |\text{MVS}(a)| \) the cardinality of \( \text{MVS}(a) \).

Distance between VCRSs can now be used to reason about differences in vowel category perception within a language community (the case where \( \ell_1 = \ell_2 \) and \( c_1 = c_2 \)), 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 \([i, a, u] \) within and across the Greek (G) and Japanese (J) language communities.

Recall the number of subjects \( n_{c} = n_{\tau} = 21 \), and let \( S = \{ s_1^\tau \mid \ell = G, J; 1 \leq \tau \leq 21 \} \) denote the set of subjects from G and J. Assuming that \( C_{G} = C_{J} = \{ i, o, a, u \} \), let \( C = \{ i, u, a \} \). For each \( a \in \text{AGES}, c \in C \), and \( s_1^a \in S \), we construct VCRSs \( R_{c}(s_1^a, a) \). The regression spline basis cardinality for each smooth function in each additive model was set to 3, and the response category extension parameter \( \alpha \) was set to 0.5. Ordered pairs \( (s_1^a, s_2^a) \), where \( s_1^a, s_2^a \in S \), 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 \( a \in \text{AGES}, c \in C \), and let \( d_{a}(s_1^a, s_2^a) \) denote the distance between \( R_{c}(s_1^a, a) \) and \( R_{c}(s_2^a, a) \).
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 \( d_u \), since the Japanese high back vowel, unlike its Greek counterpart, is generally unrounded, lowest for \( d_a \), since the high front vowel in both languages has a very peripheral cardinal vowel quality, and intermediate for \( d_i \) 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 \( c \in C \) we computed the distances \( d_a^c \) for all cross-language pairs over \( S \), and distances for all within-language pairs, excluding identity pairs \( (s_i^c, s_j^c) \), whence \( d_a^c = 0 \), trivially. Boxplots (median, interquartile range, and full range of values) over distances are shown in Figure 3.

![Boxplots](image)

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

### 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 \( a \) (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 Kyoko 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


