Reduction of vowel chaos

Lindgren, R. and Lindblom, B.

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 001-004

http://www.speech.kth.se/qpsr
Reduction of vowel chaos

Rolf Lindgren and Björn Lindblom
Department of Phonetics, Stockholm University

Abstract
A previous study showed that classifying vowel data from spontaneous speech with statistical clustering methods and artificial neural networks gave poor results.

In this study a different approach was tested. A production model (the undershoot model) was used to explain the large variation of F2 in vowel data from spontaneous speech. Expanding the model to also include regressive assimilation and formant peak velocity showed a significant improvement in predicting F2 values.

Background
The project "Speech Transforms" has as its object to study the acoustic-phonetic variation in spontaneous speech and to formulate transform rules relating the pronunciations of words and phrases in reference speech (as in a standard laboratory recording) to their variants in spontaneous speech. This enormous task has of course been only partly covered and this report deals with some aspects of the study of the vowel system in spontaneous speech.

The problem at hand is illustrated by figure 1, which shows the vowel systems of one speaker in spontaneous and reference speech. The dramatic overlap of different vowel categories is quite obvious. This "system collapse" is also, mutatis mutandis, transferable to other domains in the speech signal, when comparing to reference speech.

Figure 1. Vowel space of one speaker from spontaneous speech (left) and reference speech (right). Each ellipse represents the variation of F1 and F2 for each vowel and covers 95% of the variation.

Vowel reduction
Numerous investigations have shown a speaker's vowel system in read speech to have a large, but systematic variation, governed primarily by contextual factors. As shown in figure 1 the variation in spontaneous speech is extremely large. Often the term 'vowel reduction' is used for this variation, since the formants of the vowels seem to be "reduced" towards a central point in the vowel space. As can be seen in figure 2 this also seems to be the case for the average of F1/F2 of the vowels in our spontaneous speech material. The most common explanation for this 'centralisation' being variations on the theme that the consonants have a 'centralising effect', i.e. However, this doesn't account for the many cases where the formant values actually go away from this central point. So, the term 'vowel reduction' in this report should be interpreted as meaning 'a vowel deviating from its (theoretical) formant target'.

Figure 2. The mean of F1/F2 for each vowel in ref. speech (larger dots) connected to the mean of the corresponding vowel in spontaneous speech (smaller dots).
Classification of vowel data

Given the vowel data in figure 1, it seems like an impossible task to identify any F1/F2-point as belonging to a specific vowel. One may object that a F1/F2 plot is too simple a graph to separate the different vowel areas. But adding F3 and showing the vowel variations as volumes in a 3-D space, gives much the same result as the 2-D ellipse graph.

Artificial Neural Networks Classifiers

In a previous study (Lindgren & Eismann, 1994) we tried to resolve the vowel overlap using Artificial Neural Networks (ANN) of two different types: supervised (a back-propagation net) and unsupervised (a Kohonen Feature Map). If there are hidden patterns in the formant data, even if not revealed by statistical cluster methods, an ANN should be able to find them. However, in an exhaustive series of experiments with ANN, the amount of correctly identified vowels varied between 42-57% only. This should be compared to the 97% correct classification of vowels from reference speech. This led to the conclusion that we needed to find other parameters than formants to feed the ANN or to add more formant-related data.

Explanation of vowel data

A problem with ANN is that, even if they are very efficient making complex transformations from one set of data to another, it is often difficult to interpret the inner workings of the trained ANN and it may therefore be impossible to estimate the importance of certain parameters in detecting hidden patterns.

Another approach is to find a explanation for the variation in formant patterns via a speech production model. The advantage of this approach is that it can be independently motivated by the biomechanics of the articulators and not being a simple curve fitting or an unknown transform function.

Bio-mechanics

We decided to run a series of experiments, starting with an empirically validated production model in its most basic form, and then expanding it if necessary in trying to fit the data from spontaneous speech. The most obvious model to use in this particular case is the "undershoot model" of Stevens & House (1963) and Lindblom (1963, 1968). This model was developed to explain the variation in vowel formant patterns as a systematic shift away from a hypothetical target. A mathematical model, based on the bio-mechanics of articulation, was fitted to observed data demonstrating that the variation (or deviation from target) was dependent primarily on two factors: vowel duration and the extent of the CV transition. This model could also handle the cases where the formant does not 'centralise', but drifts away from the centre of the vowel space.

The basis for the mathematical model is that several investigations (e.g. Henke, 1966; Lindblom, 1967; Ohman, 1967; Perkell, 1974; Saltzman & Munhall, 1989) have made successful attempts to model articulatory trajectories in terms of critically damped mechanical system.

The maximum excursion ($x_{max}$) of such a system is dependent on: (1) $A$; the extent of the input force, (2) $D$; the duration of input force and (3) $\alpha$; the systems time constant:

$$x_{max} = A \{exp(\alpha D/(1-e^{\alpha D})) - \exp(\alpha D/(1-e^{\alpha D})) \}$$

If we assume that there is a straightforward relationship between the F2 and tongue movement, then the variation of F2 could be described in terms of a critically damped mechanical system. In Moon & Lindblom (1994) the analogy between a mechanical system and the speech production system is made explicit, in that $x_{max}$ and the maximum excursion of F2 is compared and $A = \text{locus-target distance}$, $\alpha = \text{F2 rate of change}$ and $D = \text{vowel duration}$. This means that the excursion of F2 is dependent upon one or more of these factors.

Model 1: duration dependency only

Not to complicate things unnecessarily, we begin with the simple assumption that only duration would have an effect on F2. Applying this to eq. 1 allows a simplification calculating the predicted value of F2 ($F_{2c1}$):

$$F_{2c1} = a * (F_{2i} - T) * e^{\alpha D} + T,$$

where $F_{2i}$ is the F2 value at the initial part of the vowel ("locus"), $T$ is the underlying target and $D$ is the vowel duration. Except for $T$, these values are all known from measurements. In order to get 'a' and 'x', eq. 2 is rewritten as:

$$\ln[(F_{2c1}-T)/(F_{2i}-T)] = \ln(a) - \alpha D$$

Using linear regression, 'a' and 'x' can be calculated since the left hand side is a linear function of D. The target T has somehow to be estimated.
Target estimation

According to eq. 2 the speaker should converge on the target, given that the vowel duration is long enough. So one approach to determine target is to simply consider target equal to \( F_2 \), i.e. \( F_2 \) at mid-point in the vowel) in vowels with long duration. However, this seems to anticipate the answer to the question we ask: does \( F_2 \) vary between \( F_2_i \) and the asymptote \( F_2_m \)?

Another approach was to calculate \( F_2c_1 \) according to eq. 2, letting target vary between a certain range, and picking as target the value that gave the best fit with observed \( F_2_m \). This method could not be used consistently, since in many cases the best fit was found with targets at very odd values (e.g. \( F_2 \) of [i:] at 1300 Hz) and the choice of a reasonable target value seemed arbitrary.

Finally we decided to simply use as target observed \( F_2_m \) values from reference speech, similar to Moon & Lindblom (1974). This choice should be considered only tentative as the question of target has yet not been resolved. It could be defended as being a rather strict test of the model.

Types of transitions

To calculate ‘a’ and ‘\( \alpha \)’, \( (F_2_m - T)/(F_2_r - T) \) must be > 0, because of the logarithm. It should also be \( \leq 1 \); otherwise it means that \( F_2 \) drifts away from target, i.e. the original ‘undershoot’ model covers transitions only of the following types:

1. \( F_2_i > F_2_m \) and \( T < F_2_m \)
2. \( F_2_i < F_2_m \) and \( T > F_2_m \)

In the spontaneous speech several other types of transitions were found (see figure 3), that were excluded when estimating ‘a’ and ‘\( \alpha \)’ in linear regression. This means that only a subset of each vowel was used to calculate ‘a’ and ‘\( \alpha \)’.

The results from using Model 1 shows a rather large spread when correlating the observed \( F_2 (=F_2_m) \) and the predicted \( F_2 (=F_2c_1) \), e.g. \( r = 0.15 \) for [a] and \( r = 0.71 \) for [e]. For detailed results, see Table 1.

Model 2: considering final locus

The ambiguous results from the experiment with Model 1, showed that it was necessary to add yet another component to the model. Model 1 as expressed with eq. 2, takes into consideration only the initial transition. It seemed logical to also consider the final transition, since the speaker should not only move from \( F_2_i \) towards \( T \), but also from the observed \( F_2_m \) to the final locus, \( F_2_f \), i.e. a regressive assimilation. Expanding the model to include also final transitions, means that the possible types of transitions (c.f. figure 3) is increased. In order to handle this, model 2 assumes that the formant movement during the vowel is composed of two parts: \( F_2_i \rightarrow F_2_m \) and \( F_2_m \rightarrow F_2_f \), which are added together. Assuming also that the final transition can be modelled in terms of a critically damped system, it should be possible to predict \( F_2 \) in this model with:

\[
\text{Eq. 4) } F_{2c2} = a \ast (F_2_r - T) \ast e^{\alpha \Delta} + b \ast (F_2_i - T) \ast e^{\beta \Delta} + T
\]

The terms ‘\( b \)’ and ‘\( \beta \)’ for the final transition corresponds to ‘a’ and ‘\( \alpha \)’ in the initial transition and can be estimated with linear regression in the same way:

\[
\text{Eq. 5) } \ln\left[ \frac{(F_2_m - T)}{(F_2_r - T)} \right] = \ln(b) - \beta \Delta
\]

Figure 3. Schematic illustration of transitions from initial to middle part of vowel. Only transitions falling into type 1a or 2a are covered by original undershoot model.
Using eq. 4 improved the correlation between observed F2 (=F2a) and the predicted F2 (=F2d) for all vowels, but most for [a], from 0.15 to 0.44. The spread was also diminished. See Table 1 for detailed results.

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>0.55</td>
<td>0.69</td>
<td>0.84</td>
</tr>
<tr>
<td>[e]</td>
<td>0.71</td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td>[i]</td>
<td>0.60</td>
<td>0.74</td>
<td>0.82</td>
</tr>
<tr>
<td>[u]</td>
<td>0.65</td>
<td>0.81</td>
<td>0.87</td>
</tr>
<tr>
<td>[ø]</td>
<td>0.15</td>
<td>0.44</td>
<td>0.70</td>
</tr>
<tr>
<td>[y]</td>
<td>0.76</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>[o]</td>
<td>0.59</td>
<td>0.65</td>
<td>0.82</td>
</tr>
<tr>
<td>[ø]</td>
<td>0.45</td>
<td>0.49</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1. Correlation between observed and predicted F2 values. The columns show the results for the used models for each vowel (rows).

Model 3: adding velocity of initial transition

In a critically damped system the maximum excursion of the system is determined not only by duration, but also by the extent of the input force (term ‘A’ in eq. 1). How then could we estimate force? Nelson (1983, 1984) has suggested that ‘articulatory effort’ could be measured by the ‘peak velocity’ of a system response. Several authors (e.g. Kuehn & Moll, 1976; Fliege, 1988) have pointed to transition velocity as an important factor in undershoot modelling. Moon & Lindblom (1994) noticed that the transition velocity is to some extent determined by the locus-target distance, i.e. the greater the distance |F2j - Tj|, the greater the velocity. If we assume that F2 velocity is used to minimise undershoot (i.e. making an articulatory effort to reach target), at least where |F2j - Tj| is large, then perhaps also F2 velocity should be included in the equation for predicting F2. The time constant of the system is already expressed with ‘a’ and ‘b’, which gives a rough estimate of the slope of the transition. The peak velocity could then be used to “fine tune” every single predicted F2 value.

Peak velocity in this case (F2v) was measured via a simple differentiation of the first 3 analysis frames in the initial F2 transition for each vowel.

F2v showed a moderate correlation (approx. 0.5-0.6) with the prediction error (F2er-F2d). Estimating the coefficient (k) and intercept (l) for F2v with linear regression, the velocity effect was added to the model, resulting in eq. 6:

(Eq. 6) F2v = a*(F2j - Tj) + e + b*(F2v - Tj) + e + c + T + (kF2d + l)

Using eq. 6 improved the correlation between observed F2 (=F2a) and the predicted F2 (=F2d) for all vowels, but most for [a] (from 0.44 to 0.68) and for [ø] (from 0.49 to 0.74). The spread was also diminished. See Table 1 for detailed results.

The results from this production oriented modelling of vowel variation is rather encouraging. An interesting continuation should be to feed a neural network with the parameters used here, too see if any improvement in recognition could be achieved.

References


