CS-BIGS 8:28-46 http://www.csbigs.fr

Modelling repeated paired phonetic measures using linear mixed models with correlated errors

Frédérique Letué
Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France, CNRS, LJK, F-38000 Grenoble, France

Marie-José Martinez
Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France, CNRS, LJK, F-38000 Grenoble, France

Sandra Cornaz
Univ. Grenoble Alpes, CNRS, GIPSA-Lab, F-38000 Grenoble, France

Nathalie Vallée
Univ. Grenoble Alpes, CNRS, GIPSA-Lab, F-38000 Grenoble, France

Nathalie Henrich Bernardoni
Univ. Grenoble Alpes, CNRS, GIPSA-Lab, F-38000 Grenoble, France

In Phonetic Sciences, statistical analysis from experimental data have to be carried out to confirm or disconfirm hypotheses. In this paper, a phonetic data set is considered and phonetic research questions are addressed. To answer these questions, a mixed model is built using a complex random effects structure and a non-diagonal residual variance-covariance matrix. Then, it is validated on the data. Finally, we focus on statistical tests in the final model allowing to compare the means between two groups of subjects, and a single mean to a reference value. The paper is accessible to an audience experienced with linear models. Some familiarity with the R software is also helpful.

Keywords *: Linear mixed models, repeated paired data, correlated errors, statistical tests,* `nlme` *and* `multcomp` *R libraries, phonetic data set.*

1. Introduction

In Phonetic Sciences, research is mostly based on experimental data to confirm or disconfirm hypotheses. For this purpose, a statistical analysis has to be carried out. The classical statistical approach consists of (i) modelling the data using an adapted model, (ii) validating the selected model, and (iii) testing statistical hypotheses to confirm or not the phonetic hypotheses.

To model the data, analysis of variance (ANOVA) is often used to explain one continuous response such as discrimination scores, detection of phonetic contrasts or boundaries (e.g. for the voicing feature), phoneme categorization, acoustic parameters (e.g. segment, syllable, and word durations, VOT, formant frequencies, pitch peak, amplitudes of harmonics,...) with respect to different experimental conditions (e.g. Kuhl et al. (1997)). Such a model assumes that the data are independent and the variance of the observations remains the same from an experimental condition to another. In studies where subjects contribute to more than one measure, the ANOVA assumption of data independence is not valid and repeated measures ANOVA may be used (e.g. Hazan and Barett (2000)). These models allow to take into account the within-subject effects. This is only valid with one factor of interest and the same number of measures per subject.

The need for advanced statistical tools in Phonetic Sciences has been recently highlighted (Bergmann et al., 2016; Roettger et al., 2019). Taking into account the variability of the response among the different individuals calls for advanced statistical approaches, such as linear mixed-effects models (Baayen et al., 2008). Moreover, in some studies, the variable of interest is not a single continuous response, but several non-independent responses. This is the case, for instance, of vowel formants. They cannot be considered as independent measurements, as they are related to vocal-tract geometry and boundary conditions. In such a case, the data modelling has to take into account the dependence between the measured formants. This can be done by introducing complex residual variance-covariance structures (Bazzoli et al., 2015). In this paper, an example of phonetic data which needs to be modelled using both a complex random effects structure and a non-diagonal residual variance-covariance matrix is considered.

Once an adapted model has been built, a model validation step is required. This important validation step is done using diagnostic plots, and in particular, residual plots are considered.

After being validated on the data, the statistical model can be used to provide answers to the scientific research questions addressed by a given study. Here, we focus on statistical tests allowing to compare the means between two groups of subjects, and a single mean to a reference value. The methodological approach is tested on a database gathered during a phonetic study which aimed at understanding how Italian native-speakers interact with the perception and the production of French as a second language (FSL) (Cornaz, 2014). Here, we focus on the realization of the French vowels /y/ and /ø/ which do not exist in the learners' native phonological systems (concerning both Italian as an official and school language and Italian native dialects).

In Section 2 details are given about the data set and research questions. Section 3 describes the statistical methodology used to fit the data set. Linear mixed-effects models with growing complexity are first elaborated, and the one that best fits the data is selected. Then the model is validated, and statistical tests are performed in the chosen model in order to answer the phonetic questions. All analyses in the paper have been performed with the R software.

2. Data set

Observations were collected in the Italian Piedmont geographic area, with fifteen Italian native speakers. They followed an 8-hours pronunciation training of FSL including phonetic correction practice. For the purpose of our statistical study, we observe the production of six women. The objective was to understand how acquisition of new phonemes (due to the lack of oral vowels /y/ and /ø/ for Italian speakers) transforms and modifies the learners' acoustic vowel space. The phonetical-acquisition assessment was twofold: (1) evaluation of how learners produce the two high front (palatal) rounded vowel of French /y/ and / ø/ before attending the course; (2) comparison with their phonetic realizations after training. In particular, formant-distance measures between vowels were addressed for each learner (F_f(before)-F_f(after) for $f = 1, \ldots, 4$). Studies have shown that focal spectral patterns due to formant frequency convergence (or focalization) induce well-defined spectral prominences which consequently increase the acoustic-perceptual salience of vowels and give rise to stable percepts (Schwartz et al., 2005). It was also demonstrated that spectral focalization plays a role in shaping the structure of vowel phoneme inventories (Schwartz et al., 1997). Therefore, the intra-vowel distance between F1 and F2, F2 and F3, and F3 and F4 were also measured and compared before and after training.

The data set contains formant values of French vowels produced by learners before and after training of FSL with phonetic correction. The formant values were computed on each segment as the average of five measurement points (located at 12%, 25%, 50%, 75%, 88% of the segment duration). The measured vowels included cardinal vowels ([i], [a], [u]), close-mid vowels ([e], [o]), and non-native target anterior vowels ([y], [ø]). Wieling (2018) suggests to use the five measurement points rather than their mean in order to model general patterns over dynamically varying data. In our paper, we follow a more classical approach based on linear mixed models in order to focus on the previously cited sources of variability (within-individual repeated measures and dependency between the measured formants).
The individual boxplots of each formant and each formant-distance before and after training are displayed in Figures 1 and 2 for non-native vowels /y/ and /ø/ respectively. Another possible data visualization would be univariate scatterplots as proposed in Politzer-Ahles and Piccinini (2018).
For each non-native vowel (/y/ and /ø/), the following questions were addressed:

- Q1. Do formants already achieve the French reference value before training?
- Q2. Do formants achieve the French reference value after training?
- Q3. Are formants similar before and after training?
- Q4. Is focalization before training already similar to that of French front vowels?
- Q5. Is focalization after training similar to that of French front vowels?
- Q6. Are distances between successive formants similar before and after training?

3. Method

To answer these questions, the following three different steps are fulfilled: (i) Data modelling taking into account the within-individual repeated measurements and the dependence between the formants; (ii) Model validation; (iii) Statistical tests in the selected model in order to answer the phonetic questions. Analyses are performed with the R software. More precisely, we use the `lme` function in the `nlme` library (Pinheiro et al., 2014) for the data modelling step and the `glht` function in the `multcomp` library (Hothorn et al., 2008) for the statistical tests step.
In this section, the methodology for the non-native vowel /y/ is presented and detailed. The same methodology has been applied to vowel /ø/ and is described more briefly afterwards.

3.1. Data modelling

For the purpose of data modelling, the data is fitted using linear mixed-effects models with complex random-effects structures and complex variance-covariance matrices of the error.

3.1.1 Modelling the random effects structure

Following Bazzoli et al. (2015), we first fit the model M_0 given in Equation (1):

$$\begin{bmatrix} F_{1sik} \\ F_{2sik} \\ F_{3sik} \\ F_{4sik} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1s} \\ \gamma_{2s} \\ \gamma_{3s} \\ \gamma_{4s} \end{bmatrix} + \xi_i \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1sik} \\ \varepsilon_{2sik} \\ \varepsilon_{3sik} \\ \varepsilon_{4sik} \end{bmatrix} \tag{1}$$

where F_{fsik} is the k^{th} measure of formant F_f for individual i and stage $s \in \{Before, After\}$, μ is the mean for formant F_1 and stage $Before$, α_f is the fixed effect of formant F_f (with $\alpha_1 = 0$), β_s is the fixed effect of stage s (with $\beta_{Before} = 0$), γ_{fs} is the interaction between formant F_f and

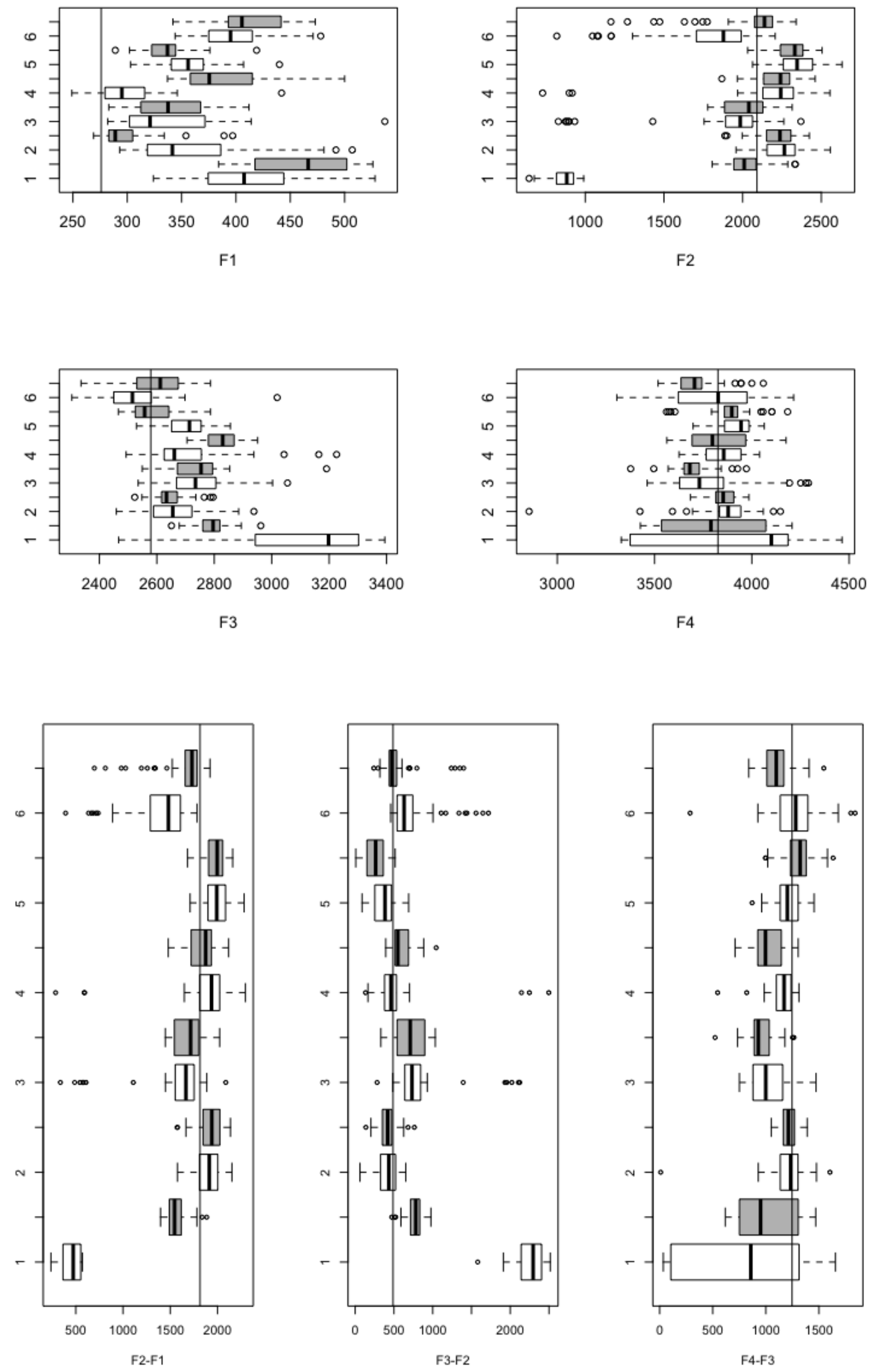

Figure 1: Individual boxplots of each formant and each formant-distance before (white) and after (grey) training for non-native vowel /y/. Solid lines correspond to reference values in the French language (Georgeton and colleagues (2012).

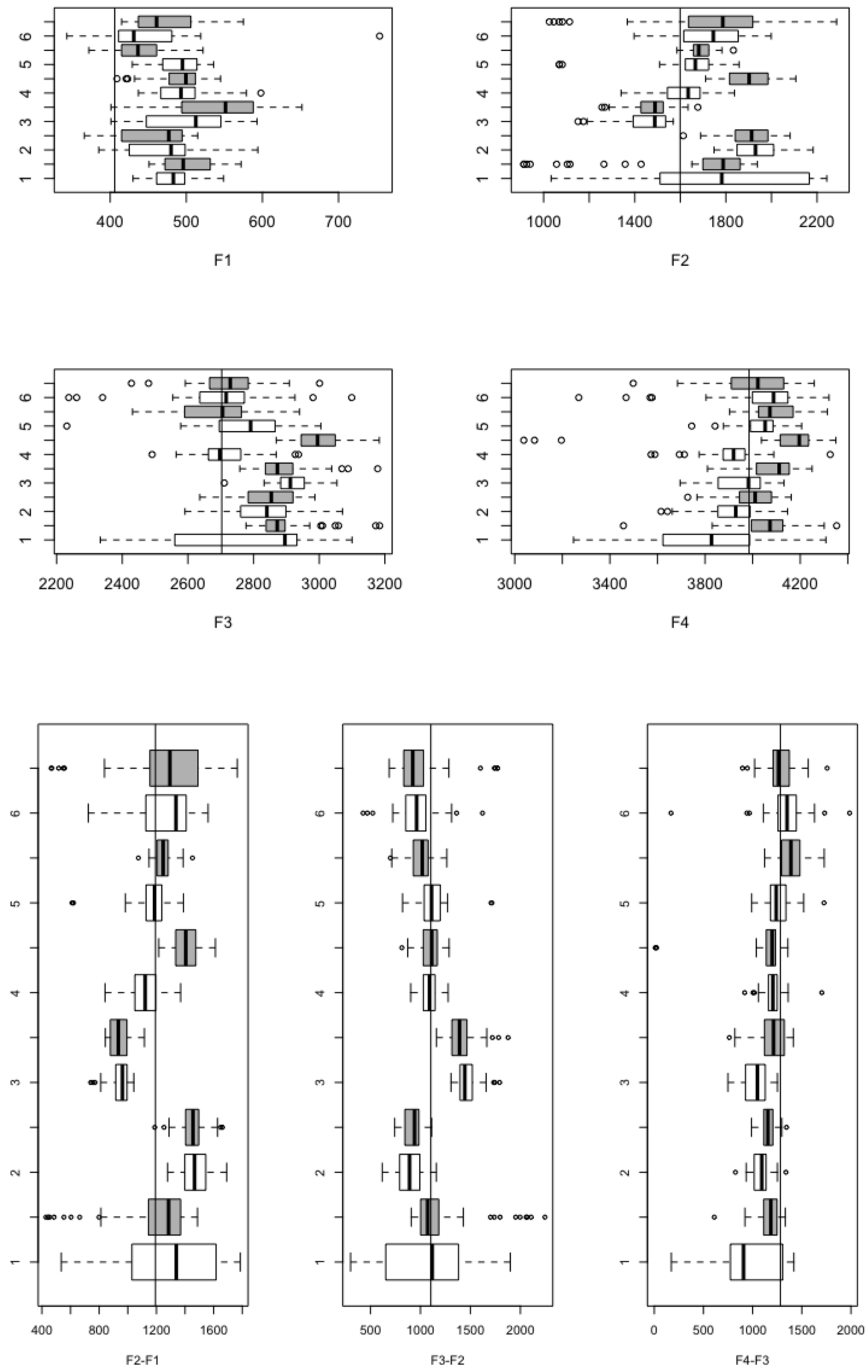

Figure 2: Individual boxplots of each formant and each formant-distance before (white) and after (grey) training for non-native vowel /ø/. Solid lines correspond to reference values in the French language (Georgeton and colleagues (2012).

stage s (with $\gamma_{1s} = 0$ for $s \in \{Before, After\}$ and $\gamma_{fBefore} = 0$ for $f \in \{1, 2, 3, 4\}$), ξ_i is the individual random effect and ε_{fsik} is the residual error. The random effect ξ_i and the residual error ε_{fsik} are supposed to be normally distributed, centered, with respective variances τ^2 and σ^2. All random effects are assumed independent from each other and independent from the error term. All residual errors are supposed to be independent. With this model, the mean measures for each formant and stage are given in Table 1.

Table 1: Mean measures for each formant and each stage.

	Stage	
Formants	**Before**	**After**
F1	μ	$\mu + \beta_{After}$
F2	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_{After} + \gamma_{2After}$
F3	$\mu + \alpha_3$	$\mu + \alpha_3 + \beta_{After} + \gamma_{3After}$
F4	$\mu + \alpha_4$	$\mu + \alpha_4 + \beta_{After} + \gamma_{4After}$

In order to evaluate the model quality, the individual boxplots of the standardized residuals by formant and stage are plotted in Figure 3. The residual analysis of model M_0 reveals that residuals are centered by stage, but not by formant. Note that residuals have different variances from a formant to another. To correct the first defect, we build a linear mixed-effects model by introducing individual random effect ξ_{if} in formant estimate, leading to model M_1:

$$\begin{bmatrix} F_{1sik} \\ F_{2sik} \\ F_{3sik} \\ F_{4sik} \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \beta_s \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \gamma_{1s} \\ \gamma_{2s} \\ \gamma_{3s} \\ \gamma_{4s} \end{bmatrix} + \xi_i \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \\ \xi_{i3} \\ \xi_{i4} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1sik} \\ \varepsilon_{2sik} \\ \varepsilon_{3sik} \\ \varepsilon_{4sik} \end{bmatrix} \quad (2)$$

with $\xi_{if} \sim \mathcal{N}(0, \tau_1^2)$.

We fit Model M_1 using the R code displayed in Annex A/Code 2. For each formant, the boxplots of the standardized residuals by individual for model M1, displayed in Figure 4, are now centered at zero. However, Figure 4 also indicates that the residual variability is different from a formant to another. To take this variability into account, a new model M_2 is defined assuming a different variance per formant for ξ_{if} i.e $\xi_{if} \sim \mathcal{N}(0, \tau_f^2)$. The R code used to fit this model is displayed in Annex A/Code 3 and the individual boxplots of the standardized residuals by formant and stage are presented in Figure 5.

The residuals for model M_2 are similar to those obtained for model M_1. Nevertheless, to compare both models, we use the ANOVA function which displays the AIC and BIC values and the p-value of the likelihood ratio test. The results displayed in Annex A/Code 4 suggest that model M_2 fits the data better. However, note that this model does not improve the residuals graphs: there still remains different residual variability from one formant to another.

To deal with this problem, we consider in the following section a more general model keeping the random-effects structure defined in model M2, but allowing different variances by formant for the within-group errors. Moreover, since the four formants are simultaneous measures, the corresponding random variables cannot be considered as independent. The correlation matrix of the errors need to be taken into account in the model.

3.1.2 Modelling the variance-covariance matrix of the errors

Since the boxplots of the standardized residuals by formant still present a different variability from a formant to another, a different variance per formant in the variance-covariance matrix of the errors $[\varepsilon_{1sik}, \varepsilon_{2sik}, \varepsilon_{3sik}, \varepsilon_{4sik}]$ is introduced, leading to a diagonal matrix with different diagonal terms. The residual error ε_{fsik} is supposed to be normally distributed, centered, with variance σ_f^2. This new model named M_3 is fitted using the code displayed in

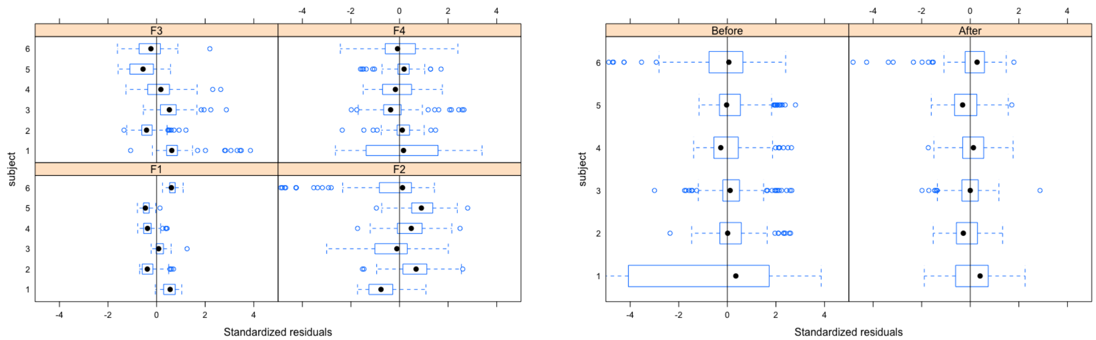

Figure 3: Standardized residuals by formant (left) and stage (right) for each subject for model M_0.

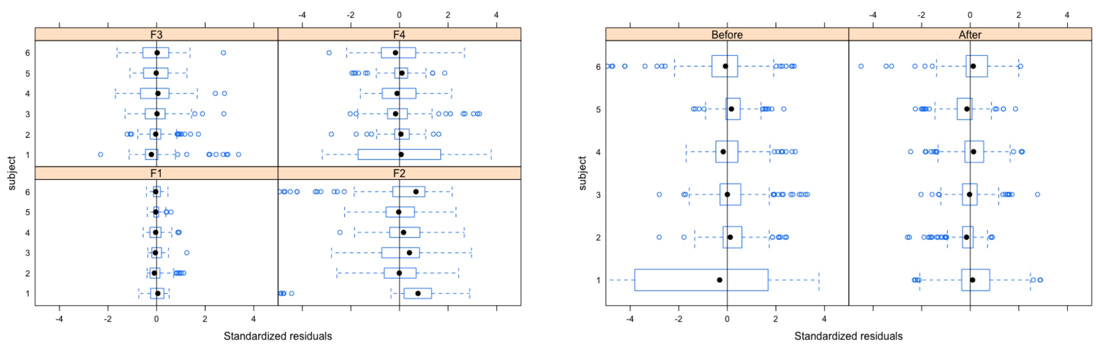

Figure 4: Standardized residuals by formant (left) and stage (right) for each subject for model M_1.

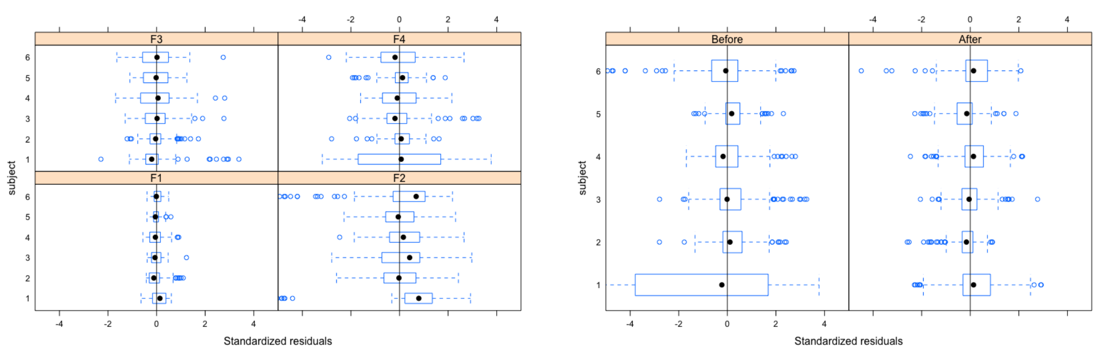

Figure 5: Standardized residuals by formant (left) and stage (right) for each subject for model M_2.

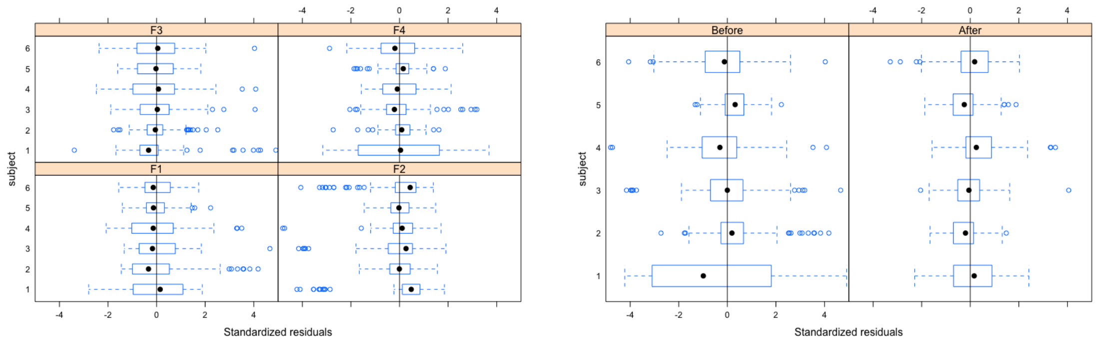

Figure 6: Standardized residuals by formant (left) and stage (right) for each subject for model M_3.

Annex A/Code 5.

The boxplots of the standardized residuals by formant and stage for model M_3 are presented in Figure 6. They show that the formant variability of the data has been captured since the standardized residuals are now similarly scattered from one formant to another. The results of the ANOVA function displayed in Annex A/Code 6 confirm that model M_3 fits the data better than model M_2.

To upgrade model M_3 by taking into account the dependence between formants, we now introduce a correlation matrix structure. In a new model M_4, the variance-covariance matrix of the errors $[\varepsilon_{1sik}, \varepsilon_{2sik}, \varepsilon_{3sik}, \varepsilon_{4sik}]$ is non-diagonal (see details in (Bazzoli et al., 2015)). Model M_4 is fitted using the code displayed in Annex A/Code 7. The ANOVA function is used to compare models M_3 and M_4. Although the residuals graphics look very similar in both models, the p-value of the likelihood ratio test statistic allows us to conclude that model M_4 fits the data better than model M_3. Thus, this variance-covariance structure of the errors is kept for the next modelling step.

3.1.3 Modelling the fixed-effects structure

Once the random-effects structure and the variance-covariance matrix of the errors are selected, we focus on modelling the fixed effects. For this purpose, we examine the estimations of fixed coefficients μ, α_f, β_s and γ_{fs}:

```
Fixed effects: Y ~ formant * stage
                       Value Std.Error
(Intercept)          364.746  18.15550
formantF2           1633.088  82.79015
formantF3           2341.679  41.75014
formantF4           3478.763  31.97609
stageAfter             7.749   3.46054
formantF2:stageAfter 140.551  20.46553
formantF3:stageAfter -15.282   9.00898
formantF4:stageAfter -53.332  14.13607

                       DF   t-value
(Intercept)          2663  20.09010
formantF2            2663  19.72563
formantF3            2663  56.08793
formantF4            2663 108.79263
stageAfter           2663   2.23916
formantF2:stageAfter 2663   6.86771
formantF3:stageAfter 2663  -1.69631
formantF4:stageAfter 2663  -3.77276

                      p-value
(Intercept)           0.0000
formantF2            0.0000
formantF3            0.0000
formantF4            0.0000
stageAfter             0.0252
formantF2:stageAfter  0.0000
formantF3:stageAfter  0.0899
formantF4:stageAfter  0.0002
```

All coefficients are significantly different from zero at level 5% except γ_{3After}. This leads us to

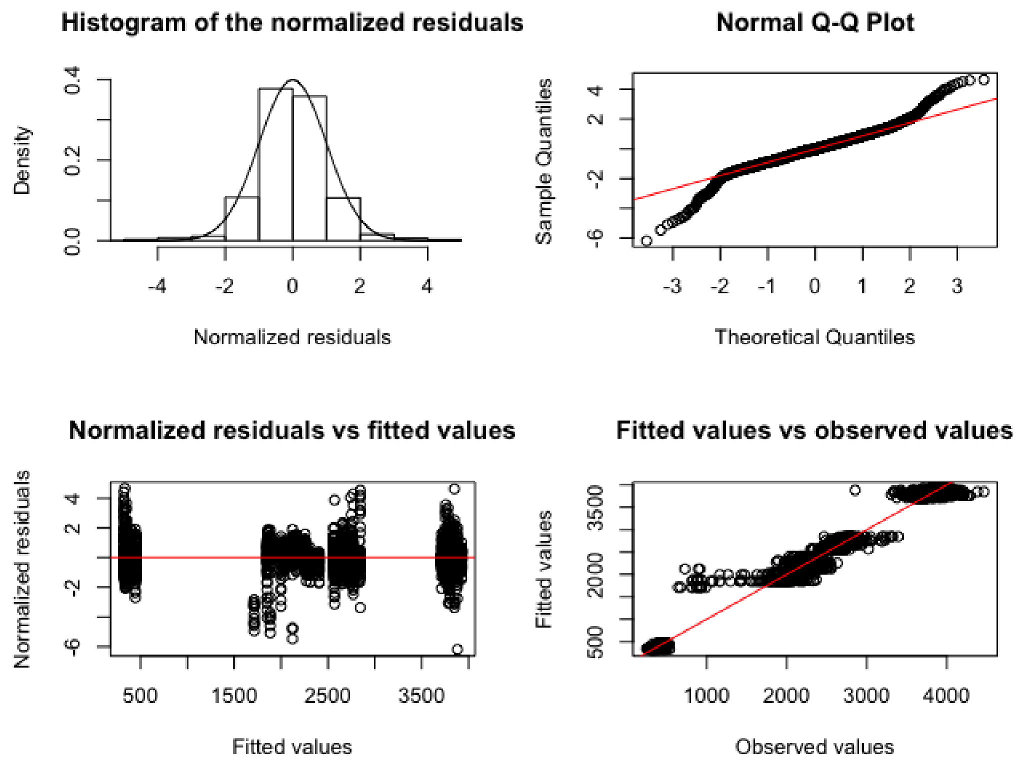

Figure 7: Diagnostic plots for model M_4.

test the simultaneous nullity of the interaction coefficients γ_{fs} by fitting model M_5 without any interaction. This is done in Annex A/Code 8 as well as the comparison of models M_4 and M_5 using the ANOVA function. The p-value of the likelihood ratio statistic leads us to conclude that the interaction is significant and thus model M_4 is preferred to M_5.

3.2. *Model validation*

To validate model M_4 selected by the data modelling step, model M_4 is fitted again by restricted maximum likelihood estimation (REML) which is often preferred to maximum likelihood estimation (ML) because it produces unbiased and non-negative variance parameter estimates (Patterson and Thompson, 1971). Classical diagnostic plots are used and displayed in Figure 7: normalized residuals histogram, normal QQ-plot, normalized residuals versus fitted values plot, fitted values versus observed values plot. The normalized residuals histogram and the normal QQ-plot suggest that there is a good fit between the normal distribution and the residuals distribution, except for the extreme tails. The normalized residuals versus fitted values plot does not highlight any residual structure.

3.3. *Statistical tests in the selected model*

To answer the phonetic questions stated in Section 2, contrast tests involving the fixed effects parameters μ, α_f, β_s and γ_{fs} are performed. First, answering the **phonetic question Q1** with regard to formant values before training amounts to comparing the mean measure for each formant at stage $Before$ to a reference value in the French language. The mean values measured by Georgeton et al. (2012) on 40 native French female speakers have been proposed as a reference in contrastive studies of French as a Foreign Language (FLE) production. They are selected as reference values here. These mean measures are displayed in column *Before* of Table 1.

From a statistical point of view, this boils down to performing the following simultaneous four tests of the null hypotheses:

$$H_{0,f}^{Before} : \mu + \alpha_f = \phi_f, \quad f \in \{1, \dots, 4\}$$

where $\phi_1 = 276$, $\phi_2 = 2091$, $\phi_3 = 2579$ and $\phi_4 = 3826$.

Note that, since four tests are simultaneously performed, the p-values need to be adjusted with respect to the case where the tests are performed separately (Dudoit and Van der Laan, 2008; Riou, 2013).

For that purpose, a contrast matrix which gives the linear combinations from the fixed effects parameters is built:

$$\begin{bmatrix} \mu \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \mu + \alpha_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_{After} \\ \gamma_{2After} \\ \gamma_{3After} \\ \gamma_{4After} \end{bmatrix},$$

and the statistical tests are performed using the R code displayed in Annex A/Code 9.

To answer the **phonetic question Q2** about achieving the target reference values after training, we perform the following simultaneous four tests of the null hypotheses:

$$H_{0,f}^{After} : \mu + \alpha_f + \beta_{After} + \gamma_{fAfter} = \phi_f, \quad f \in \{1, \dots, 4\}$$

The same methodology is followed building the new appropriate contrast matrix in order to obtain the mean measures after training (column *After* in Table 1). This matrix is built by the R code displayed in Annex A/Code 10.

To answer the **phonetic question Q3** about formant similarity before and after training, the mean measure for each formant at stage *After* (column "After" in Table 1) is compared to the mean measure at stage *Before* (column "Before" in Table 1). This amounts to comparing their differences to zero. That leads to the four simultaneous tests of the null hypotheses:

$$H_{0,f} : \beta_{After} + \gamma_{fAfter} = 0, \quad f \in \{1, \dots, 4\}$$

The new appropriate contrast matrix is built in order to obtain these differences. The R code is displayed in Annex A/Code 11.

The phonetic questions Q4, Q5 and Q6 deal with the distances between successive formants. Distances are calculated as the differences between formants frequencies. Table 2 displays the mean measures for each difference at each stage:

Table 2: Mean measures for the differences between formants at each stage.

Differences	Stage	
of formants	Before	After
F2-F1	α_2	$\alpha_2 + \gamma_{2After}$
F3-F2	$\alpha_3 - \alpha_2$	$\alpha_3 - \alpha_2 + \gamma_{3After} - \gamma_{2After}$
F4-F3	$\alpha_4 - \alpha_3$	$\alpha_4 - \alpha_3 + \gamma_{4After} - \gamma_{3After}$

To answer **Question Q4**, the mean measures of the differences before training (column *Before* in Table 2) need to be compared to same differences for target values. This comes to perform simultaneously the three following tests of the null hypotheses:

$$
\begin{aligned}
H^{Before}_{0,F2-F1} &: \alpha_2 & &= 2091 - 276 = 1815 \\
H^{Before}_{0,F3-F2} &: \alpha_3 & -\alpha_2 &= 2579 - 2091 \\
& & &= 488 \\
H^{Before}_{0,F4-F3} &: \alpha_4 & -\alpha_3 &= 3826 - 2579 \\
& & &= 1247
\end{aligned}
$$

As previously done, a new appropriate contrast matrix is built in order to obtain the mean measures of the differences before training (column *Before* in Table 2). The R code is displayed in Annex A/Code 12.

To answer **Question Q5**, the following three tests of the null hypotheses are simultaneaously performed:

$$
\begin{aligned}
H^{After}_{0,F2-F1} :& \alpha_2 + \gamma_{2After} \\
&= 2091 - 276 = 1815 \\
H^{After}_{0,F3-F2} :& \alpha_3 - \alpha_2 + \gamma_{3After} - \gamma_{2After} \\
&= 2579 - 2091 = 488 \\
H^{After}_{0,F4-F3} :& \alpha_4 - \alpha_3 + \gamma_{4After} - \gamma_{3After} \\
&= 3826 - 2579 = 1247
\end{aligned}
$$

The corresponding R code is displayed in Annex A/Code 13.

Finally, **Question Q6** consists in comparing the two columns of Table 2. This amounts to performing the following simultaneous three tests of the null hypotheses:

$$
\begin{aligned}
H_{0,F2-F1} &: \gamma_{2After} = 0 \\
H_{0,F3-F2} &: \gamma_{3After} - \gamma_{2After} = 0 \\
H_{0,F4-F3} &: \gamma_{4After} - \gamma_{3After} = 0
\end{aligned}
$$

The corresponding R code is displayed in Annex A/Code 14.

4. Results

The R outputs of the statistical tests described in the previous section are displayed in Annex B. In this section, we interpret the results. For each test, the null hypothesis is rejected at level 5% when the p-value is smaller than 0.05.

4.1. *Results for /y/ vowel*

Concerning Question 1, the results show that the formant measures differ significantly from target reference values for formants F1 and F3. For these two formants, the p-values are respectively equal to $3.11e-05$ and 0.0078, thus smaller than 0.05. In these two cases, the null hypothesis is rejected, and we can conclude that these two formants do not achieve the target reference value before training for the /y/ vowel. More precisely, F1 and F3 mean formant measures are higher than the target reference values. For formants F2 and F4, the null hypothesis cannot be rejected.

Concerning Question Q2, the null hypotheses are rejected at level 5% for formants F1 and F3 after training ($p = 4.5e-06$ and $p = 0.014$). Formants F1 and F3 are still higher than the target reference values.

Concerning Question Q3, the four tests of the null hypothesis lead to the conclusion that formants $F2$ and $F4$ have evolved during the training ($p < e-04$ and $p = 0.00204$), whereas formants $F1$ and $F3$ have not ($p = 0.097$ and $p = 0.855$). More precisely, formant F2 has increased during the training but still remains close to the target reference value, whereas formant F4 has decreased during the training even if it also remains close to the target reference value.

Results for Question Q4 show that before training, all formant distances F2-F1, F3-F2 and F4-F3 achieve the corresponding target reference values, at level 5%. Note that the p-values are very close to level 5% ($p = 0.1018$, 0.0563 and 0.0657). A focalization similar to the expected French one seems already in place before training.

Concerning Question Q5, it can be concluded that, after training, distances F2-F1 and F3-F2 still achieve their target value ($p = 0.907$ and $p = 0.795$), whereas distance F4-F3 does not achieve its target value anymore ($p = 0.007$). This result is mainly due to the decreasing of formant F4 after training.

Finally, the three tests of the null hypotheses to answer Question Q6 lead to conclude that all the distances between the formants have evolved between before and after training. The null hypothesis for each statistical test is rejected as the three p-values are smaller than 0.05. Clearly, the evolutions of distances F2-F1 and F3-F2 are both due to the evolution of formant F2.

4.2. *Results for /ø/ vowel*

For /ø/ vowel, the same methodology as for /y/ vowel has been applied. The model selection step leads to the same model M_4 as for /y/ vowel. Concerning the statistical tests step, only the target values change, they are now $\phi_1 = 406$, $\phi_2 = 1599$, $\phi_3 = 2703$ and $\phi_4 = 3985$ as referenced by Georgeton et al. (2012).

The results obtained for Question Q1 show that formants F2 and F4 have already achieved the target value before training for /ø/ vowel ($p = 0.689$ and $p = 0.567$) whereas formants F1 and F3 do not ($p < 1e-04$ and $p < 0.002$). More precisely, F1 and F3 mean formant measures are higher that their target reference values. This result is similar to what was obtained for vowel /y/.

Statistical tests for Question Q2 enable us to conclude that only formant F2 achieves the target value after training ($p = 0.202$).

Concerning Question Q3, the four tests of the null hypotheses lead to the conclusion that all formant values have increased during the training ($p = 0.0002$, $p = 3.44e-05$ and $p < 1e-05$), except formant F1 ($p = 0.999$).

The results obtained for Question Q4 show that only the difference F4-F3 has not achieved its target reference value before training ($p = 0.0003$). This is due to the fact that formant F3 is quite higher than its target reference value.

From the results obtained for Question Q5, it can be concluded that all differences between formants achieve the target reference values after training, at level 5%.

Finally, the results obtained for Question Q6 lead us to conclude that only the differences F2-F1 and F4-F3 have evolved with the training for this vowel ($p = 0.0003$ and $p < 1e-04$).

5. Discussion and conclusion

In this paper, we discuss a methodological statistical approach in the context of formant measurements expected to reflect the benefit of an 8-hour French pronunciation training proposed to native Italian female speakers. Statistical models of increased complexity have been proposed to fit the phonetical data and to take into account the dependence between the measured formants. We draw attention to the fact that all statistical tests are based on model assumptions. Test conclusions are only valid if the model assumptions are also valid, and no p-value should be interpreted before the model has been carefully validated. Graphical model validation tools such as residuals plots have been presented in detail in the paper.

The results for the two anterior non-native vowels /y/ and /ø/ have been presented. The main observation is that the native Italian speakers did not produce these two French vowels in the expected formantic areas (differences assessed in F1 and F3) prior to training. Training did not improve the matching to the target reference values, even if their vowel pronunciation evolved during training. Concerning focalization which is addressed by rather small differences between adjacent formants (F2-F1, F3-F2, and F4-F3), a similarity with French focalization was already in place prior to the training for vowel /y/, and in particular the F3-F2 distance which is perceptually relevant for this vowel. However after training, the F3-F2 distance is smaller due to higher F2 and rather stable F3 values (F2 moves away from F1 and approaches F3) reinforcing the spectral focalization. For vowel /ø/, differences were found for F4-F3 before training. Training removed these differences for vowel /ø/. Furthermore, training also modified the F4-F3 distance for vowel /y/.

The statistical modelling approach developed here can be used in all phonetical studies which make use of comparison of formant measurements.

Acknowledgements

All of the authors have been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) for this work. This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the investissements d'avenir program (ANR-15-IDEX-02).

References

Baayen, R. H., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Bazzoli, C., Letué, F., and Martinez, M.-J. (2015). Modelling finger force produced from different tasks using linear mixed models with lme R function. *Journal of Case Studies in Business, Industry and Government Statistics (CSBIGS)*, 6(1):16–36.

Bergmann, C., Nota, A., Sprenge, S. A., and Schmid, M. S. (2016). L2 immersion causes non-native-like l1 pronunciation in german attriters. *Journal of Phonetics*, 58:71–86.

Cornaz, S. (2014). *Using singing-voice tasks for outcomes in phonetic and phonological correction of a foreign language*. Theses, Université de Grenoble.

Dudoit, S. and Van der Laan, J. (2008). *Multiple Testing Procedures with Applications to Genomics.* Springer, NewYork.

Georgeton, L., Paillereau, N., Landron, S., Gao, J., and Kamiyama, T. (2012). Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de fle. In Besacier, L., Lecouteux, B., and Gérasset, G., editors, *Proceedings of the XXIXèmes Journées d'Etudes sur la Parole (Joint Conference JEP-TALN-RECITAL)*, page 145–152, Grenoble. ATALA/AFCP.

Hazan, V. and Barett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, 28(4):377–396.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):545–554.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686.

Patterson, H. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58(3):346–363.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117.

Politzer-Ahles, S. and Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics*, 70:56–69.

Riou, J. (2013). *Multiplicité des tests, et calculs de taille d'échantillon en recherche clinique.* Theses, Université de Bordeaux Segalen.

Roettger, T. B., Winter, B., and Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, 73:1–7.

Schwartz, J. L., Abry, C., Boë, L. J., Ménard, L., and Vallée, N. (2005). Asymmetries in vowel perception, in the context of the dispersion–focalisation theory. *Speech Communication*, 45(4):425–434.

Schwartz, J. L., Boë, L. J., Vallée, N., and Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25(3):255–286.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english. *Journal of Phonetics*, 70:86–116.

Correspondence: frederique.letue@univ-grenoble-alpes.fr.

Appendix A: Codes

Code 1: R code for fitting model M_0 and plotting the residuals

```
resM0.Vy.std<-residuals(fitM0.Vy,type="pearson")
plot(fitM0.Vy,subject~resM0.Vy.std|formant,abline=0,xlim=c(-5,5),
       xlab="Standardized residuals")
plot(fitM0.Vy,subject~resM0.Vy.std|stage,abline=0,xlim=c(-5,5),
       xlab="Standardized residuals")
```

Code 2: R code for fitting model M_1

```
fitM1.Vy <- lme(Y ~ formant*stage,
                  random=list(subject=pdBlocked(list(pdIdent(~1),
                                                  pdIdent(~formant-1)))),
                   method="ML")
```

Code 3: R code for fitting model M_2

```
fitM2.Vy <- lme(Y~ formant*stage,
                random=list(subject=pdBlocked(list(pdIdent(~1),
                                                   pdDiag(~formant-1)))),
                method="ML")
```

Code 4: R code for comparing models M_1 and M_2

```
> anova(fitM1.Vy,fitM2.Vy)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fitM1.Vy     1 11 34980.65 35045.46 -17479.33
fitM2.Vy     2 14 34971.34 35053.83 -17471.67 1 vs 2 15.30663  0.0016
```

Code 5: R code for fitting model M_3

```
fitM3.Vy <- lme(Y ~ formant*stage,
                random=list(subject=pdBlocked(list(pdIdent(~1),
                                                  pdDiag(~formant-1)))),
                weights=varIdent(form=~1|formant),
                method="ML",control=lmeControl(niterEM=200))
```

Code 6: R code for comparing models M_2 and M_3

```
> anova(fitM2.Vy,fitM3.Vy)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fitM2.Vy     1 14 34971.34 35053.83 -17471.67
fitM3.Vy     2 17 33340.55 33440.71 -16653.28 1 vs 2 1636.794  <.0001
```

Code 7: R code for fitting model M_4 and comparing models M_3 and M_4

```
fitM4.Vy <- lme(Y ~ formant*stage,
                random=list(subject=pdBlocked(list(pdIdent(~1),
                                                   pdDiag(~formant-1)))),
                weights=varIdent(form=~1|formant),
                correlation=corSymm(form=~1|subject/trial),
                method="ML",control=lmeControl(msMaxIter=1000))
> anova(fitM3.Vy,fitM4.Vy)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fitM3.Vy     1 17 33340.55 33440.71 -16653.28
fitM4.Vy     2 23 33253.57 33389.09 -16603.79 1 vs 2 98.97661  <.0001
```

Code 8: R code for fitting model M_5 and comparing models M_4 and M_5

```
fitM5.Vy <- lme(Y ~ formant+stage,
                random=list(subject=pdBlocked(list(pdIdent(~1),
                                                   pdDiag(~formant-1)))),
                weights=varIdent(form=~1|formant),
                correlation=corSymm(form=~1|subject/trial),
                method="ML",control=lmeControl(msMaxIter=1000))
> anova(fitM5.Vy,fitM4.Vy)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fitM5.Vy     1 20 33301.47 33419.31 -16630.73
fitM4.Vy     2 23 33253.57 33389.09 -16603.79 1 vs 2 53.89389  <.0001
```

Code 9: R code for fitting model M_4 with the REML method and for performing the statistical tests for Question 1

```
fitM4.Vy.REML <- lme(Y ~ formant*stage,
                random=list(subject=pdBlocked(list(pdIdent(~1),
                                                   pdDiag(~formant-1)))),
                weights=varIdent(form=~1|formant),
                correlation=corSymm(form=~1|subject/trial),
                method="REML",control=lmeControl(msMaxIter=1000))

library(multcomp)
valeurs.cible.y<-c(276,2091,2579,3826)
question1<-rbind('mu'=c(1,0,0,0,0,0,0,0),
              'mu+alpha2'=c(1,1,0,0,0,0,0,0),
              'mu+alpha3'=c(1,0,1,0,0,0,0,0),
              'mu+alpha4'=c(1,0,0,1,0,0,0,0))
summary(glht(fitM4.Vy.REML,linfct=question1,rhs=valeurs.cible.y))
```

Code 10: R code for performing the statistical tests for Question 2

```
question2<-rbind('mu+beta.After'=c(1,0,0,0,1,0,0,0),
                 'mu+alpha2+beta.After+gamma2After'=c(1,1,0,0,1,1,0,0),
                 'mu+alpha3+beta.After+gamma3After'=c(1,0,1,0,1,0,1,0),
                 'mu+alpha4+beta.After+gamma4After'=c(1,0,0,1,1,0,0,1))
summary(glht(fitM4.Vy.REML,linfct=question2,rhs=valeurs.cible.y))
```

Code 11: R code for performing the statistical tests for Question 3

```
question3<-rbind('betaAfter'=c(0,0,0,0,1,0,0,0),
               'betaAfter+gamma2After'=c(0,0,0,0,1,1,0,0),
               'betaAfter+gamma3After'=c(0,0,0,0,1,0,1,0),
               'betaAfter+gamma4After'=c(0,0,0,0,1,0,0,1))
summary(glht(fitM4.Vy.REML,linfct=question3))
```

Code 12: R code for performing the statistical tests for Question 4

```
question4<-rbind('alpha2'=c(0,1,0,0,0,0,0,0),
                  'alpha3-alpha2'=c(0,-1,1,0,0,0,0,0),
                  'alpha4-alpha3'=c(0,0,-1,1,0,0,0,0))
summary(glht(fitM4.Vy.REML,linfct=question4,rhs=diff(valeurs.cible.y)))
```

Code 13: R code for performing the statistical tests for Question 5

```
question5<-rbind('alpha2+gamma2After'=c(0,1,0,0,0,1,0,0),
                  'alpha3-alpha2+gamma3After-gamma2After'=c(0,-1,1,0,0,-1,1,0),
                  'alpha4-alpha3+gamma4After-gamma3After'=c(0,0,-1,1,0,0,-1,1))
summary(glht(fitM4.Vy.REML,linfct=question5,rhs=diff(valeurs.cible.y)))
```

Code 14: R code for performing the statistical tests for Question 6

```
question6<-rbind('gamma2After'=c(0,0,0,0,0,1,0,0),
                  'gamma3After-gamma2After'=c(0,0,0,0,0,-1,1,0),
                  'gamma4After-gamma3After'=c(0,0,0,0,0,0,-1,1))
summary(glht(fitM4.Vy.REML,linfct=question6))
```

Appendix B: R outputs for /y/ and /ø/ vowels

5.1. R output for Question Q1

Question Q1 : Do formants already achieve the French reference value before training?

1. **Vowel /y/**

```
Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
mu == 276            364.78      19.85   4.471 3.11e-05 ***
mu+alpha2 == 2091   1997.63      88.21  -1.058   0.7457
mu+alpha3 == 2579   2706.52      41.18   3.097   0.0078 **
mu+alpha4 == 3826   3843.52      28.40   0.617   0.9541
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                   Estimate Std. Error z value Pr(>|z|)
mu == 406            481.97      10.00   7.595  < 1e-04 ***
mu+alpha2 == 1599   1676.45      69.67   1.112  0.68895
mu+alpha3 == 2703   2794.41      26.11   3.502  0.00182 **
mu+alpha4 == 3985   3955.19      23.23  -1.283  0.56691
```

5.2. R output for Question Q2

Question Q2 : Do formants achieve the French reference reference value after training?

1. **Vowel /y/**

```
Linear Hypotheses:
                                       Estimate Std. Error z value Pr(>|z|)
mu+betaAfter == 276                      372.49      19.82   4.868  4.5e-06 ***
mu+alpha2+betaAfter+gamma2After == 2091  2146.13      87.95   0.627    0.952
mu+alpha3+betaAfter+gamma3After == 2579  2698.89      41.07   2.919    0.014 *
mu+alpha4+betaAfter+gamma4After == 3826  3797.92      28.05  -1.001    0.782
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                                       Estimate Std. Error z value Pr(>|z|)
mu+betaAfter == 406                     482.703      9.928   7.726  < 0.001 ***
mu+alpha2+betaAfter+gamma2After == 1599 1730.940     69.538   1.897  0.20205
mu+alpha3+betaAfter+gamma3After == 2703 2836.611     25.928   5.153  < 0.001 ***
mu+alpha4+betaAfter+gamma4After == 3985 4060.191     22.905   3.283  0.00419 **
```

5.3. R output for Question Q3

Question Q3 : Are formants similar before and after training?

1. **Vowel /y/**

```
Linear Hypotheses:
                          Estimate Std. Error z value Pr(>|z|)
betaAfter == 0               7.717      3.458   2.231  0.09731 .
betaAfter+gamma2After == 0  148.503     20.160   7.366  < 1e-04 ***
betaAfter+gamma3After == 0   -7.628      8.857  -0.861  0.85485
betaAfter+gamma4After == 0  -45.595     13.127  -3.473  0.00204 **
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                          Estimate Std. Error z value Pr(>|z|)
betaAfter == 0              0.7317     3.7915   0.193 0.999399
betaAfter+gamma2After == 0  54.4931    13.4606   4.048 0.000206 ***
betaAfter+gamma3After == 0  42.1969     9.4855   4.449 3.44e-05 ***
betaAfter+gamma4After == 0 105.0018    12.0893   8.686  < 1e-05 ***
```

5.4. R output for Question Q4

Question Q4 : Is focalization before training already similar to that of French front vowels?

1. **Vowel /y/**

```
Linear Hypotheses:
                      Estimate Std. Error z value Pr(>|z|)
alpha2 == 1815         1632.85      90.42  -2.014   0.1018
alpha3-alpha2 == 488    708.89      97.56   2.264   0.0563 .
alpha4-alpha3 == 1247  1137.00      49.97  -2.201   0.0657 .
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                      Estimate Std. Error z value Pr(>|z|)
alpha2 == 1193         1194.48      69.05   0.021 0.999978
alpha3-alpha2 == 1104  1117.97      72.97   0.191 0.987116
alpha4-alpha3 == 1282  1160.77      31.65  -3.830 0.000257 ***
```

5.5. *R output for Question Q5*

Question Q5 : Is focalization after training similar to that of French front vowels?

1. **Vowel /y/**

```
Linear Hypotheses:
                                                Estimate Std. Error z value Pr(>|z|)
alpha2+gamma2After == 1815                       1773.64      90.15  -0.459  0.90676
alpha3-alpha2+gamma3After-gamma2After == 488      552.76      97.23   0.666  0.79504
alpha4-alpha3+gamma4After-gamma3After == 1247   1099.03      49.70  -2.977  0.00722 **
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                                                Estimate Std. Error z value Pr(>|z|)
alpha2+gamma2After == 1193                       1248.24      68.91   0.802    0.697
alpha3-alpha2+gamma3After-gamma2After == 1104   1105.67      72.81   0.023    1.000
alpha4-alpha3+gamma4After-gamma3After == 1282   1223.58      31.37  -1.862    0.136
```

5.6. *R output for Question Q6*

Question Q6 : Are distances between successive formants similar before and after training?

1. **Vowel /y/**

```
Linear Hypotheses:
                             Estimate Std. Error z value Pr(>|z|)
gamma2After == 0               140.79      20.45   6.884   <0.001 ***
gamma3After-gamma2After == 0  -156.13      23.59  -6.618   <0.001 ***
gamma4After-gamma3After == 0   -37.97      15.59  -2.436   0.0352 *
```

2. **Vowel /ø/**

```
Linear Hypotheses:
                             Estimate Std. Error z value Pr(>|z|)
gamma2After == 0                53.76      13.87   3.876 0.000258 ***
gamma3After-gamma2After == 0   -12.30      14.85  -0.828 0.724612
gamma4After-gamma3After == 0    62.80      13.13   4.783  < 1e-04 ***
```