# KNOWLEDGE-DRIVEN VS. DATA-DRIVEN METHODS FOR FILTERING ACOUSTIC MEASURES IN PHONETICS CORPORA

Lancien Mélanie, Adda-Decker Martine, Stuart-Smith Jane

FNSRS, LPP, LISN, GULP
melanie.lancien@unil.ch, madda@limsi.fr, Jane.Stuart-Smith@glasgow.ac.uk

## ABSTRACT

This paper addresses the issue of vowel formant filtering for large scale phonetic analyses and evaluates an innovative data-driven method to remove spurious items. Automatic formant detection is error-prone due to formant jumps among other issues. A common solution is to adopt formant filters (FF) discarding tokens with measurements falling too far away from knowledge-based references. The proposed approach uses the Mahalanobis distance (MD) as a purely data-driven method. First, all vowel formant and duration values are used to compute vowel profiles. These learnt profiles supersede the need for reference values to carry out the filtering. We compare the two (knowledge-based vs data-driven) filtering approaches on the same dataset of French spontaneous speech. Results demonstrate the efficiency of MD filtering. The amount of filtered data is easily adjustable. Moreover, the data-driven status of the approach makes it well suited for less described languages.

**Keywords:** Vowels, Formants, Data Science, Automatic filtering, Corpus Phonetics

## 1. INTRODUCTION

Data derived from automatically annotated medium-size corpora are generally numerous and noisy. Different kinds of error arise from automated annotation and metrics extraction, such as formant-tracking jumps, transcription issues, or mismatches between the expected and the actual pronunciation of a word. These then lead to the presence of erroneous datapoints in data for analysis. The results from very large datasets may not be overly affected, but for medium-size datasets (e.g. with 4 or 5 tokens of each vowel type per speaker), more common in phonetic studies, erroneous values are more problematic, likely biasing subsequent results. Thus medium-size datasets need "cleaning" (filtering) as far as possible.

Unfortunately, the filtering method, and choices and outcomes linked to the process, are rarely reported. Among the few who do, [1] in a large-scale study on English /s/ showed a rejection rate of 10% with range filter methods, and [2] found 17% of her tokens were classified as non canonical with an ABX alignment tool [3] (see subsection 1.2). Thus data filter may result in rejecting up to 20%, whatever the corpus and method.

### 1.1. Post extraction methods

Several methods can be used to identify erroneous values during or after the value extraction process, some are knowledge-driven, others more data-driven. In phonetic studies, formant values are often extracted using Praat Burg algorithm [4]) and then checked via values range filters (e.g. [5]). This procedure is a knowledge-based "posthoc" procedure that consists of going through all acoustic values and identifying all likely erroneous values of e.g. F1 and F2 (e.g. F1 at 1500Hz for French /i/). However, "range filter" methods only account for expected values for the phoneme, and sometimes speaker gender; [5].

More recently, [6] tried another way of processing the extracted values, using Mahalanobis distance. This allowed the identification of erroneous datapoints based on a multi-parametric distribution and distance [7]. Mahalanobis distance is a multidimensional measure of the number of standard deviations between a point P and the mean of a distribution D; in our case P is a token of a vowel and D the distribution of the variable X for the group to which P belongs (e.g. P is an [i] and D is the distribution of the F2 of the /i/ type). The Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_p)^T$ to a set of mean value vectors $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)^T$ and having a covariance matrix $\Sigma$ is defined as follows:

$$(1) \quad Dx = (x-\mu)^T \Sigma^{-1} (x-\mu)$$

The square root of $Dx = (x-\mu)^T\Sigma^{-1}(x-\mu)$ gives the number of standard deviations between the observation and the mean of the distribution. If $P = \mu D$, the distance is 0, and increases as P moves away from the mean in a determined space. Thus by choosing a threshold for the distance ([6] chose 3 standard deviations), it is easy to discard vowel tokens with features further than $X$ standard deviations from the computed average profile.

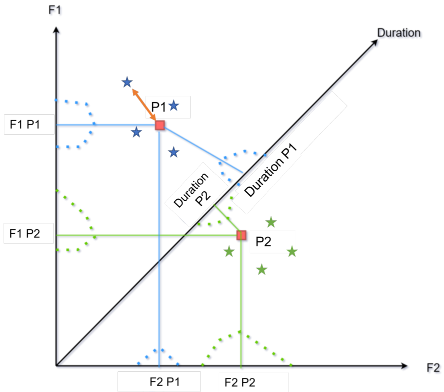

**Figure 1:** Schematic explanation of the Mahalanobis distance. Blue lines represent the distance of F1 (axis y) F2 (axis x) and duration (axis z) between the mean (orange squares) of two vowels (P1 and P2). Stars are tokens of P1 and P2 and the orange arrow indicate the distance measured.

### 1.2. Pre extraction methods

Other methods, cast in the NLP scope, can prevent the generation of some erroneous values directly during the annotation or extraction process. For instance, [2] uses sound prototypes and forced alignment to reclassify datapoints that do not match the sound *x*'s profile directly during the labelling phase, by labelling them phone *y* instead of *x*. With this method, for example, occurrences of French /b/ (whose prototype will show negative VOT) which do not have a negative VOT will be annotated by the ABX system as its unvoiced counterpart /p/. Such methods can help identify the phonetic realization of sounds actually produced, but are restricted to the profiles of variation implemented in the system. Nor can they discard tokens that are too far away from the target or are noise resulting from alignment issues, without being given a typological description.

Finally, recent propositions like [8] with its new iteration of the FAVE automated formant measurement ([9, 10]), also uses Mahalanobis distance as part of the procedure. A matrix of several re-measurements of formants on the same phone is used to eradicate errors directly during the formant detection and measurement process instead of post-processing the extracted values with filters. Each vowel is measured several times with different numbers of LPC coefficients, and the resulting measures are compared to a prototype (mean formant frequencies/bandwidths) and a covariance matrix. The selected set of measures is that with the smallest Mahalanobis distance from the prototype. This leads to a smaller amount of erroneous values in the output, given that they are corrected directly during the extraction phase. However this method cannot be applied as a posthoc process, and must be run on the raw sound files, which is quite a limitation for studies on open-source datasets, relying on previously measured files (see SPADE [11] or ESTER [12]).

### 1.3. Conclusions on the state of the art

Even though they are largely preferred (maybe because easier to use), range filters are not optimal since the boundary values must be set by the user implying two main limitations : 1) there must be literature on those values for the given language; 2) ranges are generally set in a wide and arbitrary fashion (e.g. [5] use ranges from 1500 to 2500Hz for French /i/ F2). However it does efficiently identify the most obvious erroneous values.

In this paper we compare the two posthoc (post extraction) filtering methods mentioned above to provide more background on how they operate and which results can be expected. Thus we first replicate [5]'s work with the "knowledge-based" formant range filter (FF hereafter), and then uses a filter based on Mahalanobis' distance (MD), as in [6], on the same dataset.

## 2. KNOWLEDGE-BASED FORMANT FILTERS

In [5] the authors work on manually transcribed broadcast news corpora in French and German. They extract vowel formants and want to provide a clean dataset without manual checking. To do so, they use formant range filters to get rid of errors.

Their corpus consisted in 4h of journalistic speech: 2h in French (1h ♂, 1h ♀) and 2h in German (1h ♂, 1h ♀). The corpus was automatically segmented and labelled into phones using the LIMSI speech alignment system [3].

Oral vowel formant extraction was made thanks to the Burg algorithm implemented in Praat [4].For each formant, three values were computed on a given vowel segment (corresponding to locations at the first third, the middle and the last third of the segment). Then, these three measurements were averaged to provide a single formant value per vowel token. The filter was setup to take into account the vowel type and the speaker's sex. It excluded all the tokens with formant values not falling into preset frequency ranges. These ranges were chosen in a very tolerant way, based on Calliope [13] reference values. For instance, [i] was considered as erroneous if F1>750Hz (♂) or 900Hz (♀), and/or F2 values were not between 1500-2500Hz (♂) and 1600-3100Hz (♀).

About 4% of tokens were excluded. According to the authors, most of these vowel segments were very short in duration, or displayed devoicing, thus making formant detection more complicated. Compact high vowels (such as /u/) were also particularly prone to rejection as formant extraction sometimes couldn't distinguish F1 from F2.

## 3. REPLICATION OF THE G&A FF STUDY

### 3.1. ESTER subset

We selected a similar but different subset of 2h (1h ♀ and 1h ♂) from the ESTER corpus (namely France Inter radio files). This choice allows us to get an idea of the variation intrinsic to speech and speakers independently of a change in the filtering approach. Our subset gathers 26249 oral vowel tokens (against 24000 for [5]).

### 3.2. FF approach applied to the ESTER subset

Here we replicate the G&A study using their formant range filters. The formant values were also provided by the authors of[5].

Not surprisingly, we end up with rather comparable rejection rates. Among our 26k tokens, 3.2% were excluded by the filter against 4% in [5]. However, as can be seen in Table1, a lot more of /o/ vowels were excluded in our replication. This might at least partially be related to the very low number of /o/ occurrences in our subset (733 vs. 3824 for /i/, for instance).

Similarly, the mean formant values are quite comparable to [5], and even slightly closer to the reference values of [13]. This might be partially due to the choice of France Inter radio news, which is the most normative sample in ESTER.

Figure2 gives our (ESTER-FF - in light blue) and [5] (G&A - in coral) mean formant values, as well as Calliope's values [13] (in black) which were used as references for the filters. for /i/, /e/, /ɛ/, /a/, /ɔ/, /o/, /u/ to ease the comparison (only males data and a subset of vowels are presented for the sake of readability).

| | **i** | **y** | **e** | **ɛ** | **a** | **oe** | **ø** | **ɔ** | **o** | **u** |
|---|---|---|---|---|---|---|---|---|---|---|
| G&A | 5 | 15 | 1 | 0.3 | 0.6 | 4 | 0.4 | 1 | 4.9 | 25 |
| replic. | 2 | 5 | 0.3 | .07 | 0.3 | 1 | 1 | 3 | 17 | 27 |

**Table 1:** Proportion of rejected segments for each vowel type (in %) for [5] and our replication.

## 4. DATA-DRIVEN FORMANT FILTERS

We now get back to our sample of ESTER and filter it again but with our data-driven method (MD).

We consider a three-dimensional space made of vowel tokens' duration (s), mean F1 (Hz), and mean F2 (Hz). In addition to the traditional mean F1 and F2, we take into account phone duration for it may be helpful to get more "discriminating" profiles.

The filter was coded as a R script [14], mainly using the *mahalanobis* function from *mvtoutlier* package [15].

The range filter (as set-up by [5]) considered the vowel type and the speaker's gender. However we tried different ways to set up MD calculation. The profile of vowels was computed in three fashions: accounting for the vowel type (T) only, the interaction between T and the speaker (S), or the interaction between T and the speakers' gender (G) - the later being the closest to [5]. For each three set-ups, the computed distance had a mean of $\sim 2.9$. Thus we choose $MD > 3$ as our threshold to tag values as erroneous. In other words, for each set-up, each vowel token that was further than 3 standard deviation away from its computed profile was tagged erroneous. Table 2 gives the proportion of rejection per vowel type, as a comparison with [5]'s results given in Table 1.

| | **i** | **y** | **e** | **ɛ** | **ø** | **oe** | **a** | **o** | **ɔ** | **u** |
|---|---|---|---|---|---|---|---|---|---|---|
| T | 19 | 22 | 28 | 29 | 23 | 25 | 30 | 28 | 29 | 27 |
| T*S | 25 | 31 | 29 | 34 | 32 | 30 | 31 | 34 | 35 | 33 |
| T*G | 18 | 23 | 26 | 29 | 25 | 25 | 29 | 29 | 29 | 27 |

**Table 2:** Rejection rate for each vowel type in each set-up used for MD calculation (in %). T = vowel type (e.g. /i, u, .../), S = speaker, G = gender (F/M).

The method allowing to keep the most data while excluding the most extreme values was the one using vowel type and gender (T*G, 26% rejection), closely followed by T (only the vowel type, 27%

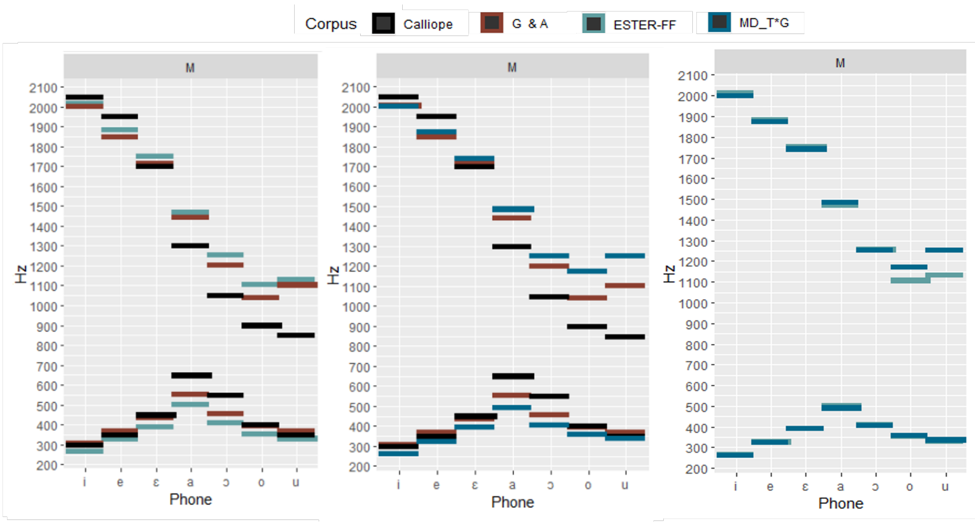

**Figure 2:** Comparison of formant structures for Calliope (Call) [13], [5] (G&A), our replication (ESTER-FF), and MD by Type and Gender (MD_T*G). Male speakers only.

rejection), T*S having higher rejection rate (31%).

For the three set-ups, we checked the resulting mean formant values and found pretty similar results both between our three set-ups, and between the MD filters and the formant range filter (see figure 2 where we only plot T*G - dark blue - for clarity).

Eventually, we ressorted to Linear Mixed Models (LMM)[16] to check the effects of the filters on the resulting formant values. To lower the number of levels for the factor âvowel typeâ, we grouped vowel types into 4 categories: Diffuse (/i, y, e, ɛ/), Compact (/o, u/), Mid (/ɔ, oe, ø/) and /a/ which was left in its own category for it is the only open and most variable vowel in French. For each of our LMMs, we compared different structures by adding fixed effects as we went along. These comparisons were made using likelihood ratio tests. When the LRT was significant ($\alpha < .05$), the most complete model was kept for further analysis. After model comparison the most parsimonious model was the one built as follow : $MeanFi \sim MethodOfFiltering * Gender * Phone + (1|Word) + (1|Speaker)$. The significance of each fixed factors was assessed using a Type III ANOVA.

Both models (on mean F1 and mean F2) showed a significant main effect of the method and its interactions with other IV as well as the 3 way interaction MethodOfFiltering*Gender*Phone (p<.03 for F1 and p<.001 for F2). However the analysis of the three way interaction and posthoc tests showed that the differences between the methods arised in very specific spots, such as males' compact vowels. The marginal estimated means for F2 differed by 152Hz max (compact ♂, FF having the lowest value), and 22Hz max for F1 (compact ♂, FF having the lowest value).

## 5. CONCLUSIONS

Mahalanobis distance was previously used by [8] as a way to choose the best measures during the formant computing phase but here we want to propose its use as a posthoc (post extraction) tool to correct datasets; as nowadays shared and open source data are available as spreadsheets, one might need a post processing tool. In this paper we wanted to put to the test the use of Mahalanobis distance, a data-driven method, as a way to filter erroneous values from acoustic phonetic datasets.

Our replication of [5] with a formant range filter and Mahalanobis distance showed that MD method have higher rejection rates then FF. The remaining tokens however have very similar formant values with both methods. One explanation is that MD method as we set it might remove more fine-grained variation that might be (socio)linguistically relevant. Thus we need to try several tightness degrees by moving the threshold for rejection (e.g. 4 or 5 standard deviations from the profile instead of 3) to have a less restrictive filter.

However MD is data-driven and gives a continuous variable (standard deviation from the profile computed on the data) as the index for rejection, meaning that the user can set it up according to the needs of the study in terms of data homogeneity. Therefor it is an efficient and adaptive tool for filtering extreme acoustic values.

## 6. ACKNOWLEDGEMENTS

The authors want to thank the Swiss National Science Fundation for the funds made available for this project, as well as the recorded speakers for the data we are using.

## 7. REFERENCES

[1] J. Stuart-Smith, M. Sonderegger, R. Macdonald, J. Mielke, M. McAuliffe, and E. Thomas, "Large-scale acoustic analysis of dialectal and social factors in english/s/-retraction," 2019.

[2] D. Amazouz, M.-A. Decker, and L. Lamel, "Variation du voisement des occlusives orales en code-switching: analyses par abx automatique et mesures acoustiques," in *Journées d'Études sur la Parole-JEP2022*, 2022.

[3] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[4] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[5] C. Gendrot and M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German." in *Interspeech 2005, Lisbon, Portugal*, 2005, pp. 2453–2456. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-00188096

[6] M. Lancien, "Le rôle de la réduction phonétique dans l'expression de la proximité sociale," Ph.D. dissertation, Université de Lausanne, 2021.

[7] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators," *Statistical Papers*, Nov 2019. [Online]. Available: http://dx.doi.org/10.1007/s00362-019-01148-1

[8] J. Mielke, E. R. Thomas, J. Fruehwald, M. McAuliffe, M. Sonderegger, J. Stuart-Smith, and R. Dodsworth, "Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several english speech corpora," 2019.

[9] K. Evanini, *The permeability of dialect boundaries: A case study of the region surrounding Erie, Pennsylvania*. University of Pennsylvania, 2009.

[10] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "Fave (forced alignment and vowel extraction) program suite," *URL http://fave. ling. upenn. edu*, 2011.

[11] M. Sonderegger, J. Stuart-Smith, M. McAuliffe, R. Macdonald, and T. Kendall, "Managing data for integrated speech corpus analysis in speech across dialects of english (spade)," 2022.

[12] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news." in *Interspeech*, 2005, pp. 1149–1152.

[13] Calliope (Firm), J. P. Tubach, and G. Fant, *La parole et son traitement automatique*. Paris ; Milan ; Barcelone : Masson, 1989.

[14] M. Lancien, "Filtering data for phonetics," 2021. [Online]. Available: https://github.com/M-Lancien/Filtering_data_for_phonetics

[15] P. Filzmoser and M. Gschwandtner, "mvoutlier: Multivariate outlier detection based on robust methods," 2018, r package version 2.0.9. [Online]. Available: https://CRAN.R-project.org/package=mvoutlier

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.