USING MAHALANOBIS DISTANCE TO FILTER ERRONEOUS VOWEL FEATURES IN LESS-RESOURCED LANGUAGES : APPLICATION TO QUEBEC FRENCH

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

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

ABSTRACT

A major challenge in building shareable datasets for phonetic studies consists of maximising data collections while minimising the errors involved, in particular when automatic processes come into play (e.g. labeling, formant detection...). Automatically extracting formants from large amounts of speech frequently produces artifacts, due to formant jumps, alignment problems, or noise. One solution is to use formant range filters. However this requires prior knowledge about the vowels, such as formant range information. Here we propose to use the Mahalanobis distance to remove erroneous values relying only on the labeled speech data. Our study is conducted on a Quebec French corpus including more than 170 k tokens of 16 vowel types. Results show that the proposed method can complement the threshold-based filter approach. Furthermore, it can be used autonomously for undocumented languages to eliminate erroneous values. The approach also makes it easy to adjust the degree of filtering.

Keywords: Vowels, Quebec French, Data Science, Automatic filtering, Corpus Phonetics

1. INTRODUCTION

In the context of automatically annotated medium to large size corpora, the data are generally numerous, heterogeneous, and noisy. Different kinds of errors arise from automated processes, such as formant-tracking jumps, transcription issues, or mismatches between the expected and the actual pronunciation of a word, among others. The said errors then lead to the presence of erroneous datapoints in the datasets. On very large datasets, erroneous values might not affect the results, but with medium-size datasets(e.g. with 4 or 5 tokens of each vowel type per speaker), more common in phonetic studies, erroneous values will likely bias the subsequent results. Thus medium-sized datasets need "cleaning" as far as possible.

Several methods can be used to identify erroneous values during or after the value extraction process (see [1] for methods applied during the formant extraction; see [2] for methods applied during the automatic alignment). Here we address the issue of "post processing" methods such as formant range filters (e.g.[3]). Formant range filters are the most common and work as follow :

1. the user sets up a series of value ranges between which formants must fall in order to correspond to the known/expected vowel profile
2. the filter go through and check the formant values extracted from the audio corpus
3. vowels with formant values oustside the ranges are discarded

The issue with those methods is that 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. [3] use ranges from 1500 to 2500Hz for French /i/ F2). However it does efficiently identify the most obvious erroneous values.

Here we present a tool for filtering datasets from low resource language, with no value ranges or vowel profiles available. Quite counter-intuitively, Quebec French (QF) can be considered as low resource when it comes to phonetic description or NLP. But its closeness to France French allow us to compare our results to some expected vowel profiles. Thus we test our tool on the QF dataset from [4].

1.1. Quebec French vowels

Quebec French vowel system includ 16 phonological vowels [5, 6] : /i, y, e, ɛ, ɜ, ø, oe, a, ɑ, ɔ, o, u/ and the 4 nasals /ẽ, ɛ̃, õe, ɑ̃/. These vowels undergo a large amount of phonological processes, namely laxing, diphtonguization, devoicing, syncope, according to specific phonological rules (described by [5, 6, 7] among others).

Most vowel qualities resulting from these phenomena can be grouped according to the notion

of vowel class [8, 9]. A vowel class represents a phoneme and its position in a syllable [10], for instance iR is the class representing phoneme /i/ followed by a lengthening consonant. It is a very important concept in QF for vowel position in the word or syllable strongly affect the allophone used.

Table 1 gives the inventory of the classes relevant for our study. Vowels in class _C they are followed by a coda (any consonant), vowels in _K are followed by a non lengthening consonant coda, those in class _R are followed by a lengthening coda (/ʁ,v,z,ʒ/), finally in class _# vowels are in absolute final. Class _# usually don't trigger any process. In class _K high vowels are laxed (e.g. "bicycle" /bisik/ [bisɪk] *bicycle*). Class _R usually triggers lengthening and diphtonguization for non-high vowels (e.g. "père" /pɛχ/ [paɛχ] *father*). Class _C is only relevant for the /A/s and can trigger dipthonguization for /ɑ/ or posteriorisation for /a/ (e.g. "pâte" /pɑt/ [paɔt] *pasta* - "là" /la/ [lo] *here*).

Phoneme	Vowel Classes
i	i# iR iK
y	y# yR yK
u	u# uR uK
e	e#
ɛ	ɛ# ɛK ɛR
ø	ø#
oe	oeK oeR
o	o# oK oR
ɔ	ɔ# ɔK ɔR
a	a# aC
ɑ	ɑ# ɑC

Table 1: Vowel classes inventory, inspired from [9] and corresponding to [4]'s corpus.

1.2. Mahalanobis distance as a filtering method

In [11], authors showed that Mahalanobis distance can be used as a tool for filtering. 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 Σ 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. In the case where $P = \mu D$, the distance is 0, this distance increases as P moves away from the mean in a determined space. Thus by choosing a threshold for the distance ([4, 11] chose 3 standard deviations), it is easy to discard any vowel token further then X standard deviation from the computed profile.

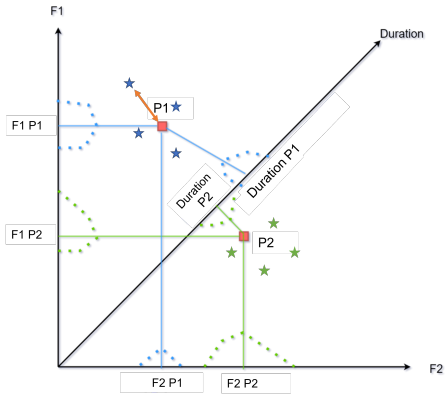

Figure 1: Schematic explanation of the Mahalanobis distance. Blue lines represent the distance of F1 (axis y) F2 (axis x) and duration (axis z) between the mean (orange squares) of two vowels (P1 and P2). Stars are tokens of P1 and P2 and the orange arrow indicate the distance measured.

2. METHODS

2.1. Quebec French data

The Quebec French dataset that we use here is part of [4]'s Ph.D work. They recorded 10 couples (10♂, 10♀), performing five tasks involving the pictures from the Diapix task [12, 13]: 1) reading, 2) image difference identification task alone, with a spouse, with an unknown investigator from the same region and finally with a stranger from France. A recording session lasted approximately 3h and the total corpus duration is 27.22h with a minimum of 1.19h of speech per speaker.

The recordings were all segmented in Speech Units and orthographically transcribed by hand. Author then used the SPPAS software (v. 2.7, adapted for QF phonetics and phonology; see [14]) for automatic segmentation and alignment (for more details see [4]). A total of 162 543 vowel tokens (belonging the 16 vowel classes introduced above) and their acoustic features were extracted from these recordings. Here we only use duration (s), mean F1 (Hz), and mean F2 (Hz).

2.2. Mahalanobis distances setup

In this work we consider a three-dimensional space made of the 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 [15], mainly using *mahalanobis* function from the *mvtoultier* package [16].

In the next section, we will investigate several settings for MD. As the data were build for a sociophonetic study we had more information than just the vowel class and the speakers' gender, and were able to try 4 different set-ups taking into account the vowel class, the speaker, their gender, or the speech condition. Thus for each 160k tokens of vowel we calculated a Mahalanobis distance between the phone (P) and:

- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class and speaker.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class and production condition.
- the mean of the distributions of mean F1, mean F2 and duration as a function of vowel class, speaker and production condition.

3. MAHALANOBIS DISTANCE AS A DATA-DRIVEN TOOL FOR PRUNING

The method allowing to keep the most data points while excluding the most extreme values was the one using only vowel class. The rejection rate was of 12% ($\sigma = 1.4$ depending on vowel classes), which is close the 10% reported by [17] on a similar work. The class ɛR shows the highest rejection rate (13.5%), which makes sense as these vowels can be both diphthongs or monophtongs, thus making the computation of a unique profile more complicated. The monologue condition lost just over 16% of its occurrences after filtering, compared to around 7.7% in Reading, 11.5% for spontaneous speech with a spouse, 13.6% with a stranger, and 13% with a foreign stranger. These style related results become very interesting in the light of phonostylistic findings and can be linked to work by [18, 19, 20, 21] (among others) on phonetic reduction and variation in speech, meaning that our method could be used efficiently for research in this field.

However, we have resorted to a more qualitative examination of the differences between the occurrences estimated as too extreme by our different filters in order to have a sharper vision of the different performances of these methods. Thus we examined the occurrences classified as valid or erroneous for three classes at the extremities of the system: i#, u#, and ɑ#. We have specifically chosen these three classes for the following reasons: i# are among the most stable in terms of phonological variation, and raise very few problems for formant detection (except for F1 detection when contextual devoicing happens); u# also have this stability, but represent a greater challenge for formants extraction due to compactness, very low F2, and the possibility of devoicing; as for ɑ#, they have a wide range of phonological variability, going from [a] to [ɔ], and can trigger detection problems when F1 and F2 are too close (e.g. $\sim$ 1000Hz).

Table 2 summarises the observations made for each filter and for the three classes. The table reads as follows: with the filter only using vowel class, out of the 3 171 tokens of i# tagged as erroneous 8% were actually valid datapoints (i.e. with mean F1s between 100Hz and 500Hz, mean F2s between 1800Hz and 2800Hz, and duration under 0.5s)[1], over the 570 u# tagged as erroneous, 21% were valid ($100 < F1 < 500$Hz, $600 < F2 < 1200$Hz, $dur < 0.5$s), and over the 1 699 tokens of ɑ# tagged as erroneous 17.5% were valid ($400 < F1 < 1$KHz, $800 < F2 < 1700$Hz, $dur. < 0.5$s). On the other hand, some tokens which should have been tagged as erroneous were not: for class ɑ# over 15 495 tokens classified as valid datapoints, about 23% should have been excluded by the filter ($F2 > 1800$Hz), for i# it drops to 8% ($F1 > 500$Hz & $F2 < 1600$Hz) of the 22,528 tokens, and about 27% of 3676 for u# ($F1 > 500$Hz & $F2 > 1500$Hz).

Filter	Diag.	i#	u#	ɑ#
class	WK	8%	27%	23%
	WD	8%	21%	17.5%
class*spkr	WK	6%	23%	21%
	WD	30%	28%	70%
class*sex* condition	WK	8%	27%	25%
	WD	30%	24%	57%
class* condition	WK	7%	28%	23%
	WD	34%	23%	52%

Table 2: Diagnostics of phones wrongly filtered/discarded (WD) or wrongly kept (WK) according to the different filters setups on the QF data.

Overall, MD filtering method did well with cases of devoicing (very common in QF), such as in the word "édifice" *building* /edifis/ where the first [i] is devoiced and drown in the affrication noise of /d/. However it missed some formants jumps in compact back vowels such as [ɔ] in [ʃprɔpoz] (reproduced in figure 2). In this example, F2

was wrongly detected around 1800Hz when it was actually around 1100Hz, merged with a high F1.

A few questionable events also arised. Figure 3 shows a case in which the vowel realization is very far from the average profile of the class: the /o/ is realized with an F2 closer to a [œ] (F1 at 370Hz and F2 at 1732Hz). This fronting movement was documented by [22, 10] as a rare but possible sociophonetic variation. Thus our filter might have removed some important tokens for the study of variation.

Finally table 3 gives the mean formants values that result from our filtered data. QF i#, u# and a# are often described as similar to France French /i, u, a/, and the formant values we report for these three vowels in QF are very close to what was found in France French reference values [23, 3], which is why we want to use them as a further evidence of the process's success.

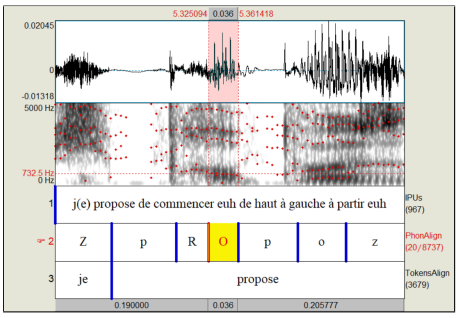

Figure 2: Spectrogram and oscillogram illustrating a case of poor detection of F2 in a [ɔ] from the word "propose" /prɔpoz/ *suggest*.

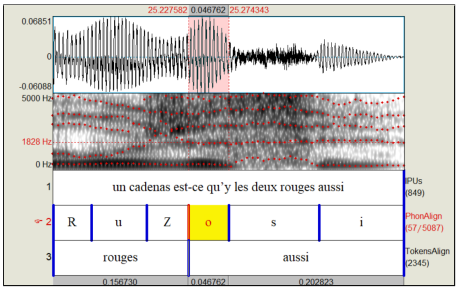

Figure 3: Spectrogram and oscillogram illustrating the non canonical realization of an /o/ in the word "aussi" /osi/ *also*.

.

4. CONCLUSIONS

The "by class" filtering process we used on the FQ corpus identified 12% of vowel tokens as erroneous. The number may seem very (too) substantial, however on a corpus of such magnitude, for which automatization was the main method, it is not unrealistic to consider that certain types of recurring errors have crept into segmentation and formant detection. In a way our rejection rate is pretty close to what [17] reported.

	mean F1		mean F2	
V. Class	F	M	F	M
i#	393	348	2184	1957
iK	435	368	2118	1896
iR	423	348	2191	2004
y#	425	396	1863	1717
yK	449	393	1788	1630
yR	474	425	1540	1473
e#	426	376	2139	1903
ɛ#	560	501	1858	1638
ɛK	539	474	1942	1692
ɛR	652	561	1748	1559
ø#	405	365	1690	1517
a#	609	532	1641	1516
aC	723	616	1627	1458
ɑ#	586	542	1325	1238
o#	442	407	1039	988
ɔ#	574	518	1436	1322
ɔK	562	513	1250	1180
u#	418	386	1299	1219
uR	443	398	1099	1012

Table 3: Mean formant values for each vowel class after the filtering by class only.

The classes experiencing the highest rejection rate were those with vowels that can be both diphthongs or monophtongs (e.g. ɛR), making the computation of a unique profile more complicated; while less variable classes had smaller rejection rates (e.g. 9.9% for ɑ#).

The differences of rejection rates between speech styles also demonstrate that MD filter can be used widely without previous information on style and will not erase style based differences. It also gives information on the variation rate of style, as the most standard style (Reading) rejection rate was of 7.7%, as opposed to 16% in self directed speech; which is in line with phonostylistic findings such as [18, 19, 20, 21].

However, more investigation is needed on the efficiency of Mahalanobis distances for filtering purpose. A comparison with [1]'s result would state whether MD is more efficient before or after features extraction. Different thresholds must be tested to evaluate which distance from the profile is the most suitable for phonetic studies and their different goals. Various type of data from high and low resources languages, different speech styles, and speaker profiles, should be tested. Further work also need to investigate on the effect of lexical categories and frequencies known to affect variation.

5. ACKNOWLEDGEMENTS

The authors want to thank the Swiss National Science Fundation as well as the recorded speakers.

6. REFERENCES

[1] 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.

[2] D. Amazouz, M. Adda-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] 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

[4] M. Lancien, "Le rôle de la réduction phonétique dans lâexpression de la proximité sociale," Ph.D. dissertation, Université de Lausanne, 2021.

[5] D. C. Walker, *The pronunciation of Canadian French*. University of Ottawa press Ottawa, 1984.

[6] L. Santerre, "Voyelles et consonnes du français québécois populaire," in *Identité culturelle et francophonie dans les Amériques*. Québec : Presses Universitaires de Laval, 1976, vol. 1, pp. 21–36.

[7] M.-H. Côté, "Laurentian french (quebec) : extra vowels, missing schwas and surprising liaison consonants," in *Phonological variation in French: Illustrations from three continents*, R. Gess, C. Lyche, and T. Meisenburg, Eds. Amsterdam : John Benjamins, 2012, pp. 235–274.

[8] W. Labov, *Principles of Linguistic Change: Internal Factors*. Oxford, Blackwell, 1994, vol. 1.

[9] C. Paradis, "An acoustic study of the variation and change in the vowel system of chicoutimi and jonquière (québec)," Ph.D. dissertation, University of Pennsylvania, 1985.

[10] M. Yaeger, "Context-determined variation in montreal french vowels." Ph.D. dissertation, University of Pennsylvania, 1979.

[11] M. Lancien, J. Stuart-Smith, and M. Adda-Decker, "Knowledge-driven vs. data-driven methods for filtering acoustic measures in phonetics corpora," in *Proceedings of the 20th International Congress of Phonetic Sciences, ICPhS 2023, Prague, Czech Republic, 7-11 August 2023*, submitted.

[12] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and speech*, vol. 53, no. 4, pp. 510–540, 2010.

[13] R. Baker and V. Hazan, "Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs," *Behavior research methods*, vol. 43, no. 3, pp. 761–770, 2011.

[14] M. Lancien, M.-H. Cote, and B. Bigi, "Developing resources for automated speech processing of quebec french," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5325–5330. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.655

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

[16] 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

[17] 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.

[18] J.-L. Rouas, M. Beppu, and M. Adda-Decker, "Comparison of spectral properties of read, prepared and casual speech in french," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., 2010.

[19] N. Audibert, C. Fougeron, C. Gendrot, and M. Adda-Decker, "Duration-vs. style-dependent vowel variation: a multiparametric investigation," in *Proceedings of the 18th International Congress of Phonetic Sciences, ICPhS 2015, Glasgow, UK, 10-14 August 2015*, 2015.

[20] M. Lancien and M.-H. Côté, "Phonostyle et réduction vocalique en français laurentien," in *6e Congrès Mondial de Linguistique FranÃ§aise, 9-13 juillet 2018, Mons, Belgique*, vol. 46, 2018.

[21] M. Lancien, "Variation stylistique en français québécois : l'effet de l identité de l'interlocuteur," in *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, and S. Schneider, Eds. Nancy, France: ATALA, 2020, pp. 308–316. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02812571

[22] J.-D. Gendron, *Tendances phonétiques du français parlé au Canada*. Paris : Klincksieck, 1966.

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

[1] We willingly take a wide range of variation to take into account the different possible degrees of reduction.