Towards the automatic identification of /l/-vocalisation in English speakers in Australia

Simon Gonzalez[#1], *Gerard Docherty*[*2]

[#]Australian National University, Australia
[*]Griffith University, Australia
[1]u1037706@anu.edu.au, [2]gerry.docherty@griffith.edu.au

Abstract

The aim of this paper is to describe the initial development of a computational framework designed to automatically recognize and classify vowel-/l/ rhyme realisations produced by Australian English speakers as either consonantal or vocalised. We implemented a Random Forest model as the main classificatory technique. This allowed us to explore in a hierarchical way the contribution to the classification of a wide a range of potential predictors. The test classification accuracy of the Random Forest model was 82.1% overall, with its sensitivity estimated to be 73.7% (consonantal realisations) and the specificity to be 89.1% (vocalised realisations).

Index Terms: /l/-vocalisation, machine learning, Australian English, sociophonetics

1. Introduction

Advances in forced alignment and segmentation have transformed the toolkit available to researchers interested in tracking the variable performance of speakers in large-scale corpora of natural speech. The spectral and time-domain properties of vowels are well-understood and key parameters such as formant frequencies can be extracted automatically from corpora using off-the-shelf tools such as FAVE [1] or MAUS [2]. In recent years progress has been made in automatically extracting some of the variable realisational properties of consonants (e.g. Voice Onset Time (VOT) [3] and the spectral properties of fricative realisations [4]). Some phonetic properties of speech, however, present greater challenges for automatic identification and classification, especially in spontaneous speech styles.

One such property is the variable realisation of post-vocalic /l/ across different varieties of English. In syllable-coda position (most notably in pre-pausal position or in the context of a following syllable that begins with a consonant) /l/ can be realised as a canonical lateral approximant with a central occlusion of the oral cavity and uni- or bi-lateral airflow. With concurrent voicing, this gives rise to a spectral signature not unlike that of a vowel segment, but typically with discontinuities associated with the formation of the central occlusion. However, if the occlusion for a coda /l/ is not fully formed, the /l/ is said to have a vocalised realisation and its auditory and acoustic properties are rendered much more vowel-like with the result that it can be hard to distinguish a vocalised /l/ from back rounded vowels (which have a very similar articulatory configuration).

This phenomenon has been reported in a number of languages across the world (e.g. Swiss German [5], Dutch [6], and Portuguese [7]). In English, /l/-vocalisation is a widely observed phenomenon across many varieties [8] [9] [10], including those spoken by speakers in Australia [11] [12] where its occurrence has been found to be influenced by speaker age and provenance, as well as by a range of context-dependent factors. For example, as well as finding differences across different cities and age-groups, Horvath and Horvath [11] found that high vowels are more likely to trigger /l/-vocalisation than low vowels, and that back vowels are more likely to be associated with vocalisation than front and central vowels. The same study reported a greater likelihood of vocalisation following phonologically long monophthongs than following diphthongs and short monophthongs, and differences as a function of the place of articulation of an adjacent consonantal context. It was also found that tokens followed by a consonant were more likely to be vocalised than those occurring in pre-pausal and pre-vocalic contexts. In general, however, it is fair to say that this is a rather under-documented feature of the phonetic variability characterising English speakers in Australia.

Experimental studies of the occurrence of /l/-vocalisation in English have largely made use of auditory [13][14] and/or articulatory methods [15][16][17] [18][19]. These studies bring into question the nature of the phenomenon in itself, pointing to vocalisation being less obviously a categorical process as opposed to being located at one end of an articulatory light-dark articulatory/auditory continuum, albeit a continuum that seems to embed an auditory discontinuity given researchers' confidence in detecting at least some V-/l/ tokens as clearly vocalised [14], with the latter sort of auditory judgement constituting the basis for most work to date on the socio-phonetic/-linguistic aspects of this phenomenon.

This represents quite a challenge for sociophonetic research which generally seeks to quantify the occurrence of discrete variants within large speech corpora across speakers, styles, locations, or to track gradient acoustic measures that are known to be valid metrics (e.g. formant values, VOT, or spectral Center of Gravity (CoG) for fricatives) of a particular gradient realisational variant. In order to achieve this for /l/-vocalisation, what would be ideally be required is an acoustic measure mapping to the articulatory continuum allowing investigators to test for the occurrence of any socially-correlated variation within the gradient realisation of coda /l/, but in reality, for now, investigators remain largely reliant on a by-hand auditory analysis that seems ill-suited to the phenomenon which it is attempting to capture (leading [14] to note that "the precise phonetic difference between velar (L) and vocalized (L) is one of the more subtle variable distinctions in sociophonetic research and presents one of the biggest methodological challenges").

In this study, we investigate the possibility that a machine learning method might provide a means of addressing some of the methodological challenges associated with vocalised /l/. The question addressed is whether there are patterns in the signal that can differentiate even conservatively-defined vocalised/non-vocalised /l/s that can be learned and applied automatically with a reasonable degree of effectiveness? If this proves to be the case, it would potentially offer a means of enabling researchers to extend the prevalent binary classification of /l/-vocalisation to larger corpora. While our study is not designed to resolve the question of whether /l/-vocalisation is best characterised as a binary variant or as one manifestation of a more complex realisational continuum, it does nevertheless provide a means of gauging if there is mileage in pursuing an automatic approach to the binary auditory judgement that predominates in this line of research.

2. Model Development

The approach adopted in this study drew on pre-existing reports of the application of machine learning methods in automatically classifying a range of different types of realisational variant, such as clear or dark /l/ in speakers of English from the USA [20], and variants of post-vocalic /r/ and medial /t/ in NZE [21]. In the approach described below, a statistical model is first trained to map between a set of acoustic parameters and the vocalised/non-vocalised labels that are to be classified. Subsequently the effectiveness of the model to classify new material is tested in order to gauge the predictive power and effectiveness of the statistical model [22]. To our knowledge, this is the first application of this type of machine learning methodology to the classification of /l/-vocalisation. This particular realisational variation is a good testbed for this approach to phonetic classification, given the potential that the specific classification method adopted has to deal with a wide range of candidate predictive parameters.

2.1. Creation of a test dataset

This study is part of a larger project investigating sociophonetic variability in the performance of English speakers from Perth in Western Australia. The dataset around which our analytic approach was developed and tested therefore comprised tokens of post-vocalic /l/ produced by young speakers of West Australian English (from the Perth metropolitan area). Recordings were obtained from two different speech styles: word-lists (12 native speakers of West Australian English – 6 females and 6 males, aged 18-22, producing a 165 item wordlist designed to elicit citation form/carefully produced tokens of key variables) and conversational speech (2 * 4 same-sex speakers recorded while participating in unscripted dyad conversations of around 30 minutes in duration).

Audio files were first segmented in Elan [23] and subsequently force-aligned within LaBB-CAT [24], using CELEX [25] and HTK [26] with manual correction of alignments. Using Praat [27], we extracted formant estimates (F1, F2, and F3), intensity, F0, and Mel Frequency Cepstral Coefficient (MFCC) tracks for each speech recording.

In order to create a dataset that could be used for benchmarking, training and testing the classification methodology, we selected from the above material 334 tokens of /l/ preceded by monophthongs and followed by either a pause or a consonant. For the purposes of this initial development phase of our work, we excluded tokens where /l/ was followed by an approximant, vowel or other liquids, in order to avoid dynamic formant trajectories beyond the offset of /l/. Table 1 summarises the characteristics of this reference dataset.

Table 1: *Count of all tokens in the reference dataset including the benchmark type of realisation as determined by the trained listeners' auditory classification*

Speech Style	Gender	Realisation	%	Count
Word List	Female	Vocalised	13.4	45
		Consonantal	16.4	55
	Male	Vocalised	13.8	46
		Consonantal	16.2	54
Conversation	Female	Vocalised	17.4	58
		Consonantal	9.6	32
	Male	Vocalised	10.2	34
		Consonantal	3	10
Total			100	334

A perceptual task was carried out to create the benchmark classifications necessary for evaluating the automatic coding framework, i.e., whether the automatic prediction from the machine learning model matches with the corresponding type of /l/-realisation. For this task three phonetically and sociolinguistically trained listeners coded each of the tokens containing the vowel+/l/ rhymes. Each rater listened to all of the tokens played in random order, and classified them into one of the following categories: Consonantal /l/, Intermediate, or Vocalised /l/ [28]. Where two of the three raters agreed on a particular classification, the majority view was adopted. If raters had no agreement on their auditory classification, all three raters consulted and reached a consensus decision after further listening to the token concerned. For the purposes of this initial test of the automatic classification methodology, we grouped the Consonantal and Intermediate classifications into one single group, following [28], thereby provide a binary-classified benchmark reference dataset, with the Vocalised set comprising those tokens that were most readily classified auditorily as containing vocalised /l/ variants.

Figure 1 presents /l/-realisation rates as coded in the auditory test pooling across the two speech styles investigated. It shows that all speakers vocalise /l/, but with substantial inter-speaker variation, ranging from speakers who mainly vocalize to speakers who mainly produce canonical realisations.

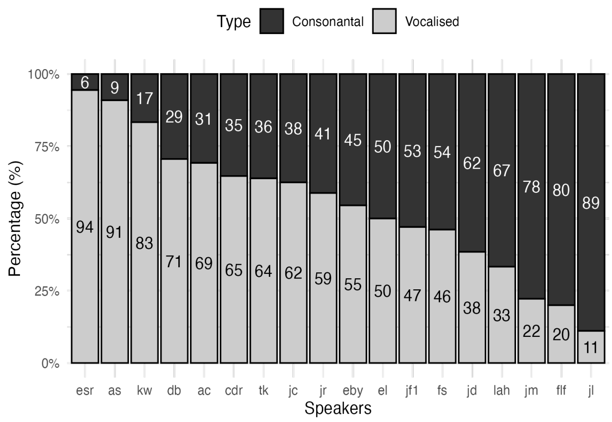

Figure 1: *% /l/-realisation type rates per speaker as per auditory test pooled across speech styles*

As has previously been noted in the literature (see above), the patterns of /l/-realisation are influenced by the identity of the vowel component of the rhyme (as shown in Figure 2). Back and central vowels (GOOSE, FOOT, THOUGHT, NURSE, SCHWA) are associated with relatively high levels of vocalised realisations across the two speech styles (bottom lighter colors), whereas front and low vowel contexts yield relatively fewer tokens of vocalisation. There are also differences across speech styles for some vowel contexts, with a relatively higher level of vocalisation found in word-list style for low and front vowels but a more even distribution across styles for the back and central vowel contexts.

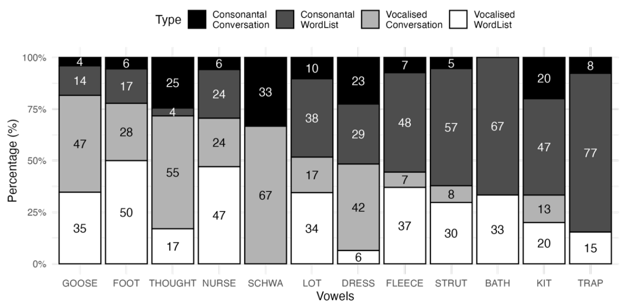

Figure 2: *% /l/-realisations by Vowel and Speech Style. The top two darker colours are consonantal realisations and the bottom two lighter colours are vocalised realisations.*

2.2. Parameterisation of the data

Following [29] and [30], for all of the selected vowel+/l/ sequences, we captured acoustic data from the onset of the vowel to the offset of the /l/. For each vowel+/l/ rhyme token the acoustic parameters referred to above (formants, intensity, F0 and MFCCs) were extracted at 11 equidistant points through an R [31] script developed for this purpose. After this, the dynamic trajectory of each acoustic measure was parameterised using a Discrete Cosine Transformation (DCT) of the 11-point trajectories. This then gave for each token a set of single values (the DCT coefficients) characterizing the trajectory for each acoustic parameter. The DCT output consists of C0-C3. C0 captures the mean amplitude, C1 the linear slope (whether it is flat or not), C2 the curvature, and C3 the amplitude at higher frequencies. This allows the learning algorithm to test which aspect of the signal may have the strongest predictive power in the classification task being undertaken. For all trajectories, the DCT transformation also smoothed out point extraction errors from the signal. A set of non-acoustic parameters accessible from the forced alignment were also built into the model including factors such as quality of the preceding vowel (on a front-back, close-open dimension), and information regarding the place and manner of articulation and voicing of adjacent consonants. This multi-faceted parameterisation of the vowel+/l/ dataset generated many potential predictors, acoustic and non-acoustic. The learning task was to determine which if any provided an effective basis for classifying tokens as vocalised or not.

2.3. Predictor Optimisation

The modelling was carried out in training, testing and assessment phases. In the training phase, given our initial broad set of variables extracted from the data (including speaker characteristics, vowel information, phonological context, and acoustic features, as described above), we applied the Boruta feature selection algorithm [32] as a Variable Importance Measure designed to identify the features of a dataset most relevant to the classification task. We ran the Boruta package in R [31], implementing a feature selection algorithm that identifies the relevant variables in a dataset [33] by iteratively removing the features which are less relevant for classifying the data. The power of this approach lies in its application to datasets in which a large number of variables need to be modeled in the absence of prior knowledge regarding which factors are likely to be the most relevant. Another crucial reason is that it has been demonstrated that there are machine learning algorithms that show a decrease in accuracy when the number of variables is higher than optimal [34]. The Boruta analysis pointed to 11 variables which were most important in distinguishing between the categories to be classified, as shown in Table 2.

Table 2: *Important parameters from the Boruta algorithm (Gini Importance). The ** represent those parameters that were not significant in the final Random Forest Model.*

Variable	**Description**	**Boruta Importance**
MFCC 5 c1	DCT Slope (coeff. 5)	**21.2**
F1 (50%)	F1 value at 50%	**20.6**
Intensity c0	DCT Mean amplitude	**20.5**
MFCC 5 (80%)	MFCC coeff. 5 at 80%	**19.6**
Intensity (20%)	Intensity at 20%	**19.2**
MFCC 8 (80%)	MFCC coeff. 8 at 80%	**16.7**
Advancement	Vowel advancement	**5.7**
MFCC 2 (20%)	MFCC coeff. 2 at 20%	**4.2**
MFCC 7 c2	DCT Curvature (coeff. 7)	**
Duration (ms)	Duration of trajectory	**
F1 c0	DCT Mean amplitude	**

2.4. Classification model training and testing

The 11 variables referred to above were chosen to be input parameters in a Random Forest (RF) model [35] implemented by using the randomForest package in R. This runs a combination of tree predictors, with each tree randomly selecting multiple predictive variables and evaluating their contribution to the classificatory task. In the end, all variables are ranked and the ones with lower error rates are classified as more important for classifying the categories specified. For this, the data was split into a training dataset and a test dataset used for validation and prediction. For the training dataset, we randomly selected 75% of the full dataset. For the test dataset, we selected the remaining 25%. In the first stage, the RF model is trained using the training dataset. Then in a second stage, the algorithm tries to predict the /l/-realization type on the test dataset, which it has not seen before. With this, we avoid overfitting on the predictability power of the model if we only test accuracy on data that the predictor has already seen.

The RF model was run using a classification method to predict the benchmark 'correct' classification of the /l/-realisation (i.e. the outcome that the RFs aims to predict) provided by the auditory assessment. The results from this model showed that out of the eleven variables initially identified in the parameter optimisation stage of our analysis, eight proved to be important variables when running the classification, as shown in Table 2.

2.5. Estimating model effectiveness

The RF modelling was run with 500 trees, with the number of variables tried at each split (mtry) evaluated at 2. The overall test classification accuracy of the RF model was 82.1%. The correct classification of consonantal types (sensitivity) had 73.7% accuracy; i.e. over 7/10 Consonantal realisations were correctly classified as true positives by the model (with 3/10 as false negatives), while the correct classification of vocalised types (specificity) had 89.1% accuracy; i.e. almost 9/10 of the Vocalised realizations were correctly classified (with 1/10 as a false negative). The effectiveness of the model classification was somewhat variable across conditions; the most effective overall classification (87%) was found following a front vowel nucleus – probably an environment where the spectral differentiation of the two variants is relatively high.

3. Discussion

The integration of machine learning and linguistic analysis in applications such as automatic speech recognition has demonstrated that robust phonetic classification models can be developed with a great degree of accuracy. In this work, we have developed a promising model for applying these techniques to the automatic classification of /l/-realisational variants in speakers of one variety of English, an approach that approximates the 'conservative' categorical judgements of /l/-vocalisation that have prevailed in the sociophonetic literature to date. While caveats of course apply to our findings (not least in relation to the relatively small-scale benchmark dataset that we have deployed), our results suggest that there would be value in further refining this approach. A first step would be to test the modelling framework with a significantly larger benchmark dataset. It would also be important to further explore ways of optimizing the statistical model and the criterial acoustic parameters, with a view to reducing false positive classification of vocalized tokens, and we need to systematically assess reasons why classification effectiveness varies across different phonological contexts. In pursuing the latter question, one challenge is perhaps the somewhat opaque mapping between MFCC representations and the acoustic/articulatory space that we are most familiar with in accounting for contextual and coarticulatory variation in speech performance [36].

A further limitation of our approach is that it is predicated on a binary categorization of /l/-realisations and therefore does not lead us any closer to automatically capturing from large corpora socially-correlated variability in the realisation of the broader /l/-colouring continuum (of which vocalisation is most likely just one aspect). Having said that, as pointed out in [14], a binary classification of this phenomenon does seem to map to a certain extent to listener's auditory judgements of the vocalised /l/, so there is some merit in attempting to achieve automatic coding of this variable that approximates that which listeners can provide. This exemplifies a more general point of debate within sociophonetic research; namely a tension between insights gained from acoustic and/or articulatory investigations with methods that can identify deep layers of fine-grained cross-speaker variation, and other insights gained from auditory analysis based on less granular categories. This is a tension that has been captured by Labov's [37] concern about an "endless pursuit of detail", but ultimately it is a tension that underscores the need to investigate sociophonetic variability from the point of view of both speakers and listeners.

4. Conclusions

Our findings demonstrate that a Random Forest approach to modelling, deploying a range of acoustic and other parameters together with the application of a Variable Importance Measure, is a promising basis for undertaking automatic acoustic coding of vocalised /l/-realisations by speakers of Australian English. Future development of this approach will explore the relative effectiveness of the model in different phonological contexts, as well as with a much larger dataset than that which was used here in this exploratory proof-of-concept study. There is no doubt that /l/-vocalisation warrants much more investigation across many varieties/languages (including the many other varieties of English in Australia and NZ). The work reported above suggests that, with effective further refinement, automatic classification methods have potential to form part of the toolkit for this task.

5. Acknowledgements

This research was undertaken with support from the Australian Research Council (DP130104275). We also wish to thank three anonymous reviewers for their helpful comments on the shape and content of the submission.

6. References

[1] Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. and Yuan, J., "FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2", 2014.

[2] Strunk, J., Schiel, F. and Seifart, F., "Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS", in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis [Eds], Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14, 3940-3947, Reykjavik, Iceland, 2014.

[3] Arias-Vergara, T., Arguello-Velez, P., Vásquez-Correa, J.C., Nöth, E., Schuster, M., Gonzalez-Rátiva, M.C. and Orozco-Arroyave, J.R., "Automatic detection of Voice Onset Time in voiceless plosives using gated recurrent units", Digit. Signal Process., 104, 102779, 2020.

[4] Frid, A. and Lavner, Y., "Spectral and textural features for automatic classification of fricatives", XXII Annual Pacific Voice Conference (PVC), 1-4, 2014.

[5] Leemann, A., Kolly, M., Werlen, I., Britain, D. and Studer-Joho, D., "The diffusion of /l/-vocalization in Swiss German", Language Variation and Change, 26(2): 191-218, 2014.

[6] Jongkind, A. P. and van Reenen, P., "The vocalization of /l/ in standard Dutch", in A. Timuska [Ed], Proceedings of the IVth International Conference of Dialectologists and Geolinguists, University of Latvia, Riga, 28.7.2003. Riga: Latvian Language Institute, 1-6, 2007.

[7] Callou, D. and Leite, Y., "Iniciação à Fonética e a Fonologia", Editira Zahar, Rio de Janeiro, 1990.

[8] Hardcastle, W. and Barry, W., "Articulatory and perceptual factors in /l/ vocalizations in English", Journal of the International Phonetic Association, 15:3-17, 1985.

[9] Ash, S., "The vocalization of /l/ in Philadelphia", Doctoral dissertation, University of Philadelphia, Pennsylvania, PA, 1982.

[10] Bauer, L., "English in New Zealand", in Burchfield, R. [Ed], The Cambridge History of the English Language. English in Britain and Overseas: Origins and Development, vol. 5. Cambridge University Press, Cambridge, 82-429, 1994.

[11] Horvath, B. and Horvath, R. J., “The geolinguistics of /l/-vocalization in Australia and New Zealand”, Journal of Sociolinguistics, 6(3): 319-346, 2002.
[12] Borowsky, T. and Horvath, B., “L-vocalization in Australian English”, in F. Hinskens, R. van Hout, and W. L. Wetzels [Eds], Variation, change and phonological theory, 101-123. Amsterdam: John Benjamins, 1997.
[13] Loakes, D., “An Investigation of the /el/-/ael/ Merger in Australian English: A Pilot Study on Production and Perception in South-West Victoria”, Australian Journal of Linguistics, 34: 436-452, 2014.
[14] Hall-Lew, L. & Fix, S., “Perceptual coding reliability of (L)-vocalization in casual speech data”, Lingua, 122: 794-809, 2012.
[15] Lin, S. and Demuth, K., “Children’s Acquisition of English Onset and Coda /l/: Articulatory Evidence”, Journal of Speech, Language, and Hearing Research, 58: 13-27, 2015.
[16] Lin. S., Palethorpe, S. and Cox, F., “An ultrasound exploration of Australian English /CVl/ words’, ASSTA, Sydney, Australia, 2012.
[17] Turton, D., “Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English”, Journal of the Association for Laboratory Phonology 8(1): 1-31, 2017.
[18] Szalay, T., Benders, T., Cox, F. and Proctor, M., “Lingual configuration of Australian English /l/”, Proceedings ICPhS, 2019.
[19] Scobbie, J.M. and Pouplier, M., “The role of syllable structure in external sandhi: an EPG study of vocalization and retraction in word-final English /l/”, Journal of Phonetics, 38: 240-259, 2010.
[20] Yuan, J. & Liberman, M., “Investigating /l/ variation in English through forced alignment”, Proceedings Interspeech, 2215-2218, 2009.
[21] Villarreal, D., Clark, L.S., Hay, J. and Watson, K., “From categories to gradience: Auto-coding sociophonetic variation with random forests”, Laboratory Phonology, 11(1): 1-31, 2020.
[22] Malakar, M., and Keskar, R.B., “Progress of machine learning based automatic phoneme recognition and its prospect”, Speech Commun., 135: 37-53, 2021.
[23] Sloetjes, H. and Wittenburg, P., “Annotation by category – ELAN and ISO DCR”, in Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.
[24] Fromont, R. and Hay, J., “LaBB-CAT: An annotation store”, University of Otago, Dunedin, New Zealand: Australasian Language Technology Workshop (ALTA), 4-6 Dec 2012, in Proceedings, 10: 113-117, 2012.
[25] Baayen, R. H., Piepenbrock, R. and Gulikers, L., “CELEX2 LDC96L14. Web Download”, Philadelphia: Linguistic Data Consortium, 1995.
[26] Young, S., Evermann, G., Gales, M. et al., “The HTK Book (for HTK Version 3.4)”, Cambridge University Engineering Department, 2006.
[27] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer [Computer program]”. Version 6.2.19, retrieved 12 September 2022 from http://www.praat.org/.
[28] Durian, D., “The vocalization of /l/ in urban blue collar Columbus, OH African American Vernacular English: a quantitative sociophonetic analysis”, OSU Working Papers in Linguistics 58: 30-51, 2008.
[29] Stuart-Smith, J., Lennon, R., Macdonald, R., Robertson, D., Sóskuthy, M., José, B. and Evers, L., “A dynamic acoustic view of real-time change in word-final liquids in spontaneous Glaswegian”, 18th ICPhS Proceedings, Glasgow, UK, 2015.
[30] Plug, L. and Ogden, R., “A parametric approach to the phonetics of postvocalic /r/ in Dutch”, Phonetica, 60: 159-86, 2003.
[31] R Development Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2021.
[32] Kursa, M.B. and Rudnicki, W.R., “Feature Selection with the Boruta Package”, Journal of Statistical Software, 36(11): 1–13, 2010.
[33] Stoppiglia H, Dreyfus G, Dubois R. and Oussar, Y., “Ranking a Random Feature for Variable and Feature Selection”, Journal of Machine Learning Research, 3: 1399-1414, 2003.
[34] Kohavi, R. and John, G.H., “Wrappers for Feature Subset Selection”, Artificial Intelligence, 97: 273-324, 1997.
[35] Breiman, L., “Random Forests”, Machine Learning, 45: 5-32, 2001.
[36] Iskarous, K., “The encoding of vowel features in Mel-Frequency Cepstral Coefficients”, 2018, downloaded 10.7.22 from https://www.aisv.it/StudiAISV/2018/vol_4/001_Iskarous.pdf
[37] Labov, W., “A sociolinguistic perspective on sociophonetic research”, Journal of Phonetics, 34: 500-515, 2006.