Acoustic analysis of L2 English phonemes for language identification

Samantha Williams, Vincent Hughes, and Paul Foulkes

University of York, York, UK
samantha.williams@york.ac.uk, vincent.hughes@york.ac.uk, paul.foulkes@york.ac.uk

ABSTRACT

This study provides an initial step towards developing a phonetically interpretable method of language identification to determine a speaker's L1 given their L2 English speech. 112 male L2 English speakers representing five L1s were selected from the Speech Accent Archive. F1, F2, and Duration were measured for four monophthongal vowel phonemes (/ɪ/, /i/, /ɛ/, and /ɔ/). These phonetic features were then used to calculate by-phoneme probabilities using the multivariate kernel density method. Findings suggest that for computational analysis, individual segmental acoustics in isolation are unlikely to be sufficient for making good predictions. Additionally, the features that are most effective for identifying a language depend on both the segment and comparison under investigation.

Keywords: L2 English, speech acoustics, language identification, computational analysis

1. INTRODUCTION

Language identification systems, as the name suggests, attempt to automatically identify the native language (L1) of a speaker based on a sample of his or her speech. Such systems have wide usage for commercial purposes (e.g., as a front-end for speech recognition) and for forensic and security purposes (e.g., to identify the language being spoken by a person of interest in an incriminating recording).

Most language identification systems approach this task from either an acoustic-phonetic approach (e.g. [1,2]) exploiting segmental realisations, or a purely acoustic approach which works on more abstract, mathematical representations of the speech signal extracted across an entire recording (such as MFCCs; e.g. [3]).

Current state-of-the-art systems which use Deep Neural Networks (DNN), have been shown to out-perform other approaches when the number of possible L1s is high. For example, previous studies have cited classification rates (CR) of 72% for a 6-way comparison task using DNN [4] and 84% for a 10-way comparison using TDNN [5].

One drawback of purely acoustic systems, and particularly those that utilise deep learning, is that they lack the direct interpretability that segmental approaches provide. Such transparency is essential for use in forensic applications. However, interpretability often means sacrificing performance. For example, [1] provides a recent example of an acoustic-phonetic approach which uses distance measures between segments to determine the L1. The system resulted in a CR of 61% in a 7-way comparison.

Despite the many recent advances in speech technology, language identification based on a speaker's L2 speech remains a difficult and under-researched area. Factors like proficiency, regional variations in the L1, along with other social factors such as attitude towards the target variety, impact L2 realisation and which features are acquired [6-8]. These factors introduce variability within a group, thus making the L2 variety classes difficult to model. Additionally, similar target L2 pronunciations mean between-group differences are expected to be more subtle than between dialects or with language identification based on L1 speech. Whether the L2 varieties should be even be grouped by their L1 and not a finer grained label is a separate issue not discussed in this paper (see [9]).

The present study specifically deals with General American (GA) English as the target L2 of a set of speakers with a wide range of L1s. We aim to explore L2 English phoneme data through the lens of language identification. This presents an initial step towards developing a larger (i.e., based on a wider range of features at different linguistic levels), phonetically interpretable system that can recognise a speaker's L1 based on samples of their L2 English. This approach has forensic applications as a priority.

This paper first explores the distributions of F1, F2, and Duration values for four vowel phonemes, followed by an evaluation of classification rate and probability scores. Finally, we consider whether the phonetic realisations of the different L1 groups are different enough that they can be accurately identified and whether classification performance can be explained in a phonetically interpretable way.

2. METHODS

2.1. Data

The speakers used for this study were 112 male speakers of five varieties of L2 GA English selected from the Speech Accent Archive [10]. The groups are referred to by their L1 as labelled in [10]: Arabic, French, German, Mandarin, Portuguese. Each recording contained approximately 30 seconds of a speaker reading the "Stella" passage. The Montreal Forced Aligner (MFA) [11] was used to segment each sample at the phone level and generate a time aligned Praat TextGrid. The General American acoustic model and grapheme-to-phoneme dictionary were used (english_arpa_us and english_us_ipa_g2p). Therefore, all segment labels are based on the expected phoneme, not the actual realisation. Some alignment error is to be expected but should be sufficient for the purposes of this paper. A study of the performance of the MFA on a subset of this dataset found that 80-93% of boundary placements were within 20 ms and 67-84% were within 10 ms [12]. Taking measurements at the midpoint of the segment mitigates the impact of alignment errors.

2.2. Feature extraction

Using a Praat script [13], F1 and F2 measurements were automatically extracted from a 25 ms midpoint frame for all monophthongal vowel segments contained in the passage, along with duration of the segment. The reference formant values were based on manually corrected values for 2-3 male speakers from each of the varieties.

Four vowel phonemes were selected for this study as a means of focusing on segments that were (i) relatively easy to measure automatically, and (ii) showed high or low variability between varieties based on the manually corrected values. The following four vowels were selected: /ɪ/, /i/, /ɛ/, /ɔ/.

Each feature was z-score normalised, based on the mean across all speakers, to bring the features onto the same scale while maintaining original differences between distributions. This was done to allow for the use of a single smoothing parameter in formula (1) (see below; [14]). Tokens with an absolute z-score above 3 for any of the features were considered measurement errors and removed from the analysis. 97% of the data was retained: 1187 /i/, 893 /ɪ/, 439 /ɛ/, and 556 /ɔ/ tokens.

2.3. Language identification

For within-language comparisons, each language set was randomly split into 80% training and 20% test data using K-fold cross validation (K=5). An adaptation of Aitken and Lucy's [15,16] multivariate kernel density (MVKD) likelihood ratio (LR) formula was used to calculate the probability of a given speaker's data given a model for each language group. This was done via a modified version of Morrison's [17] MATLAB MVKD implementation. MVKD provides a transparent method of comparing

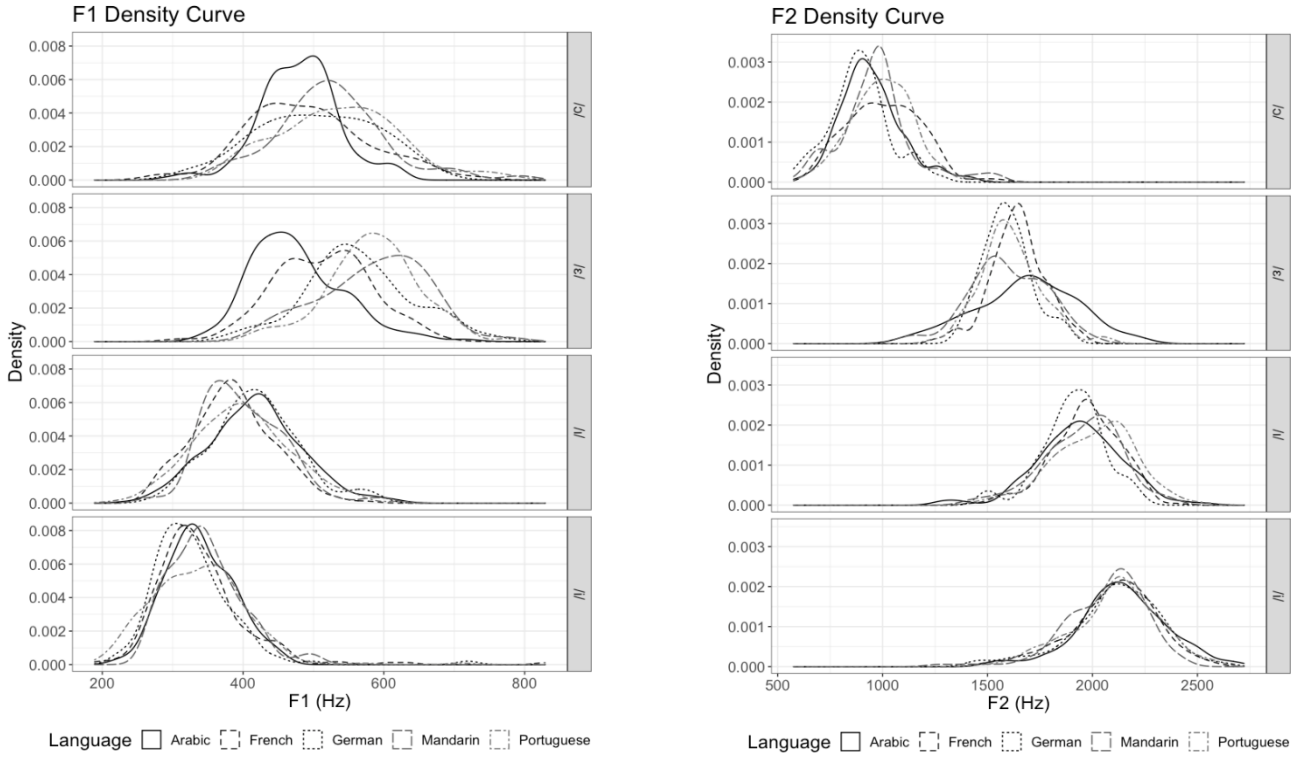

Figure 1: Density distributions for raw F1 and F2 values by vowel and L1.

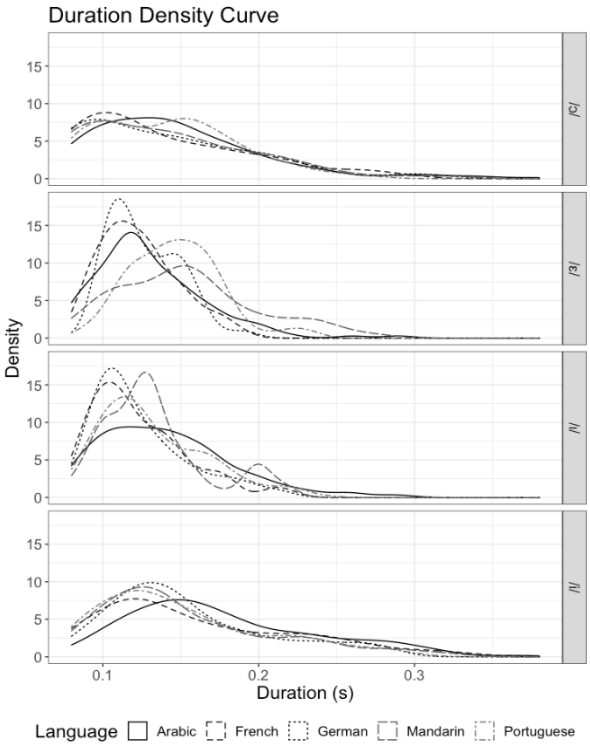

Figure 2: Density distributions for raw Duration values by vowel and L1.

distributions and has been used widely for the purposes of speaker comparison research (e.g., [18,19]). Here, only an adapted version of the denominator of the MVKD LR was used. This involves modelling data for each language group with a multivariate kernel density distribution, made up of equally weighted Gaussians from each speaker in the training data. This also accounts for correlation between features. The mean probability of each feature vector (F1, F2, Duration) from each token for each speaker given each language model is then calculated, as in (1):

$$(1)\ f(z|\mu_1, C, U, H_k) = \int_\theta f(\boldsymbol{z}|\theta_1, U) f(\theta_1|\boldsymbol{\mu}, C) d\theta$$

Where $\boldsymbol{z}$ is the measurements from the speaker, $\boldsymbol{\mu}_1$and C are the within-group mean and covariance, U is the within-speaker covariance and operating under the assumption of hypothesis k (where H_k is the hypothesis that the speaker comes from a given language model).

2.3.1 Evaluation

For each test speaker, probabilities were ranked by language model and then compared against the ground truth to arrive at a conclusion which could be classified in binary terms as correct or incorrect. This was then used to generate an overall classification rate (CR). Additionally, logLRs, which estimate the strength of the evidence, ($\boldsymbol{z}$; i.e., the acoustic measures for a test speaker), were calculated for comparisons between target (H_1) and non-target (H_2) language pairs, as in (2):

$$(2)\ logLR_{1,2} = \frac{P(\boldsymbol{z}|H_1)}{P(\boldsymbol{z}|H_2)}$$

LogLRs are widely used as the output of many automatic speech technologies and are particularly used within the forensic domain. For ease of interpretation, numerical logLRs here are converted to verbal equivalents using the scale in [20].

3. RESULTS

3.1. Feature distributions

Figures 1 and 2 display the distributions of raw F1, F2, and Duration values for the five language groups and four vowels. Larger images and reference formant values are provided as supplementary material [21].

The vowels demonstrate high and low between-language variability (see 2.2). For instance, /i/ shows considerable overlap in the feature space for all selected varieties except Portuguese (for F1) and Arabic (for Duration). However, due to the wider distribution, if a Portuguese speaker's features fall towards the centre of the distribution (around 300Hz along F1 for example), there will always be a higher probability that the speaker belongs to one of the other L1s. This suggests there will be a lot of confusability between the varieties for this segment that will not result in a strong decision in any direction. Conversely, /ɛ/ shows variability across all three features. /ɪ/ displays variability mainly in Duration and F2, while /ɔ/ varies mainly in F1 and F2.

The results predict the performance ranking of the segments, from best to worst, will be: /ɛ/, /ɔ/, /ɪ/, /i/.

3.2. Language identification results

The overall classification rates, where chance is 20%, were: 15% (French), 9% (Arabic), 17% (Mandarin), 11% (Portuguese), and 48% (German). German had a noticeably better classification rate, which is largely attributed to the performance on /ɔ/ and /i/. The performance rankings of each vowel can be found in Table 1. Within a language group, /i/ was most frequently classified correctly, while /ɔ/, /ɪ/ and /ɛ/ tended to be correctly classified less frequently.

Table 2 displays the most common misidentification for each L1 and vowel. In most instances, the probability of the speaker having Mandarin as an L1 was higher than for their correct L1. However, it is worthwhile to note that pairwise logLRs mostly fell between 0 and 2, but more often

between 0 and 1. This indicates a very weak classification, meaning even if the segment was classified incorrectly, the strength of the evidence is limited to moderate on the verbal scale [20].

L1	Vowel			
	1st	2nd	3rd	4th
Arabic	/i/	/ɛ/	/ɔ/	/ɪ/
French	/ɪ/	/ɛ/	/i/	/ɔ/
German	/ɔ/	/i/	/ɪ/	/ɛ/
Mandarin	/i/*	/ɔ/*	/ɪ/	/ɛ/
Portuguese	/i/	/ɪ/	/ɔ/	/ɛ/

Table 1: Vowel phonemes ranked by performance (best to worst) for each L1. * Indicates a tie.

L1	L1 (model)			
	/i/	/ɔ/	/ɪ/	/ɛ/
Arabic (A)	P	G	M	M
French (F)	A/M	G	M	M
German (G)	M	-	M	M
Mandarin (M)	P	G	F	P
Portuguese (P)	M	G	M	M

Table 2: Most common misidentification for each L1 and phoneme label.

Some pairwise comparisons showed greater strength of evidence. For example, speakers with French L1 tended to show logLRs around 3 in support of either hypothesis for /ɛ/. However, higher strength of evidence generally appeared to be speaker-specific rather than patterning with specific L1s.

3. DISCUSSION

The difficulty with classifying L2 varieties is evident from the poor classification rates found in this study. By having the same target L2 English variety, we are minimizing the between-group variability, especially at the phoneme-level.

The performance ranking of the phonemes was not as expected from 3.1. Surprisingly, /ɛ/, which showed the largest differences in Figures 1 and 2 between varieties tended to be the least well-classified. We suspect this is due to where in the feature space a speaker falls. For example, /ɔ/ has very wide distributions for F1 across all varieties. Therefore, despite large differences between the overall distributions, the tails of a given distribution overlap with means of others. Conversely, the vowel with the smallest between-language variation in mean values, /i/, had the best performance for most varieties. The tested varieties all have a phoneme with this label in their native inventory [22-29]. It is possible, however, that /i/ is realised differently across languages, and that such differences are maintained in the L2, in addition to less within-speaker variability. This is reflected in the strength of evidence for pairwise comparisons which often fell into *limited* or *moderate* strength of evidence on the verbal scale [20]. Then again, inclusion of a segment in the speaker's native inventory does not seem to necessarily impact performance. For instance, compare the overall performance of Mandarin and Arabic: 17% and 9% respectively. Of the four vowels in this study, only the phoneme /i/, or one labelled as such, is shared with the native variety, yet they show very different classification rates. There were also a lot more tokens of /i/ than the other phonemes, meaning better representation of the group.

Surprisingly, many of the varieties were confused with Mandarin and German (Table 2). We suspect this has more to do with the relatively smaller number of speakers compared with the other L1s used to form the training set [30]. The German and Mandarin groups contained 12 and 13 speakers respectively, compared with 39 Arabic speakers. Even with only three features, dimensionality could be an issue.

Low strength of evidence was consistent across all comparisons and suggests that while it is possible to correctly identify a speaker's L1 based on their pronunciation of individual vowel phonemes, the classification based on this modelling method is not accurate or consistent enough. However, it does demonstrate issues that will translate to methods using more opaque measurements (such as MFCCs) in higher dimensions, in particular, large within-group variability and substantial overlap between varieties in the phonetic subspace. Also of note is that speakers were not consistent in their performance across segments. This again highlights the challenge of inherent variability of L2 speakers due to differences in factors such as proficiency.

Many other factors could have further affected the results, including alignment error from the MFA, measurement error from Praat formant tracking, and scarcity of data.

4. CONCLUSION

The lack of strong results leads us to conclude that segments in isolation do not provide enough language-specific information to make good predictions. However, some phonemes work better for some languages (and likely some speakers) than others. So, the way forward is to (a) consider a wider range of features at different linguistic levels, and (b) tailor the choice of linguistic features to the language (this may be determined by a combination of knowledge of the languages and through empirical testing). The most effective features for identifying a language, along with the strength of evidence, depend on the pairwise comparison.

5. REFERENCES

[1] Brown, G., Franco-Pedroso, J., & González-Rodríguez, J. 2021. A segmentally informed solution to automatic accent classification and its advantages to forensic applications. *International Journal of Speech, Language and the Law*, 28(2), 201-232.

[2] Kumpf, K., & King, R. W. 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Fifth European Conference on Speech Communication and Technology*.

[3] Bahari, M. H., Saeidi, R., & Van Leeuwen, D. 2013, May). Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7344-7348). IEEE

[4] Upadhyay, R., & Lui, S. 2018, January. Foreign English accent classification using deep belief networks. In *2018 IEEE 12th international conference on semantic computing (ICSC)* (pp. 290-293). IEEE.

[5] Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y. and Xie, L. 2021, June. The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6918-6922). IEEE.

[6] Drummond, R. 2013. The Manchester Polish STRUT: Dialect Acquisition in a Second Language. *Journal of English Linguistics*, 41(1), 65–93. https://doi.org/10.1177/0075424212449172

[7] Gut, U. 2009. *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang.

[8] Rindal, U. 2010 Constructing identity with L2: Pronunciation and attitudes among Norwegian learners of English 1. *Journal of Sociolinguistics* 14, no. 2: 240-261.

[9] Markl, N. 2022, June. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 521-534).

[10] Weinberger, S. 2015. *Speech Accent Archive*. http://accent.gmu.edu/index.php.

[11] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502. doi: 10.21437/Interspeech.2017-1386

[12] Williams, S., Foulkes, P., & Hughes, V. 2023. Analysis of forced aligner performance on L2 English speech. [Manuscript submitted for publication] Language and Linguistic Science, University of York.

[13] Harrison, P. 2022 measureAllFormantsTracked.praat v 1.2 (Praat script)

[14] Silverman, B. W. 2018. *Density estimation for statistics and data analysis*. Routledge.

[15] Aitken, C. G. G., & D. Lucy. 2004. Evaluation of Trace Evidence in the Form of Multivariate Data. *Journal of the Royal Statistical Society. Series C, Applied Statistics* 53 (1): 109–22.

[16] Aitken, C., Taroni, F., and Bozza, S. 2020. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd Ed. Wiley.

[17] Morrison, G. S. 2007. Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. http://geoff-morrison.net/#MVKD

[18] Enzinger, E., & Morrison, G. S. 2017. Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277, 30-40.

[19] Morrison, G. S. 2011. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 532), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

[20] Champod, C. and Evett, I. W. 2000 Commentary on A. P. A. Broeders (1999) Some observations on the use of probability scales in forensic identification. *Forensic Linguistics* 6(2): 228–41.

[21] https://github.com/sejw/Williams_ICPhS_2023

[22] Bahrani, N., & Modarresi Ghavami, G. 2021. Khuzestani Arabic. *Journal of the International Phonetic Association*, 51(2), 299-313. doi:10.1017/S0025100319000203

[23] Barbosa, P., & Albano, E. 2004. Brazilian Portuguese. *Journal of the International Phonetic Association*, 34(2), 227-232. doi:10.1017/S0025100304001756

[24] Cruz-Ferreira, M. 1995. European Portuguese. *Journal of the International Phonetic Association*, 25(2), 90-94. doi:10.1017/S0025100300005223

[25] Fougeron, C., & Smith, C. 1993. French. *Journal of the International Phonetic Association*,23(2), 73-76. doi:10.1017/S0025100300004874

[26] Kessler, B., Leben, W. R., & Lyovin, A. 2017. *An introduction to the languages of the world*. Oxford University Press.

[27] Kohler, K. 1990. German. *Journal of the International Phonetic Association*, 20(1), 48-50. doi:10.1017/S0025100300004084

[28] Thelwall, R., & Sa'Adeddin, M. 1990. Arabic. *Journal of the International Phonetic Association*, 20(2), 37-39. doi:10.1017/S0025100300004266

[29] Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press. doi:10.1017/CBO9780511753459

[30] Hughes, V. 2017. Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough? *Speech Communication*, 94, 15-29.