Anglophonia
French Journal of English Linguistics

30 | 2020
Phonétique et phonologie : représentations et variabilité

Labiodentals /r/ here to stay: Deep learning shows us why

Hannah King and Emmanuel Ferragne

Electronic version
URL: https://journals.openedition.org/anglophonia/3424
DOI: 10.4000/anglophonia.3424
ISSN: 2427-0466

Publisher
Presses universitaires du Midi

Electronic reference
Hannah King and Emmanuel Ferragne, "Labiodentals /r/ here to stay: Deep learning shows us why", *Anglophonia* [Online], 30 | 2020, Online since 20 December 2020, connection on 07 June 2021. URL: http://journals.openedition.org/anglophonia/3424 ; DOI: https://doi.org/10.4000/anglophonia.3424

Labiodentals /r/ here to stay: Deep learning shows us why

Hannah King and Emmanuel Ferragne

AUTHOR'S NOTE

Articulatory data were collected with the generous help from the Clinical Audiology, Speech and Language Research Centre at Queen Margaret University, as well as with the funding provided by the Direction de la recherche, de l'internationalisation, de la valorisation et des études doctorales (DRIVE) at Université Paris Diderot obtained by the first author. We wish to thank Ioana Chitoran for her invaluable input throughout this project.

Introduction

Anglo-English approximant /r/

1 It is well reported that non-lingual labiodental productions (e.g., [ʋ]) of the English pre-vocalic approximant /r/ commonly occur in the variety of English spoken in England, Anglo-English. Along with /l/-vocalisation, /t/-glottaling and TH-fronting, the labiodentalisation of /r/ appears to be part of a general accent levelling process, which typically affects consonants and is spreading across England from its epicentre in the south east of England (Foulkes and Docherty). Indeed, in non-standard south-eastern accents, labiodental /r/ has been established as a relatively widespread feature (Wells; Foulkes and Docherty). However, instances of [ʋ] have been reported all over England including Norwich (Trudgill), Milton Keynes, Reading, Hull (Williams and Kerswill), Derby (Foulkes and Docherty), Leeds (Marsden), Middlesborough (Llamas) and Newcastle (Foulkes and Docherty). Up until the early 2000s, labiodentalisation was regarded as a speech defect or infantilism – due to its presence as a development feature in children acquiring English /r/ (Kerswill; Knight, Villafaña Dalcher, and

Jones) – or as an affectation of upper-class speech (Foulkes and Docherty).[1] However, dialectological evidence suggests that perceptions of labiodental /r/ are changing, particularly in the popular media (Foulkes and Docherty). Indeed, where the labiodental variant was once stigmatised as defective, it is now treated with greater tolerance to such an extent that 'many parents may now be less ready to correct this variant as defective in their children's speech' (Armstrong and Pooley 142).

2 It is generally assumed that labiodental variants have emerged by speakers retaining the secondary labial component of the post-alveolar approximant [ɹ] at the expense of the lingual one (Jones; Foulkes and Docherty; Docherty and Foulkes). The lingual articulation of the post-alveolar approximant has been well-studied, particularly in rhotic Englishes. Post-alveolar /r/ is produced with a multitude of possible tongue shapes, which are categorised on a continuum between two extreme configurations: retroflex, with a raised, curled-up tongue tip and a lowered tongue body; and bunched, with a lowered tongue tip and a raised tongue body (e.g., Delattre and Freeman; Zawadzki and Kuehn; Tiede, Boyce, Holland, et al.). Non-rhotic Englishes (e.g., Anglo-, Australian and New Zealand English) have received comparatively less attention from phoneticians than rhotic ones, but articulatory evidence indicates that pre-vocalic /r/ in non-rhotic English is produced with the same lingual variation as in rhotic varieties, although retroflex tongue shapes appear to be more common in non-rhotic than in rhotic Englishes (Heyne et al.; King and Ferragne 2020).

3 Despite the diversity of possible tongue shapes for post-alveolar /r/, its acoustic profile is paradoxically very stable, at least with regards to the first three formants (Espy-Wilson et al.), which is reflected in perception as American English listeners are reported to be unable to distinguish between the two most extreme tongue shapes (Twist et al.). It is generally agreed that the most salient acoustic feature of the post-alveolar approximant is its low frequency third formant (F3), which is usually below 2 000 Hertz (Hz) (e.g., Delattre and Freeman; Boyce and Espy-Wilson; Proctor et al.) and some researchers have remarked on the close proximity of the third to the second formant (Lisker; O'Connor et al.; Stevens; Guenther et al.). It has also been suggested that the acoustic correlate of the rhotic approximant is not a low-frequency F3 in itself, but rather the dominance of a single perceptual peak in the frequency region of the second formant (Heselwood and Plug). Theoretical articulatory-acoustic models have affiliated the low F3 typical of post-alveolar /r/ with the size of the front cavity, i.e., between the palatal constriction and the lips (Alwan, Narayanan, and Haker; Stevens; Espy-Wilson et al.). These models predict that larger front cavity volumes produce lower F3 frequencies. Increasing the volume of the front cavity – and F3 lowering – may be achieved by backing the palatal constriction; by creating space underneath the tongue tip via increased retroflexion; or by increasing the length of the accompanying lip protrusion channel. Labialisation can thus be considered an articulatory enhancement strategy for post-alveolar /r/ as it acts to extend the front cavity (Smith et al.; King and Ferragne 2020).

4 Although phonetic descriptions are few and far between, non-lingual labiodental pronunciations of /r/ are associated with much higher F3 frequencies than their lingual counterparts. Spectrographic and formant analysis of /r/ in Anglo-English speakers revealed that energy in the higher frequencies beyond F3 is much clearer for labiodental than it is for post-alveolar /r/ (Foulkes and Docherty). While F3 is in close proximity to F2 in lingual post-alveolar productions (at around 1 700 Hz), non-lingual

labiodental variants have a markedly higher F3 at around 2 200 Hz (Foulkes and Docherty). Therefore, acoustically speaking, [ʊ] may actually be closer to [w] than [ɹ] and perceptual confusion between [ʊ] and [w] is indeed widely attested (Foulkes and Docherty; Villafaña Dalcher, Jones, and Knight). Historical evidence of labiodental /r/ shows a tendency for it to be represented orthographically as <w> both in classical literature – including works by Charles Dickens and George Orwell – as well as in more contemporary media such as in television and advertising (Foulkes and Docherty). For example, the English television presenter Jonathan Ross' notable use of labiodental /r/ has awarded him the affectionate nickname 'Wossy', which is also his Twitter handle. Anecdotally, Ross tweeted about a funny encounter he had with Apple's voice-prompted service Siri, which misperceived his labiodental /r/ as /w/. When Ross set a reminder in his diary to call BBC Radio 2, he was presented with a note telling him to call 'Wadey 02'! It would seem that perceptual confusion between [ʊ] and [w] occurs not only in human listeners but in machines too.

5 Contrary to the lingual articulation, phonetic accounts of the labial articulation which accompanies the post-alveolar approximant are markedly absent from the literature. Indeed, as justly observed, 'if its labial component is mentioned at all, it is only *en passant*' (Foulkes and Docherty 182 emphasis original). In actual fact, detailed articulatory accounts of the implementation of labialisation in consonants are surprisingly hard to come by more generally. This is perhaps because the articulatory dimensions used for labialisation in consonants are the same as those used for rounding in vowels (e.g., Marchal). Indeed, Laver explains that the label 'labialisation' has been used so extensively that the only articulation to which the various usages refer is likely a 'horizontal compression of the interlabial space' (Laver 38), which he indicates is the articulatory property that rounded vowels also share. However, in other accounts of lip rounding in vowels, this 'horizontal compression' is markedly absent. In one of the earliest phonetic accounts of rounding in vowels, two types of rounding are distinguished: 'inner' and 'outer' (Sweet). According to Sweet's descriptions, inner rounding, which is typical of back vowels, involves a lateral compression of the lip corners (i.e., with horizontal compression), while outer rounding, typical of front vowels, involves vertical lip compression (i.e., without horizontal compression) (Mayr). More modern accounts of lip rounding have been proposed, although they have all taken inspiration from Sweet's original descriptions of inner and outer rounding by distinguishing between two main types of rounding, associated with back and front vowels respectively (e.g., 'horizontal' and 'vertical' rounding - Heffner; 'endolabial' and 'exolabial' rounding - Catford). Although we might assume that the different lip rounding postures have a similar acoustic effect by lengthening the front cavity, front vowels tend to be rounded without horizontal compression, which may prevent over-lowering F2 and preserve their front quality (Catford). Similarly, it has been suggested that by not being produced with the close lip rounding typically associated with back vowels, the vowel [y] can contrast with [i] as it results in F3 coming in close proximity to F2, whereas for [i], F3 is high and close to F4 (Wood).

6 As for existing articulatory descriptions of labialisation in post-alveolar /r/, most accounts simply state that pre-vocalic /r/ may involve lip rounding in both American English (e.g., Delattre and Freeman; Mielke, Baker, and Archangeli; Proctor et al.) and Anglo-English (e.g., Abercrombie; Jones; Scobbie). In rhotic Englishes (e.g., American and Scottish English), pre-vocalic /r/ presents lower formant values than post-vocalic /

r/, which is generally assumed to be the result of lip rounding pre-vocalically (Delattre and Freeman; Lehiste; Zawadzki and Kuehn). In non-rhotic Anglo-English, it has been observed that between 25 and 50 percent of non-broadcasters interviewed on the radio and television in the UK labialise their /r/ at least some of the time (Scobbie). Some argue that variability in the degree of lip rounding in pre-vocalic English /r/ is largely determined by co-articulation with the following vowel, with /r/ preceding rounded vowels displaying more rounding than /r/ preceding non-rounded vowels (Gimson). However, by examining the articulation of Anglo-English word-initial /r/ from ultrasound tongue images and profile video camera images, we concluded that co-articulation with the following vowel cannot entirely account for the degree of labialisation for /r/, at least with regards to lip protrusion (King and Ferragne 2020). No significant difference in lip protrusion was observed between occurrences of /r/ in the context of the (unrounded) FLEECE vowel and the (rounded) GOOSE, THOUGHT, and LOT vowels. However, a significant correlation was observed between the degree of lip protrusion and the size of the buccal cavity in front of the palatal constriction: constrictions resulting in small front cavities were accompanied by more lip protrusion than those with larger cavities. The size of the front buccal cavity appears to be dictated by tongue shape – tip-up tongue shapes with their substantial sublingual space induce larger cavities than tip-down ones (Zhang et al.) – and tongue position – the palatal constriction for /r/ before front vowels is more anterior than those before back vowels due to co-articulation (King and Ferragne 2020). We thus proposed a trading relationship between the size of the front buccal cavity and the size of the lip protrusion channel for English /r/, which allows speakers to maintain a relatively stable acoustic output for /r/ in different articulatory contexts, particularly with regards to F3 (King and Ferragne 2020). Although we only examined productions of /r/ in Anglo-English, we see no reason why this trading relationship might not exist in other varieties of English, and a positive correlation between tip-down tongue shapes and lip protrusion has indeed been observed in American English (Tiede, Boyce, Espy-Wilson, et al.).

7 English pronunciation manuals touch on the labialisation of /r/ but vary in the advice they give to second-language learners. O'Connor recommends learners approach [ɹ] from [w], and then curl the tip of the tongue back until it is pointing towards the hard palate, which implies that the lip postures for [ɹ] and [w] are identical. However, other manuals explicitly warn learners not to exaggerate lip rounding for /r/ because it would produce the percept of a [w] (e.g., Lilly and Viel; Roach). Others even go as far as to inform learners that using their lips to help them form the /r/ sound is 'wrong' and recommend that they use their fingers to hold their lips still in order to force them to concentrate on only using the tongue to produce the sound (Ashton and Shepherd 49).[2]

8 The most detailed articulatory description of the lip posture for /r/ is arguably the impressionistic account provided by Brown, who suggests that the lip postures for /r/ and /w/ are not the same in Standard Southern British English. She observes that the lip corners are compressed horizontally but not necessarily pushed forwards for /w/, while for /ʃ, ʒ, tʃ, dʒ/ and /r/, the lips are opened and pushed forwards, showing their soft inner surfaces. Her account of lip rounding for /w/ echoes Sweet's description of inner rounding in back vowels, while her description of lip rounding for /r/ parallels the outer rounding observed in front vowels. Although not discussed by Brown, from an acoustic standpoint, these observations make sense. Unlike the post-alveolar approximant /r/, the labial-velar approximant /w/ is categorised as having a high F3

and a very low F2, resulting in a large gap between the two formants (Espy-Wilson; Stevens). Acoustic modelling indicates that in the case of a backed tongue constriction, such as the one produced for /w/ and for its vocalic counterpart /u/, the condition of minimum F2 is only achieved with close lip rounding, forming a narrow lip opening (Fant; Stevens). Close lip rounding for post-alveolar /r/ could result in over-lowering of F2, which would potentially cause perceptual confusion with /w/. Consequently, as with front rounded vowels, rounding without a horizontal contraction of the inter-labial space (i.e., with outer rounding) may allow F2 to remain in close proximity to F3 for /r/, thus producing a maximal auditory contrast with /w/ (as discussed in King and Ferragne 2020).

Aims and predictions

9 It has been proposed that non-lingual labiodental variants have emerged in England by speakers retaining the labial articulation of /r/ at the expense of the lingual one (Jones; Foulkes and Docherty; Docherty and Foulkes). A key assumption underlying this hypothesis is that Anglo-English /r/ is produced with a labiodental articulation even when it is accompanied by a post-alveolar lingual gesture. This proposal seems contrary to the general agreement held by phoneticians that English /r/ may be produced with 'lip rounding'. However, we do not yet know how this so-called lip rounding is implemented for /r/. A detailed articulatory investigation of the labial posture in Anglo-English speakers who still present a lingual gesture for /r/ may allow us to provide a phonetic account as to how and why non-lingual labiodental variants are becoming increasingly widespread in England. In this paper, we assess whether lingual productions of the approximant /r/ are accompanied by a labiodental-like lip posture by analysing lip camera data of /r/ and /w/ from Anglo-English speakers who still produce lingual /r/. Given that /w/ is the semi-vocalic counterpart for the back vowel /u/, it is generally agreed that /w/ is produced with the same lip rounding posture as /u/, i.e., with inner rounding (e.g., Catford; Marchal). For inner rounding, the lip corners are brought together laterally away from the teeth. This lip configuration seems entirely incompatible with a labiodental-like posture, in which the lower lip moves towards the top teeth (Ladefoged and Maddieson). If post-alveolar /r/ is produced with a labiodental-like posture, we thus predict that the inner rounded labial gesture for /w/ will differ considerably.

Procedure

10 We present video camera data from 23 Anglo-English speakers (21F, 2M) who produced one repetition of 18 monosyllabic minimal pair words contrasting /r/ and /w/ word-initially (see Appendix A for full word list). The speakers, aged between 18 and 55 (mean = 30.34 ±11.27), come from all over England (south west: *n*=1; south east: *n*=6; midlands: *n*=3; north west: *n*=7, north east: *n*=6) and were recorded at Queen Margaret University, Edinburgh, where ethical approval had been obtained. The participants self-identified as speaking with an English accent and the first author, who is a native Anglo-English speaker, verified that this was true by conversing with them. Before participating, the speakers signed an informed consent form and completed a background questionnaire. They were financially compensated £20 for their participation. An NTSC micro-camera was fixed in place relative to the speaker's head,

capturing front-view colour images of the bottom half of the face at a rate of circa 60 frames per second. The data come from an existing study in which we examined the articulation and acoustics of /r/ using synchronised ultrasound tongue imaging, front and profile lip cameras, as well as the auditory signal (King and Ferragne 2020). All 23 speakers had an observable lingual gesture for /r/, which was visually classified from ultrasound tongue images on a continuum from tip-down bunched to curled-back retroflex (see King and Ferragne 2020 for more details). As the dataset contains limited data from male subjects (*n*=2) and as it is well established that speaker sex influences formant values, acoustic analysis was performed on the remaining 21 female speakers. In women, F3 was around 800 Hz lower and F2 500 Hz higher for productions of /r/ than for /w/ on average, thus confirming a clear phonetic difference between the productions of /r/ and /w/. For the present study, the image corresponding to maximal labial constriction was manually selected from 414 lip videos of word-initial /r/ and /w/ by visually examining sequential video frames.

A deep learning-based approach

11 If post-alveolar /r/ is inherently more labiodental than /w/, their lip postures should be recognisably different. In our previous study, the lip postures for /r/ and /w/ were measured by hand, which was both time consuming and prone to human error (King and Ferragne 2020). A logical alternative would be to measure the lips automatically. Given the visual nature of the dataset and the great success deep neural networks (DNNs) have enjoyed in recent years in the field of image recognition (Simonyan and Zisserman), it seemed like a good opportunity to apply deep learning-based methods to answer phonetic questions. The most common class of DNNs applied to image classification and recognition is Convolutional Neural Networks (CNNs). The technical details concerning the inner workings of CNN architectures go far beyond the scope of this paper (see Goodfellow, Bengio, and Courville for an introduction)[3] However, for image recognition, the idea behind them is relatively straightforward. To put it very simply, the aim is to replicate the basic human skill of recognising and classifying objects within an image. For example, if a person is presented with an image of a cat, their brain will automatically recognise the object and classify it as 'cat'. The brain is also able to distinguish a cat from another object or animal, such as a dog. By providing the computer with lots of images of cats and dogs, it should be able to learn the attributes that distinguish the two animals within an image. So, when the computer is presented with a new image of a cat, it should be able to tell with a certain degree of certainty that there is a cat, and not a dog, in the image. The 'convolutions' in CNNs filter the images pixel by pixel. The pixels that are important for a cat to be classified as a cat are 'enhanced' by the model, whereas non relevant pixels for the cat class receive negligible weight.

12 We aimed to use a CNN to automatically classify /r/ and /w/ tokens from our 414 front camera static lip images. In a way, we may consider the results from this CNN analysis as an alternative to inferential statistics (Ferragne). If the CNN is able to classify /r/ and /w/ with a high level of accuracy, we may conclude that /r/ and /w/ present sufficiently discriminant features which allow the programme to distinguish between them. As far as we are aware, we are the first phoneticians to analyse the lips using techniques from deep learning. As a result, this investigation is highly exploratory in

nature. For ease of presentation, we will present our analyses and results together in one combined section.

Analyses and results

Automatic classification of /r/ versus /w/ with deep learning

13 The image corresponding to maximum labial constriction was manually located and extracted from the 414 lip videos in our dataset by visually examining sequential video frames, resulting in 207 colour images of /r/ and of /w/. The well-known CNN architecture ResNet-18 (He et al.) was trained to automatically learn the difference between /r/ and /w/ from these 414 still images of the lips with the input resized to match the size of the images in our dataset using *MATLAB Deep Learning Toolbox*. To validate the model, 10-fold cross validation was applied. In this type of model validation, the dataset is randomly split into 10 equal subsets, 9 of which are used for training the model, while the remaining subset is set aside for testing. This process is repeated 10 times with each of the 10 subsets used for testing and the results are averaged to produce one single model estimation. With 10-fold cross validation, the CNN achieved 99.52% mean correct classification of /r/ and /w/ tokens with a standard deviation of 1.02%. An initial interpretation of this extremely high model accuracy is that the front lip images for /r/ and /w/ must differ. However, although 10-fold cross validation is methodologically valid, given that the dataset split is random, data from all speakers is present in both the training and the test sets. This means that the models may have relied on speaker-specific information to distinguish between /r/ and /w/, rather than differences occurring across speakers.

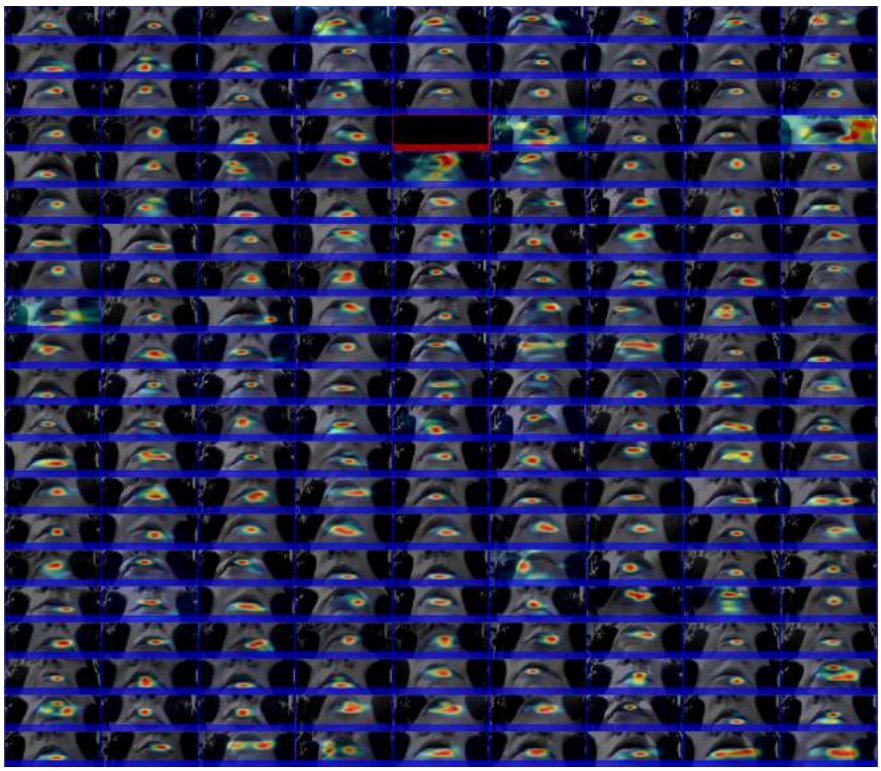
14 To challenge the generalisation ability of the model, we supplemented our validation technique by testing a leave-one-out validation procedure. In this type of model validation, a speaker's whole dataset is left out for the model testing stage. Training is therefore carried out with data from the remaining 22 participants. With this more demanding procedure, mean correct classification was 92.27% with a standard deviation of 14.86%. Model accuracy varied from one speaker to the next, ranging from 50 to 100%, which is reflected in the relatively high standard deviation. Visualising the data revealed that the camera angle obscured the top lip in two speakers who obtained some of the lowest accuracy scores, which may explain why the model struggled to classify these speakers' /r/ and /w/ productions (see Appendix B for an example). Although model accuracy was slightly lower and more variable with this leave-one-out validation procedure than with the previous 10-fold cross validation procedure, a classification rate of over 90% on average still indicates that the lip postures for /r/ and /w/ tokens differ, as visual inspection also indicates.

Deciphering the models

15 Both of our validation techniques seemed to confirm that a CNN is able to recognise a /r/ token from a /w/ token from front-facing images of the bottom half of Anglo-English speakers' faces with extremely high accuracy, which follows our prediction that the labial postures for /r/ and /w/ are different. However, how do we know that the CNNs based their decisions on linguistically relevant information, i.e., the configuration of the lips? It is frequently remarked that one of the shortcomings of deep learning is the

difficulty in understanding what exactly DNNs learn from the data. Indeed, DNNs are often described as 'black boxes' due to the opaqueness of their inner mechanisms (Ferragne, Gendrot, and Pellegrini). Luckily for us, solving this problem has been the focus of many researchers in the deep learning community and as a result, effective methods of visualising what DNNs learn now exist (Ferragne). One such technique is 'occlusion sensitivity' (Zeiler and Fergus), whereby a mask is placed to cover a small area of each image and the resulting drop in the probability that the image will be correctly classified is recorded. The mask position is then changed slightly and the probability drop of the new mask position is computed until the mask has occluded all possible positions in the image. The default settings of the occlusionSensitivity function in *MATLAB Deep Learning Toolbox* were used to implement an occlusion analysis on our models. Mask size was 60 × 160 pixels (height × width) and step size (aka 'stride') was 30 × 80 pixels. To visualise the results, each image was overlaid with a heatmap showing the areas on which the models based their decisions. Red regions in the heatmaps highlight the most relevant areas for the classification, while regions in blue (or those with no overlaid colour) show parts of the image whose influence on the classification is small to negligible. Example heatmaps are presented in Figure 1 for the first model in which the 10-fold cross validation procedure was employed. Visualising the heatmaps indicated that much more often than not, it is the lips that are highlighted. Occlusion analysis was also performed on the model with the more demanding leave-one-out validation procedure and for the speakers whose model accuracy score was high, the salient regions of interest for the model were again the lips. We can thus conclude with a reasonable degree of certainty that the lip configurations for /r/ and /w/ have sufficiently discriminant features which allow the programme to distinguish between them.

Figure 1 Example heatmaps from occlusion analysis of a CNN trained to automatically classify /r/ and /w/ from 414 front lip images using 10-fold cross validation. The image with a red frame shows the only misclassified item in this batch.



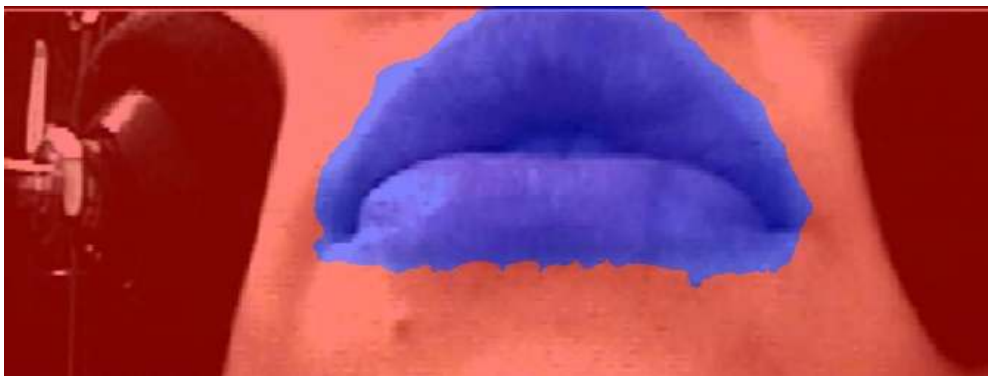
Segmenting and measuring the lips with deep learning

16 The previous analysis indicated that a CNN can tell a /r/ from a /w/ token simply by 'looking' at static images of the speaker's lips. However, we do not yet know how exactly the lip postures differ in articulatory terms. Efforts were made to find a technique capable of segmenting the mouth from the rest of the image automatically. Attempts to use traditional colour segmentation techniques proved unsuccessful due to poor image quality. Indeed, video acquisition was done in rather adverse conditions. Camera angle could not be controlled due to limitations caused by video camera stabilisation and the lighting conditions were not optimised. We again looked to deep learning for a potential solution. A technique called 'semantic segmentation'[4] was applied to teach a CNN to detect the lip area using *MATLAB Computer Vision Toolbox* and *MATLAB Deep Learning Toolbox*. 100 of the 414 images in the dataset were randomly selected and the mouth was manually segmented. Manual segmentation for semantic segmentation involves labelling the pixels within an image which correspond to a particular object or class. In our data, we had two classes: the mouth and everything else (the background, henceforth). A DeepLab v3+ (Chen et al.) based on ResNet-18 (He et al.) was trained using 60 of the 100 segmented images with their corresponding pixel labels (i.e., mouth and background). The remaining 40 images were used to test the model on what it had learnt. For each image, the CNN selects the pixels it has learnt to associate with the mouth, which are then compared with the pixel values obtained from manual segmentation. The model's performance may thus be evaluated both at a global and at a class level.

17 Global segmentation accuracy[5] with semantic segmentation of the lip area was very high at 94.29% suggesting that the CNN performed very well. However, the resulting performance metrics also indicated that the model performed less well at detecting the

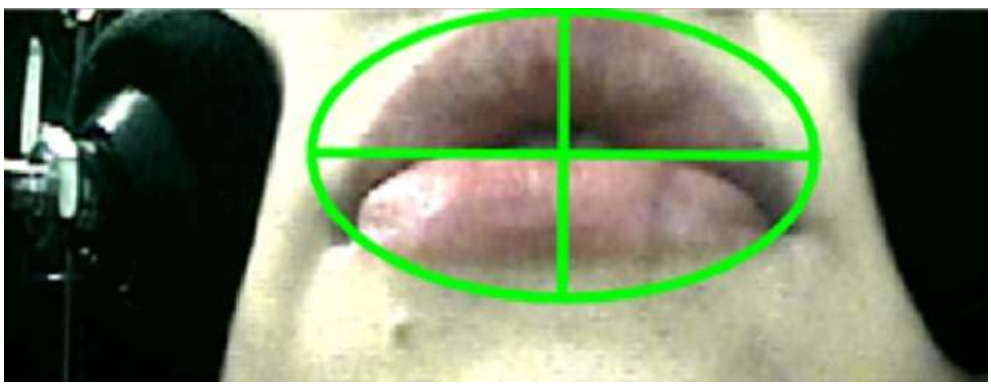
boundary between the mouth and the background because accuracy was only 56.23% on average according to the mean BF score[6]. These results suggest that globally, the model was able to segment the mouth from the background but was less successful at detecting the boundary between them i.e., the lip contour. An example image of the resulting automatic segmentation is presented in Figure 2. Automatic segmentation of the mouth is presented in blue. As in the figure, despite the high global accuracy score, automatic segmentation of the mouth does present stray pixels, although the CNN was generally able to localise the mouth quite well.

Figure 2 Automatic segmentation of the mouth (in blue) via semantic segmentation using a CNN.



18 Given the high global accuracy achieved by the CNN at nearly 95%, the resulting model was used to automatically detect the mouth in all 414 front view images. In order to prevent bias caused by stray pixels, an ellipse was automatically fitted to the region identified as the mouth in each image,[7] an example of which is presented in Figure 3. The ellipse then allowed us to compute four measurements of the lips (in pixels). These measures were based on the length of the horizontal and vertical axes and the position of the ellipse centroid (i.e., where both axes meet). The four measures and their corresponding lip dimensions are presented in Table 1.

Figure 3 Ellipse fitted to the automatically segmented mouth, which is used to compute mouth width, height and centroid.



19 Table 1 Ellipse measures and their corresponding dimensions resulting from automatic semantic segmentation of the mouth using a CNN.

Ellipse measure (in pixels)	Dimension
length of horizontal axis	lip width

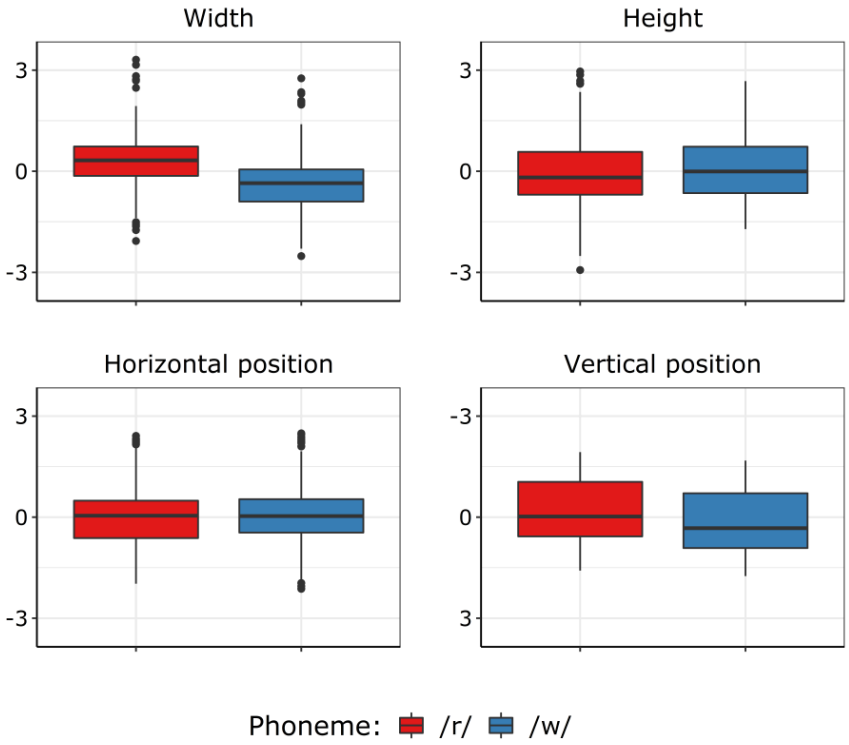
length vertical axis	lip height
position along x-axis of centroid	horizontal lip position
position along y-axis of centroid	vertical lip position

20 Table 2 presents descriptive statistics for the four lip dimension measures (in pixels) acquired via semantic segmentation of the mouth using a CNN. These measures suggest that /w/ tokens have a smaller lip width on average than /r/ tokens, while lip height does not seem to greatly differ between the two. With regards to the horizontal position of the lips, there is unsurprisingly very little difference between /r/ and /w/ tokens. We had no reason to believe that labialisation would result in the lips being positioned more to one side than the other, so this result was expected. However, the mean values for vertical position suggest that productions of /r/ and /w/ vary along this dimension. It is important to specify here that lower values in vertical position correspond to pixels located closer to the top of the image pane, i.e., a higher lip position. /r/ tokens have a lower vertical lip measure than /w/ ones on average, suggesting that the lips are higher for productions of /r/ than they are for /w/. The box plots in Figure 4, in which the lip dimensions were converted to z-scores, paint a similar picture. The interquartile range (presented in boxes) for /r/ and /w/ clearly overlap for lip height and horizontal position, suggesting that the labial articulation of the two phonemes does not greatly differ across these two dimensions. The main difference for /r/ and /w/ seems to involve the width and the vertical position of the lips. The y-axis has been reversed in the box plots for vertical position to reflect the fact that lower values correspond to a higher lip position. In other words, a higher position in the graph indicates a higher lip position. According to the graph, /r/ seems to be articulated with a higher lip position than /w/, which may be suggestive of labiodentalisation: the bottom lip moves up towards the upper front teeth.

21 Table 2 Mean and standard deviation (in parentheses) lip dimensions (in pixels) of /r/ and /w/ from automatic semantic segmentation using a CNN.

Phoneme	Width	Height	Horizontal position	Vertical position
/r/	289.53 (34.47)	166.50 (23.60)	375.40 (29.44)	138.74 (31.97)
/w/	316.42 (33.29)	163.96 (26.29)	374.26 (28.65)	129.21 (32.95)

Figure 4 Box plots showing the interquartile range (box), median (horizontal black line), extreme values (whiskers) and outliers (dots) for lip dimensions converted to z-scores for /r/ and /w/ acquired from semantic segmentation of the mouth using a CNN. The y-axis has been reversed for the vertical lip position measure (bottom right) to reflect the fact that lower values correspond to a higher lip position.



22 To analyse whether the lip dimensions for /r/ and /w/ significantly differ, a generalised linear mixed effects model was performed in R (R Core Team) using the lmer() function of the *lme4* package (Bates et al.) to predict the probability that a token is a /w/ based on the four automatic lip measures: width, height, vertical position and horizontal position. All four measures were z-scored to improve model fit and to allow us to measure the relative impact of each on predicting a token is a /w/. The maximal set of successfully converging random intercepts for subjects and for the following vowel were included, which turned out to be random intercepts for subjects. The addition of random intercepts for the following vowel resulted in a singular-fit. Model residuals were plotted to test for deviations from homoscedasticity or normality and assumptions were met. To test the significance of main effects to model fit, likelihood ratio tests, which were implemented with the mixed() function of the *afex* package (Singmann et al.), revealed that horizontal lip position was the only effect which failed to reach significance (Horizontal Position: $\chi^2(1) = 2.60$, $p = 0.11$). The three other dimensions were highly significant (Width: $\chi^2(1) = 159.93$, $p < .001$; Height: $\chi^2(1) = 25.75$, $p < .001$; Vertical Position: $\chi^2(1) = 80.06$, $p < .001$). The *lmerTest* library (Kuznetsova, Brockhoff, and Christensen) was used to calculate indications of significance within the model, which uses values derived from Sattherthwaite's approximations for degrees of freedom. The resulting *p*-values are presented in the model output in Table 3. The table also presents model estimates which indicate that for an average speaker, the log-odds of observing a /w/ token are 4.45 higher when lip width decreases, 1.68 higher when lip height increases, and 7.27 higher when the vertical lip measure increases. We stress again that a positive vertical lip measure corresponds to a *lower* lip position. In other words, the model predicts lip width to be smaller, lip height to be larger and vertical lip position to be lower for /w/ than for /r/. Although these three dimensions are statistically significant, a comparison of their *t* values indicates that width and vertical position are the strongest predictors of phoneme category. Indeed, a comparison of

each speaker's mean lip dimensions for /r/ and /w/ revealed that the most robust difference between /r/ and /w/ occurs in these two dimensions. On average, lip width is wider and vertical lip position is higher for /r/ than for /w/ tokens in 22 of the 23 speakers. Differences in lip height are less robust in that only 13 of the speakers have a larger lip height for /w/ than for /r/ on average.

23 Table 3 Output of a generalised linear mixed-effects model predicting the probability a token is a /w/ according to the lip dimensions acquired from semantic segmentation using a CNN. Main effects were z-scored. A positive estimate for the effect of Vertical Position corresponds to a lower lip position. R syntax used to build the model is presented directly below the table.

Predictor	Estimate (log-odds)	Std. Error	*t* value	*p* value
(Intercept)	-0.08	1.67	-0.05	0.97
Width	-4.45	0.51	-8.74	< .001***
Height	1.68	0.36	4.62	< .001***
Horizontal Position	1.18	0.75	1.58	0.12
Vertical Position	7.27	1.09	6.66	< .001***
Phoneme ~ Width_z + Height_z + HorizontalPosition_z + VerticalPosition_z + (1|*Speaker*)				

Summary of results

24 By using automatic methods from deep learning, we have shown that the lip postures which accompany lingual /r/ and /w/ differ in Anglo-English speakers. A deep convolutional neural network used static front view images from 23 speakers at the point of maximum labial constriction to automatically learn the difference between /r/ and /w/. The very high accuracy of the model (near 100% for most subjects) supported the sufficiently discriminant role of lip configuration; and occlusion analysis confirmed that the model relied on the lips to categorise each image as /r/ or /w/ token. Another deep neural network was trained to automatically segment the mouth from the rest of the images. This technique allowed us to obtain consistent measurements of the lip dimensions and provided us with a more detailed understanding of the labial postures for Anglo-English productions of /r/ and /w/ in articulatory terms. Statistical analysis of the resulting lip dimensions revealed that the most robust indicators of phoneme class involved the width and the vertical position of the lips: the lips are wider and higher for /r/ than they are for /w/ productions.

Discussion

A phonetic account of labiodentalisation

25 The results from this study indicate that in Anglo-English speakers who still produce a post-alveolar lingual constriction for /r/, the accompanying labial posture is different from that of /w/. Measurements of the lips acquired using techniques from deep learning indicate that at the point of maximum constriction, the lips are less wide and are positioned higher up for /r/ than they are for /w/ tokens. If we return to the lip rounding distinctions described in Section 1.1, notably concerning inner versus outer rounding, we propose that while /w/ is produced with an articulation which resembles the former, /r/ is likely produced with the latter. Inner rounding is typically produced with a horizontal contraction of the lip corners and the smaller lip width observed for /w/ seems to correspond to such a posture. In contrast, the higher vertical lip position for /r/ suggests that the lips are protruded upwards without horizontal contraction, i.e., with outer rounding. Incidentally, an upward movement of the lips may result in labiodentalisation: the bottom lip is raised resulting in an approximation of the inner surface of the bottom lip with the front surface of the upper incisors. As a result, the vertical lip position measure presented in this paper seems to indicate that /r/ is accompanied by a labiodental-like lip posture. Our investigation has therefore provided phonetic evidence to support the supposition in the literature that the post-alveolar approximant /r/ is generally produced with a labiodental-like lip posture in Anglo-English. We may therefore account for the change in process to labiodental /r/ (e.g., [ʋ]) with the loss of the lingual component of the articulation of /r/, leaving the remaining labial one behind to form the primary constriction (as proposed by Jones; Foulkes and Docherty; Docherty and Foulkes).

26 The question remains as to why the labial postures for /r/ and /w/ might vary. It has been proposed that the lip rounding in front and back vowels differs in order to enhance the perceptual contrast between them: front vowels are produced with less lip corner contraction to avoid over-lowering F2 (Catford; Wood). It may be that different lip rounding strategies are employed to enhance the acoustic contrast between /r/ and /w/. By using outer rounding, speakers who produce /r/ with an observable tongue body gesture may enhance the lowering of F3 by lengthening the cavity in front of the palatal constriction, all the while maintaining a small distance between F3 and F2 by avoiding the close lip rounding associated with inner rounding. An alternative explanation for the observed difference in labial configurations between /r/ and /w/ in Anglo-English could be that by using distinctive articulatory cues, speakers are able to enhance the visual contrast between the two phonemes. Indeed, speech has been shown to be visually optimised in cases where pressure to maintain a phonological contrast is high. For example, the visual lip rounding cue has been found to enhance the perception of the /ɑ/-/ɔ/ contrast, which is currently undergoing a merger in some accents of American English (Havenhill and Do). Similarly, it has been found that Swedish listeners heavily rely on visual cues in the perception of the /i/-/y/ contrast in Swedish (Traunmüller and Öhrström). With regards to /r/, it has been suggested that the loss of the lingual gesture in Anglo-English /r/ may be due to the heavy visual prominence of the lips (Docherty and Foulkes). As such, our future research will investigate to what extent the visual cue of the lips may influence the perception of the /r/-/w/ contrast. We have shown in this study that a CNN can distinguish between /r/

and /w/ with very high levels of accuracy just by 'looking at' images of the lips. Although the human brain has long served as a source of inspiration for machine learning and the best algorithms today for learning structure in data are artificial neural networks (Fong, Scheirer, and Cox), we do not yet know if the difference in the lip postures for /r/ and /w/ that is recognisable in machines is perceptually salient to human perceivers.

Methodological implications

27 On a methodological level, we have used techniques from deep learning to not only train models to learn articulatory differences from raw lip camera images, but also to automatically segment and measure the lips. The fact that convolutional neural networks learn their own representations from the data is a promising research avenue for future phonetic studies. We see no reason why analyses with CNNs may not be extended to any relatively large image-based dataset, such as those containing spectrograms, fundamental frequency curves or ultrasound tongue imaging to name a few. This study has shown that it may be possible to partly overcome the 'black box' problem and make what DNNs learn from the data more explicit by using occlusion analysis. Incidentally, we came to the same conclusion in a previous study which used other equivalent techniques (Class Activation Maps - King and Ferragne 2019). We have illustrated how the visualisation of heatmaps not only makes neural networks' decisions more interpretable, but can also draw researchers' attention to potential biases in their studies. We were able to show that the models drew on articulatory plausible information within the images (i.e., the mouth) to classify tokens as /r/ or /w/. Had this not been the case, the high accuracy obtained by the model would have been very hard to interpret.

28 Semantic segmentation using a CNN was able to accurately segment the mouth from the rest of the images and provided us with measurements of lip dimensions and position, despite the quality of the lip images being rather poor. This approach was less time consuming and is more reproducible than taking measurements of the lips by hand. Although we have presented results from static data, a logical extension will be to train models with whole lip videos rather than selected frames, which we are currently working on implementing.

Conclusion

29 In this study, we have used techniques from deep learning to help us provide a phonetic account as to why increased labiodental variants of /r/ may have emerged in Anglo-English. It has been suggested that labiodentalisation of English /r/ may be due to speakers retaining the labial gesture of post-alveolar /r/ at the expense of the lingual one (Jones; Foulkes and Docherty; Docherty and Foulkes), implying that Anglo-English /r/ is always labiodental even in lingual productions. We verified this assumption by comparing the labial postures of /r/ and /w/ in Anglo-English speakers who still present a lingual component for /r/. We hypothesised that if post-alveolar /r/ is labiodental, the labial gesture for /w/, which is unequivocally considered rounded, should differ substantially. A deep convolutional neural network automatically learnt the difference between /r/ and /w/ from still images of the bottom half of the face in

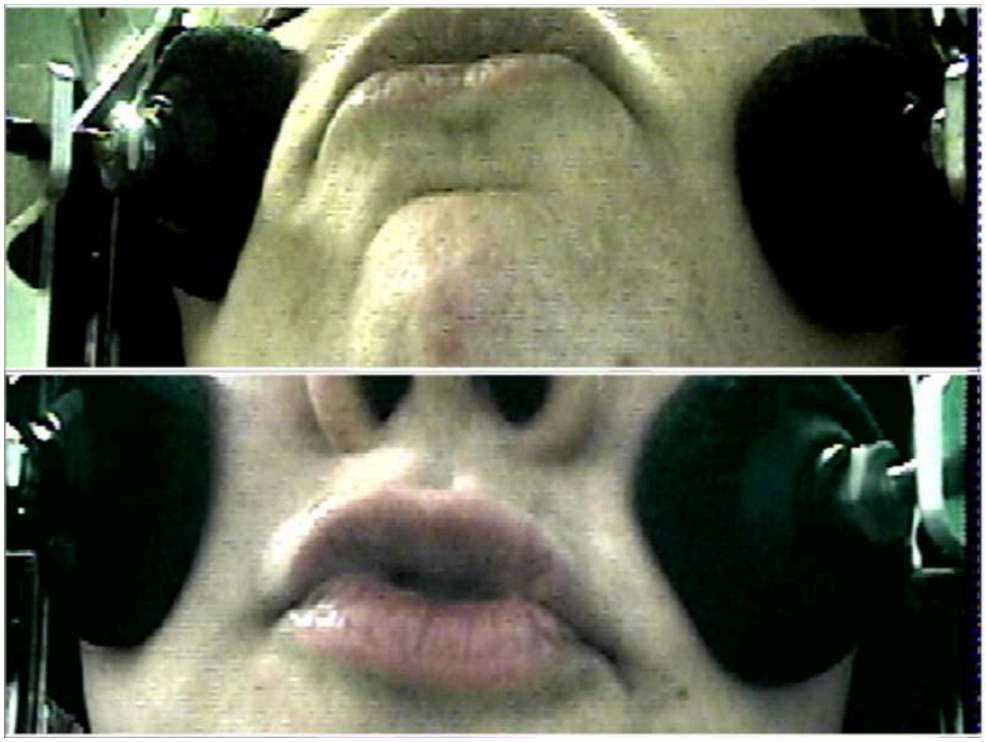
23 speakers. The very high accuracy of the model (near 100% for most subjects) supported the sufficiently discriminant role of lip configuration; and occlusion analysis confirmed that the model relied on the lips. In order to get a more detailed understanding in articulatory terms, another deep neural network was trained to automatically segment the mouth from the rest of the images. Semantic segmentation allowed us to obtain consistent measurements of lip dimensions and position. To the best of our knowledge, this is the first time such techniques have been used for the analysis of articulatory phonetic data. Our results indicated that the lip postures differ significantly for /r/ and /w/. The lip corners are brought together at the centre for /w/, whereas for /r/, the lips are protruded upwards, presumably resulting in the bottom lip approaching the upper teeth. We thus conclude that a labiodental-like lip posture accompanies post-alveolar approximant articulations of /r/ in Anglo-English and that labiodentalisation may be due to the loss of the lingual articulation.

Appendices

30 Appendix A Test words including minimal pairs contrasting /r/ and /w/ word-initially. Phonological transcriptions are for Standard Southern British English.

/r/-initial		/w/-initial	
/riːd/	‘reed’	/wiːd/	‘weed’
/riːp/	‘reap’	/wiːp/	‘weep’
/red/	‘red’	/wed/	‘wed’
/ruːm/	‘room’	/wuːm/	‘womb’
/rɪŋ/	‘ring’	/wɪŋ/	‘wing’
/ræk/	‘rack’	/wæk/	‘whack’
/rʌn/	‘run’	/wʌn/	‘won’
/rɔː/	‘raw’	/wɔː/	‘war’
/rɒt/	‘rot’	/wɒt/	‘what’

31 Appendix B Examples of front camera images of /r/ tokens of varying quality. The top image comes from a speaker who achieved relatively poor automatic classification scores, while the bottom image comes from a speaker who achieved perfect classification of /r/ versus /w/ using a CNN with a leave-one-out validation procedure.



BIBLIOGRAPHY

Abercrombie, David. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

Alwan, Abeer, Shrikanth Narayanan, and Katherine Haker. 'Toward Articulatory-Acoustic Models for Liquid Approximants Based on MRI and EPG Data. Part II. The Rhotics.' *The Journal of the Acoustical Society of America* 101.2 (1997): 1078–1089.

Armstrong, Nigel, and Tim Pooley. 'Levelling, Resistance and Divergence in the Pronunciation of English and French.' *Language Sciences* 39 (2013): 141–150. Universalism and Variation in Phonology: Papers in Honour of Jacques Durand.

Ashton, Helen, and Sarah Shepherd. *Work on Your Accent*. London: Collins, 2012.

Bates, Douglas et al. 'Fitting Linear Mixed-Effects Models Using Lme4.' *Journal of Statistical Software* 67.1 (2015): 1–48.

Boyce, Suzanne, and Carol Y. Espy-Wilson. 'Coarticulatory Stability in American English /r/.' *The Journal of the Acoustical Society of America* 101.6 (1997): 3741–3753.

Brown, Gillian. 'Consonant Rounding in British English: The Status of Phonetic Descriptions as Historical Data'. *Towards a History of Phonetics*. Ed. R.E Asher and Eugénie J.A. Henderson. Edinburgh: Edinburgh University Press, 1981. 67–76.

Catford, John C. *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press, 1977.

Chen, Liang-Chieh et al. 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.' *Computer Vision – ECCV 2018*. Ed. Vittorio Ferrari et al. Cham: Springer International Publishing, 2018. 833–851. Lecture Notes in Computer Science.

Delattre, Pierre, and Donald C. Freeman. 'A Dialect Study of American r's by X-Ray Motion Picture.' *Linguistics* 6.44 (1968): 29–68.

Docherty, Gerard J., and Paul Foulkes. 'Variability in (r) Production-Instrumental Perspectives.' *'R-Atics: Sociolinguistic, Phonetic and Phonological Characteristics of /r/*. Ed. H. Van de Velde and R. van Hout. Brussels, Belgium: Université Libre de Bruxelles, 2001. 173–184.

Espy-Wilson, Carol Y. 'Acoustic Measures for Linguistic Features Distinguishing the Semivowels / w j r l/ in American English.' *The Journal of the Acoustical Society of America* 92.2 (1992): 736–757.

Espy-Wilson, Carol Y et al. 'Acoustic Modeling of American English /r/.' *The Journal of the Acoustical Society of America* 108.1 (2000): 343–356.

Fant, Gunnar. *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton, 1960.

Ferragne, Emmanuel. 'Phonetics and Artificial Intelligence: Ready for the Paradigm Shift?' *Presented at the Phonology of Contemporary English Conference*. Aix-en-Provence, France, 2019.

Ferragne, Emmanuel, Cédric Gendrot, and Thomas Pellegrini. 'Towards Phonetic Interpretability in Deep Learning Applied to Voice Comparison.' *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Ed. Sasha Calhoun et al. Canberra, Australia: Australasian Speech Science and Technology Association Inc, 2019. 790–794.

Fong, Ruth C., Walter J. Scheirer, and David D. Cox. 'Using Human Brain Activity to Guide Machine Learning.' *Scientific Reports* 8.1 (2018): 1–10.

Foulkes, Paul, and Gerard J. Docherty. 'Another Chapter in the Story of /r/: "Labiodental" Variants in British English.' *Journal of sociolinguistics* 4.1 (2000): 30–59.

Gimson, Alfred Charles. *An Introduction to the Pronunciation of English*. London: Arnold, 1980.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

Guenther, Frank H. et al. 'Articulatory Tradeoffs Reduce Acoustic Variability during American English /r/ Production.' *The Journal of the Acoustical Society of America* 105.5 (1999): 2854–2865.

Havenhill, Jonathan, and Youngah Do. 'Visual Speech Perception Cues Constrain Patterns of Articulatory Variation and Sound Change.' *Frontiers in Psychology* 9 (2018): 728.

He, Kaiming et al. 'Deep Residual Learning for Image Recognition.' *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: The Institute of Electrical and Electronics Engineers, Inc, 2016. 770–778.

Heffner, Roe-Merrill Secrist. *General Phonetics*. Madison WI: University of Wisconsin Press, 1950.

Heselwood, Barry, and Leendert Plug. 'The Role of F2 and F3 in the Perception of Rhoticity: Evidence from Listening Experiments.' *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*. Ed. Wai-Sum Lee and Eric Zee. Hong Kong: City University of Hong Kong, 2011. 867–870.

Heyne, Matthias et al. 'The Articulation of /ɹ/ in New Zealand English.' *Journal of the International Phonetic Association* (2018): 1–23.

Jones, Daniel. *An Outline of English Phonetics*. 9th Edition. Cambridge: Cambridge University Press, 1972.

Kerswill, Paul. 'Children, Adolescents, and Language Change.' *Language Variation and Change* 8.2 (1996): 177–202.

King, Hannah, and Emmanuel Ferragne. 'Loose Lips and Tongue Tips: The Central Role of the /r/-Typical Labial Gesture in Anglo-English.' *Journal of Phonetics* 80 (2020): 100978.

King, Hannah, and Emmanuel Ferragne. 'The Contribution of Lip Protrusion to Anglo-English /r/: Evidence from Hyper- and Non-Hyperarticulated Speech'. *Proceedings of Interspeech* (2019): 3322–3326.

Knight, Rachael-Anne, Christina Villafaña Dalcher, and Mark J. Jones. 'A Real-Time Case Study of Rhotic Acquisition in Southern British English.' *Proceedings of the 16th International Congress of Phonetic Sciences*. Ed. Jürgen Trouvain and William J. Barry. Saarbrücken, Germany: Universität des Saarlandes, 2007. 1581–4.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune Haubo Bojesen Christensen. 'LmerTest Package: Tests in Linear Mixed Effects Models.' *Journal of Statistical Software* 82.13 (2017).

Ladefoged, Peter, and Ian Maddieson. *The Sounds of the World's Languages*. Oxford: Blackwell, 1996.

Laver, John. *The Phonetic Description of Voice Quality: Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press, 1980.

Lehiste, Ilse. *Acoustical Characteristics of Selected English Consonants*. Ann Arbor: University of Michigan Communication Sciences Laboratory, 1962.

Lilly, Richard, and Michel Viel. *La Prononciation de l'Anglais : Règles Phonologiques et Exercices de Transcription*. Paris, France: Hachette, 1977.

Lisker, Leigh. 'Minimal Cues for Separating /w, r, l, y/ in Intervocalic Position.' *Word* 13.2 (1957): 256–267.

Llamas, Carmen. 'Language Variation and Innovation in Middlesborough: A Pilot Study.' *Leeds Working Papers in Linguistics and Phonetics* 6 (1998): 97–114.

Marchal, Alain. *From Speech Physiology to Linguistic Phonetics*. Vol. 145. London: Wiley, 2009.

Marsden, Sharon. 'A Sociophonetic Study of Labiodental /r/ in Leeds.' *Leeds Working Papers in Linguistics and Phonetics* 11 (2006): 153–172.

MATLAB and Computer Vision Toolbox Release 2019b. Natick, MA, USA.: The MathWorks, Inc.

MATLAB and Deep Learning Toolbox Release 2019b. Natick, MA, USA.: The MathWorks, Inc.

MATLAB and Image Processing Toolbox Release 2019b. Natick, MA, USA.: The MathWorks, Inc.

Mayr, Robert. 'What Exactly Is a Front Rounded Vowel? An Acoustic and Articulatory Investigation of the NURSE Vowel in South Wales English.' *Journal of the International Phonetic Association* 40.1 (2010): 93–112.

Mielke, Jeff, Adam Baker, and Diana Archangeli. 'Individual-Level Contact Limits Phonological Complexity: Evidence from Bunched and Retroflex /ɹ/.' *Language* 92.1 (2016): 101–140.

O'Connor, Joseph D. et al. 'Acoustic Cues for the Perception of Initial /w, j, r, l/ in English.' *Word* 13.1 (1957): 24–43.

O'Connor, Joseph D. *Better English Pronunciation*. Cambridge: Cambridge University Press, 1967.

Proctor, Michael et al. 'Articulatory Characterization of English Liquid-Final Rimes.' *Journal of Phonetics* 77 (2019): 100921.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna Austria: R Foundation for Statistical Computing, 2018.

Roach, Peter. *English Phonetics and Phonology: A Practical Course*. Second. Cambridge: Cambridge University Press, 1983.

Sattherthwaite, F.E. 'An Approximate Distribution of Estimates of Variance Components.' *Biometrics Bulletin* 2.6 (1946): 110–114.

Scobbie, James M. '(R) as a Variable.' *The Encyclopaedia of Language and Linguistics.* Ed. Keith Brown. 2nd Edition. Vol. 10. Oxford: Elsevier, 2006. 337–344.

Simonyan, Karen, and Andrew Zisserman. 'Very Deep Convolutional Networks for Large-Scale Image Recognition.' *Proceedings of the 3rd International Conference of Learning Representations (ICLR)*. San Diego, CA, USA, 2015.

Singmann, Henrik et al. *Afex: Analysis of Factorial Experiments*, 2015. R Package, Version 0.13–145.

Smith, Bridget J. et al. 'Sound Change and Coarticulatory Variability Involving English /ɹ/.' *Glossa: a journal of general linguistics* 4(1).63 (2019): 1–51.

Stevens, Kenneth N. *Acoustic Phonetics*. Vol. 30. Cambridge, MA: MIT press, 1998.

Sweet, Henry. *A Handbook of Phonetics*. Vol. 2. Oxford, UK: Clarendon Press, 1877.

Tiede, Mark K., Suzanne E. Boyce, Christy K. Holland, et al. 'A New Taxonomy of American English /r/ Using MRI and Ultrasound.' *The Journal of the Acoustical Society of America* 115.5 (2004): 2633–2634.

Tiede, Mark K., Suzanne E. Boyce, Carol Y. Espy-Wilson, et al. 'Variability of North American English /r/ Production in Response to Palatal Perturbation.' *Speech Motor Control: New Developments in Basic and Applied Research*. Ed. Ben Maassen and Pascal van Lieshout. Oxford: Oxford University Press, 2010. 53–68.

Traunmüller, Hartmut, and Niklas Öhrström. 'Audiovisual Perception of Openness and Lip Rounding in Front Vowels.' *Journal of Phonetics* 35.2 (2007): 244–258.

Trudgill, Peter. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press, 1974.

Twist, Alina et al. 'Are "Covert" /ɹ/ Allophones Really Indistinguishable?' *Penn Working Papers in Linguistics* 13.2 (2007): article 16.

Villafaña Dalcher, Christina, Mark J. Jones, and Rachael-Anne Knight. 'Cue Switching in the Perception of Approximants: Evidence from Two English Dialects.' *Penn Working Papers in Linguistics* 14.2 (2008): Selected Papers from NWAV 36.

Wells, John C. *Accents of English*. 3 vols. Cambridge: Cambridge University Press, 1982.

Williams, Ann, and Paul Kerswill. 'Dialect Levelling: Change and Continuity in Milton Keynes, Reading and Hull.' *Urban Voices: Accent Studies in the British Isles*. Ed. Paul Foulkes and Gerard Docherty J. London: Arnold, 1999. 141–162.

Wood, Sidney. 'The Acoustical Significance of Tongue, Lip, and Larynx Maneuvers in Rounded Palatal Vowels.' *The Journal of the Acoustical Society of America* 80.2 (1986): 391–401.

Zawadzki, Paul A., and David P. Kuehn. 'A Cineradiographic Study of Static and Dynamic Aspects of American English /r/.' *Phonetica* 37.4 (1980): 253–266.

Zeiler, Matthew D., and Rob Fergus. 'Visualizing and Understanding Convolutional Networks'. *Computer Vision – ECCV 2014*. Ed. David Fleet et al. Cham: Springer International Publishing, 2014. 818–833. Lecture Notes in Computer Science.

Zhang, Zhaoyan et al. 'Acoustic Strategies for Production of American English "retroflex" /r/.' *Proceedings of the 15th International Congress of Phonetic Sciences*. Ed. M.J. Solé, D. Recasens, and J. Romero. Barcelona: Universitat Autònoma, 2003. 1125–1128.

NOTES

1. Notable examples include the depiction of Pontius Pilate in Monty Python's 'Life of Brian' and the treatment of the English football manager Roy Hodgson in the tabloid media with front page headlines including 'Bwing on the Euwos (We'll see you in Ukwaine against Fwance)'.
2. As suggested by an anonymous reviewer, the observed disparity in the treatment of labialisation in English /r/ may be due to differences in the manuals' target audience and date of publication. However, we note that the most recent second edition of Ashton and Shepherd's manual was published this year (2020) and targets all learners of English regardless of their first language.
3. In the interest of readability, some of the technical aspects of our deep learning analyses have been reduced in the current paper. However, we invite interested readers to consult the accompanying GitHub which includes both of the models presented in this paper (https://github.com/emmanuelferragne/labiodentalsRHere).
4. We note that the term 'semantic' in 'semantic segmentation' simply refers to the fact that this particular technique assigns labels to objects in a picture.
5. Global accuracy is a measure of the ratio of correctly classified pixels to the total number of pixels.
6. Mean BF score is a measure of how close the predicted boundary of an object matches the manually segmented boundary.
7. The 'regionprops' function in the *MATLAB Image Processing Toolbox* was used to compute ellipse parameters.

ABSTRACTS

The secondary labial articulation which accompanies the post-alveolar approximant /r/ in English has attracted far less attention from linguists than the primary lingual one. However, the lips may be particularly important in the variety of English spoken in England, Anglo-English, because non-lingual labiodental articulations ([ʋ]) are on the rise. Labiodentalisation may be due to speakers retaining the labial gesture at the expense of the lingual one, implying that /r/ is always labiodental even in lingual productions. We verify this assumption by comparing the labial postures of /r/ and /w/ in Anglo-English speakers who still present a lingual component. If post-alveolar /r/ is labiodental, the labial gesture for /w/, which is unequivocally considered rounded, should differ considerably. Techniques from deep learning were used to automatically classify and measure the lip postures for /r/ and /w/ from static images of the lips in 23 speakers. Our results suggest that there is a recognisable difference between the lip postures for /r/ and /w/, which a convolutional neural network is able to detect with a very high degree of accuracy.

Measurements of the lip area acquired using an artificial neural network suggest that /r/ indeed has a labiodental-like lip posture, thus providing a phonetic account for labiodentalisation. We finish with a discussion of the methodological implications of using deep learning for future analyses of phonetic data.

L'articulation labiale secondaire qui accompagne l'approximante post-alvéolaire /r/ en anglais a beaucoup moins suscité l'intérêt des linguistes que son articulation primaire, linguale. Or les lèvres peuvent présenter un intérêt tout particulier dans la variété d'anglais parlée en Angleterre car les réalisations labiodentales sans geste lingual ([ʋ]) sont en voie d'expansion. La labiodentalisation résulte probablement de la préservation d'un geste labial aux dépens du geste lingual, ce qui impliquerait que /r/ soit toujours labiodental, y compris dans les productions linguales. Nous vérifions cette hypothèse en comparant la configuration des lèvres du /r/ et du /w/ chez des locuteurs d'anglais d'Angleterre qui ont conservé la composante linguale dans leur production. Si le /r/ post-alvéolaire est labiodental, le geste labial du /w/, qui est unanimement considéré comme arrondi, devrait être très différent. Nous avons utilisé des techniques de l'apprentissage profond afin de classer automatiquement et de mesurer la configuration labiale de /r/ et /w/ à partir d'images des lèvres de 23 locuteurs. Nos résultats suggèrent qu'il existe bel et bien une différence nette de configuration labiale entre /r/ et /w/, qu'un réseau de neurones artificiels à convolution est capable de détecter avec une très grande précision. Des mesures effectuées automatiquement au niveau des lèvres au moyen d'un réseau de neurones artificiels montrent que /r/ a effectivement une configuration des lèvres de type labiodental, ce qui nous permet de décrire précisément la réalisation phonétique de cette labiodentalisation. Nous finirons avec une discussion des implications méthodologiques de l'utilisation de l'apprentissage profond dans les analyses phonétiques.

INDEX

Mots-clés: labialisation, rhotiques, apprentissage profond, changements linguistiques, anglais d'Angleterre
Keywords: labialisation, rhotics, deep learning, sound change, Anglo-English

AUTHORS

HANNAH KING

CLILLAC-ARP, EA 3967, Université de Paris
hannahhmking@gmail.com

EMMANUEL FERRAGNE

Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Université Sorbonne Nouvelle – Paris 3
emmanuel.ferragne@u-paris.fr