INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The contribution of lip protrusion to Anglo-English /r/: Evidence from hyper- and non-hyperarticulated speech

Hannah King[1], *Emmanuel Ferragne*[2]

[1]CLILLAC-ARP, EA 3967, Université de Paris, France
[2]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS - Sorbonne Nouvelle), France
hannah.king@univ-paris-diderot.fr, emmanuel.ferragne@univ-paris-diderot.fr

Abstract

Articulatory variation of /r/ has been widely observed in rhotic varieties of English, particularly with regards to tongue body shapes, which range from retroflex to bunched. However, little is known about the production of /r/ in modern non-rhotic varieties, particularly in Anglo-English. Although it is generally agreed that /r/ may be accompanied by lip protrusion, it is unclear whether there is a relationship between tongue shape and the accompanying degree of protrusion. We present acoustic and articulatory data (via ultrasound tongue imaging and lip videos) from Anglo-English /r/ produced in both hyper- and non-hyperarticulated speech. Hyperarticulation was elicited by engaging speakers in error resolution with a simulated "silent speech" recognition programme. Our analysis indicates that hyperarticulated /r/ induces more lip protrusion than non-hyperarticulated /r/. However, bunched /r/ variants present more protrusion than retroflex variants, regardless of hyperarticulation. Despite some methodological limitations, the use of Deep Neural Networks seems to confirm these results. An articulatory trading relation between tongue shape and accompanying lip protrusion is proposed.

Index Terms: rhotics, Anglo-English, lips, hyperarticulation, articulatory-acoustic trading relations

1. Introduction

The approximant consonant /r/ has been described as one of the most complex phones in the English language, particularly regarding its articulatory characteristics [1]. Its lingual constriction is described as belonging to a continuum of possible tongue configurations between two extreme configurations: tip-up, curled-up retroflex; and tip-down bunched [2, 3, 4]. Some speakers use one tongue configuration exclusively, while others present consistent but individual allophones, conditioned by syllable position, coarticulation and prosody [5]. The extreme form of retroflexion involving a curled-up tongue tip is generally considered more common in non-rhotic British English varieties than rhotic ones [2, 6]. Despite potential articulatory differences between rhotic and non-rhotic varieties, the vast majority of recent articulatory research involving /r/ has focused on rhotic English (North American [5, 7, 8]; Scottish [9, 10, 11]). However, similar articulatory patterns to those found in American English /r/ have been observed in non-rhotic New Zealand English [12] and in a small-scale study of Anglo-English [13].

It is generally agreed that English /r/ may be accompanied by a lip constriction [2, 14]. It has been observed, at least in rhotic varieties, that lip rounding is likely to occur in prevocalic and pre-stress syllable positions [2, 3]. On the other hand, [6] suggests that lip rounding is largely a function of the quality of the following vowel in Anglo-English. Despite these observations, most accounts of English /r/ focus on its lingual features [15]. However, the lips are of particular interest in Anglo-English. [16] informally observed that between 25 and 50 percent of nonbroadcasters interviewed on United Kingdom radio and television labialised /r/ at least some of the time. Furthermore, labiodental variants ([ʋ]) are becoming increasingly common in younger speech [15]. It is generally implied that labiodental variants lack a lingual gesture [17, 18]. Indeed, [13] observed one participant who produced labiodental /r/ with no obvious tongue body gesture. However, another participant presented labiodentalisation accompanied by a tip-up tongue configuration. It seems unlikely then that labiodental variants always lack a lingual gesture, but more research is needed regarding the exact relationship between the tongue and lips.

The acoustic profile of the different articulatory configurations for /r/ are remarkably indistinguishable, at least with regards to the first three formants. As a result, /r/ is considered to exhibit a "many-to-one articulatory-acoustic relationship" [19]; its main acoustic correlate being a very low third formant (F3). Acoustic modelling has associated this low F3 with a large front cavity volume, i.e. between the palatal constriction and the lips [19, 20]. These models predict that extending the front cavity will lower F3, which may be achieved through a more posterior placement of the tongue, the addition of a sublingual space, or lip protrusion. Trading relations have been observed between various articulatory manoeuvres which reciprocally contribute to the lowering of F3, enabling speakers to produce a stable acoustic output with different configurations [20, 21, 22]. As bunched /r/ is formed with the tongue-tip down, it has negligible sublingual space. As a result, [20] posit a trading relation between the sublingual space (for retroflexes) and a more posterior palatal constriction for bunched /r/. It does not seem unlikely then that similar articulatory-acoustic trade-offs may be observed for lip protrusion in /r/. However, there is no study to date that investigates this idea. We hypothesise that in order to compensate for its lack of sublingual space, bunched /r/ will be accompanied by more lip protrusion than retroflex /r/. To our knowledge, two existing studies have indeed observed a positive correlation between lip protrusion and bunching in both Anglo-English [13] and American English [23], although detailed accounts as to why have yet to be given.

To test to what extent lip protrusion contributes to /r/, this paper presents lip, tongue and acoustic data from Anglo-English productions in both non-hyper- and hyperarticulated speech. If the final goal of speech movements is the correct perception of speech by the listener, the goal of hyperarticulation must be to enhance the discriminability of phonetic categories (as expressed in Lindblom's H&H Theory [24]). If the acoustic goal of English /r/ is indeed a low F3, we hypothesise that hyperarticulated /r/ will reach even lower F3 values than those observed in non-hyperarticulated speech. If lip protrusion contributes to

http://dx.doi.org/10.21437/Interspeech.2019-2851

the lowering of F3, and therefore to the discernibility of /r/, we expect to find more lip protrusion in hyperarticulated speech than in non-hyperarticulated speech. Finally, if a trading relation between a sublingual space and lip protrusion exists, we may observe a larger degree of lip protrusion in bunched /r/ than in retroflex. In hyperarticulated speech, retroflexers may attain lower F3 values by increasing the size of the sublingual space (i.e. more retroflexion), a strategy which would not be available to bunchers. We therefore hypothesise that hyperarticulated bunched /r/ will be accompanied by more protrusion than hyperarticulated retroflex variants.

2. Methodology

2.1. Participants

29 native speakers of Anglo-English were recorded at Queen Margaret University, Edinburgh. Some speakers were excluded due to articulatory data visualisation issues (n=4) and one English-Punjabi bilingual was excluded because Punjabi also has retroflex consonants in its inventory. We present data from the remaining 24 speakers (22F, 2M) aged between 18 and 55 (M=29.71 $\pm$ 11.07) who come from all over England.

2.2. Equipment

Simultaneous articulatory and acoustic data were recorded using Articulate Assistant Advanced (AAA) software [25]. Participants wore a headset to ensure ultrasound probe stabilisation [26]. Attached was an Audio-Technica AT803 microphone and two NTSC micro-cameras to capture front and profile lip videos. Ultrasound recordings were recorded at a rate of circa 121 fps using a Sonix RP system. Lip videos were obtained at a rate of circa 60 fps. Audio files were digitalized as PCM mono files with a 22050 Hz sampling rate and 16-bit quantization.

2.3. Procedure

Speech has been found to be hyperarticulated in computer- compared with human-directed speech [27], particularly in speech following recognition errors [28, 29]. If only one segment is incorrectly identified, or is likely to be misunderstood, speakers may limit and target their adaptations to that particular segment [30, 31, 32]. Adaptations may occur at the prosodic level by speaking more slowly and loudly, modifying pitch, and adding more pauses. At a segmental level, speakers have been shown to replace reduced or assimilated forms with more canonical ones [30]. For this study, in order to elicit targeted hyperarticulation specifically at a segmental level, we engaged speakers in error resolution with a simulated speech recognition programme. Speakers were deceptively informed that the aim of the experiment was to test a new automatic "silent speech" reader, which used information from speech movements to recognise the word they had said. The experiment was divided into two parts. During the first, speakers were informed the computer had access to visual and audio cues from speech and as a result, the programme correctly "identified" every word uttered. This first part provided us with baseline, non-hyperarticulated (or at least non-contrastive) productions of /r/. During the second part, participants were informed that the audio would be turned off and that the programme would only have access to visual speech information. During this second part, the computer "incorrectly" identified one third of the stimuli. Whenever computer errors occurred, participants were instructed to repeat the word to try to make the computer understand. Each "incorrectly" identified word was repeated two more times. Recording sessions lasted no longer than 30 minutes. Steps were taken to ensure the believability of the simulated programme. After each production, participants saw the message "processing...please wait", which gave time for the experimenter, who was in an adjacent control room, to select the appropriate computer response. A phoney programme interface was created and presented to speakers on a separate screen throughout the recordings. Fake on/off buttons were shown next to the words "audio", "video" and "ultrasound". Just before the second "silent speech" part started, the experimenter "turned off" the audio by clicking on the corresponding fake button.

2.4. Stimuli

Stimuli comprised of nine /r/-initial monosyllabic words followed by the vowels /iː, uː, ɪ, ɛ, æ, ʌ, ɔː, ɒ/. Fillers were /w/-initial words followed by the same monophthongs. In the non-hyperarticulated session, stimuli were "correctly" identified by the simulated programme. To ensure believability, one repetition per item was recorded in the first session. For the second hyperarticulated session, /r/ in the words "reed", "red", and "room" were "incorrectly" identified as /w/ and /l/ (e.g. "red" was identified as "wed" or "led"). When an incorrect response was given, the original word was repeated two more times. The same method was used for /w/-initial filler words, where /w/ was mistaken for /r/ or /l/. A total of 24 productions of /r/ were recorded in the second session. All stimuli were randomised.

2.5. Data analysis

2.5.1. Ultrasound tongue imaging

One ultrasound frame was selected per recording depicting the maximum of the anterior lingual gesture for /r/ just before movement into the vowel. Tongue configurations for /r/ were visually classified into five types: Mid Bunched, Front Bunched, Front Up, Tip Up, Curled Up. The first two are bunched (i.e. tip down), while the remaining three are retroflex (i.e. tip up). We supplemented the classification of the first four types presented in [11] with the final Curled Up type to distinguish between the tip up /r/ without curling up and the curled-up retroflex. Curling up of the tongue tip results in a near-parallel orientation of the tongue surface to the ultrasound scanlines, which produces artefacts in the ultrasound image [33]. In particular, we observe a bright white region above where the tongue tip is expected [5] and a discontinuity in the tongue contour where the tongue is curled up [34].

2.5.2. Lip protrusion

Lip protrusion was calculated in relation to a neutral lip position before speech in AAA. The image corresponding to maximum protrusion for each /r/ production was visually selected from the profile lip video. One image was also chosen per speaker depicting a neutral lip position prior to speech. A horizontal fiducial line was positioned to intersect the lip corner during each speaker's neutral image and was used as a reference for protrusion measures. Another line was positioned to touch the lower and upper lip edge, intersecting the neutral lip corner fiducial. As the fiducial had previously been scaled (in centimetres) to a physical ruler positioned along the midline of the headset, AAA could calculate the distance from the origin of the fiducial to where the lip edge line crossed. The neutral lip distance measurement was subtracted from the maximum protrusion distance for /r/ yielding final protrusion values, as depicted in Figure 1.

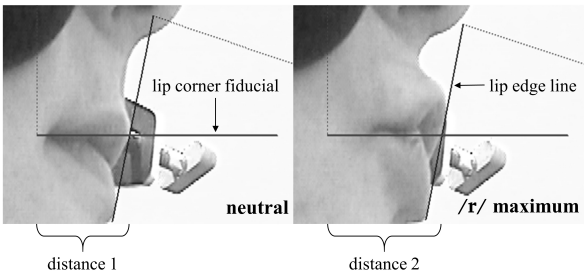

Figure 1: *Lip protrusion measure. Distance 1 is subtracted from distance 2.*

2.5.3. Acoustics

The acoustic data were exported as wav files from AAA and analysed in Praat [35]. Using Praat's Burg algorithm, the first three formants (F1, F2, F3) were extracted at the point of minimal F3 during /r/ [21] and at the midpoint of a steady state of the following vowel, avoiding obvious transitions to and from surrounding consonants. Formant parameters were manually adjusted in order to reach an optimal match between Praat's formant estimation and the underlying spectrogram.

2.5.4. Statistical analysis

Statistical analysis was implemented in R [36] using the *lme4* package [37] to perform linear mixed effects (LME) and generalised linear mixed effects (GLMM). Where appropriate, fixed effects were centred to improve model convergence. We tested the significance of main effects and interactions (when justified) to model fit using likelihood ratio tests. Model comparison was carried out using likelihood ratio tests and a comparison of Akaike Information Criterion (AIC). Model residuals were plotted to test for deviations from homoscedasticity or normality.

2.5.5. Deep Neural Networks

ResNet-18 [38], a well-known convolutional neural network architecture in image recognition, was retrained from scratch to perform two classification tasks – predict /r/ type and predict hyperarticulation – based on three types of images: ultrasound, front camera, and profile camera. The images were resized to 224×224 pixels to match the expected input size of ResNet-18. Overfitting being a common issue with such small datasets, 10-fold cross-validation – each model was re-trained ten times with a different test set – and data augmentation – training images were subject to random rotation, scaling, and translation – were used. Our analysis was complemented with visualisations of Class Activation Maps (CAM) [39] in order to ensure that the model had learnt meaningful representations. The CAM technique was applied to visualise the activations of the ReLU layer obtained after the last convolution layer in ResNet-18. The whole workflow was implemented with Matlab Deep Learning Toolbox [40].

3. Results and discussion

3.1. Predicting /r/ type

In both hyper- and non-hyperarticulated speech, seven speakers use only bunched /r/ types, fourteen only retroflex, and three use both. Therefore, our findings do not support the suggestion that Anglo-English /r/ is exclusively retroflex [2, 6]. If we use the more detailed /r/ types we established in Section 2.5.1, nine speakers had one canonical /r/ type. All other speakers used at least two types with one even presenting all five. Previous studies have found that tongue shape for /r/ may be influenced by the following vowel. For example, it has been observed that retroflexion is favoured by back and perhaps by low vowels [23]. We performed a LME analysis to examine to what extent /r/ type can be predicted based on lip protrusion (in mm), the following vowel (centred F1 & F2) and context (non- vs. hyperarticulated). As the different /r/ types are said to be on a continuum, the outcome variable was coded on a scale from one to five, one being the most bunched (Mid Bunched) and five being the most retroflex (Curled Up). The final model we present in Table 1 had F1, F2, lip protrusion and context as significant main predictors, a random slope for context within speaker and a random intercept for item. As the table suggests, the higher the F1 and the lower the F2 of the following vowel, the higher the chances of retroflexion. If we accept that F1 is an acoustic correlate of tongue height and F2 of tongue position, our results corroborate those in [23]: retroflexion is favoured by low and by back vowels. Interestingly, we find a significant effect of context: retroflexion rate significantly increases in hyperarticulated speech. Finally, we find the opposite effect for protrusion: the more protrusion, the lower the chances of retroflexion. This result is in line with our hypothesis that there is a trading relation between lip protrusion and sublingual space. A significant interaction was not observed between protrusion and context. We therefore conclude that protrusion is favoured by lower values on the bunched-retroflex continuum (i.e. in bunched configurations), regardless of context.

Table 1: *Fixed effects predicting /r/ type. Positive values indicate greater retroflexion.*

Predictors	Estimates	CI	*p*
Vowel F1	0.138	0.031	$< 0.001^{***}$
Vowel F2	-0.209	0.031	$< 0.001^{***}$
Context hyper	0.227	0.079	0.004^{**}
Protrusion	-0.084	0.039	0.032^{*}
(Intercept)	3.500	0.320	$< 0.001^{***}$

Turning to neural networks, mean test accuracy (and standard deviation) after 10-fold cross-validation in the bunched-vs-retroflex classification task were: 97.82 % (2.29 %) ultrasound; 97.60 % (2.76%) front image; 96.64 % (3.23 %) profile image. On the face of it, these very high and consistent scores seem to strongly support the view that the model had learnt a reliable dichotomy between bunched and retroflex /r/ from articulatory differences in tongue and/or lip shape. However, a close inspection of CAM images reveals that this is not necessarily the case. While the highlighted region in the left panel of Figure 2 looks like a plausible basis for (correctly) deciding that this /r/ is bunched, the right panel tells a different story for this correctly classified retroflex token. The highlighted zone shows that the model has used a piece of hardware to make its decision. As it turns out, /r/ type and speaker are partly confounded (i.e. some speakers produce one type only); so, it comes as no surprise that the model has learnt whatever was available to tell a buncher from a retroflexer, including non linguistic features (camera angle, hardware), and made good decisions for bad reasons. Contrary to our lip images, there were presumably much fewer spurious artefacts for the model to fall back on in

the ultrasound images. In other words, the model was forced to turn its attention to relevant articulatory regions of interest – all the more so as data augmentation was probably more efficient to erase speaker-specific attributes from ultrasound data compared to other images – and to discard linguistically irrelevant information. We can therefore be quite confident that the model had learnt robust representations of tongue shape to differentiate a bunched from a retroflex /r/, but a preliminary look at the CAM images we obtained warrants further investigation.

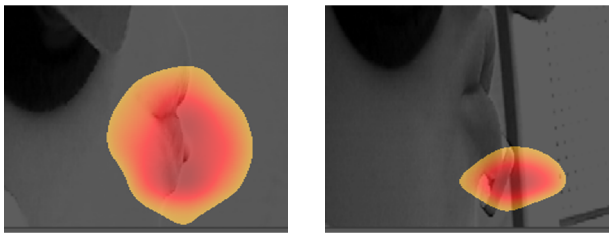

Figure 2: *CAM visualisations: bunched vs. retroflex.*

3.2. Predicting hyperarticulation

To assess to what extent hyperarticulation can be predicted based on /r/ type, lip protrusion and acoustics (F1, F2, F3 of /r/), we performed a GLMM analysis with context (non- vs. hyperarticulated) as the binary outcome variable. To improve model convergence, /r/ types were divided into two categories: retroflex and bunched. The final model we present in Table 2 had F3, protrusion, and /r/ type as significant main predictors, and random intercepts for speaker and the following vowel. The model follows our prediction that hyperarticulated /r/ exhibits more lip protrusion and lower F3 values. We previously remarked that an alternative hyperarticulation strategy for /r/ could be the use of more retroflexion, which this model corroborates, as it predicts significantly more instances of retroflexion in hyperarticulated speech. Neither F1 nor F2 of /r/ came out as significant predictors for hyperarticulation. This is interesting because the lowering of all formants would be the expected acoustic consequence of greater lip protrusion, and not just F3. Although this analysis cannot tell us to what extent F3 lowering is the result of changes in lip protrusion or in tongue configuration, our results suggest that speakers actively control articulatory parameters in order to enhance the discriminability of /r/.

Table 2: *Fixed effects predicting hyperarticulation.*

Predictors	Estimates	CI	*p*
F1	-0.067	0.186	0.719
F2	-0.115	0.180	0.522
F3	-0.713	0.216	$< 0.001^{***}$
Protrusion	2.701	0.302	$< 0.001^{***}$
/r/ type RETROFLEX	1.683	0.533	0.002^{**}
(Intercept)	2.485	1.988	0.212

Neural network mean test accuracy (and standard deviation) after 10-fold cross-validation in the non- vs. hyperarticulated classification task were: 77.82 % (12.04 %) ultrasound; 88.44 % (4.90 %) front image; 70.94 % (16.99 %) profile image. Although all scores are statistically significant according to binomial tests, it should be noted that they are not only smaller than those found in Section 3.1, but they are also more variable, in particular when the models are trained with the ultrasound and profile images. However, contrary to the bias mentioned in Section 3.1, CAM visualisations are more consistent with phonetic expectations. For example, the left panel of Figure 3 shows that the lips are highlighted for this correctly classified non-hyperarticulated /r/, which is phonetically plausible and interpretable. The right panel shows a misclassified token, and the reason for this is credible given the region the model looked at to make a decision.

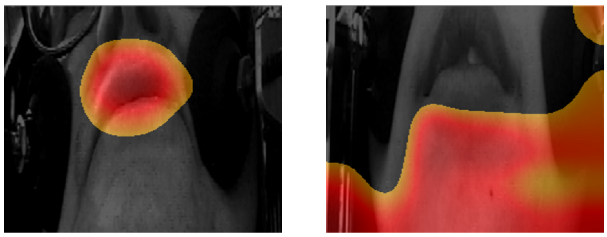

Figure 3: *CAM visualisations: non-hyperarticulated.*

4. Conclusions

Our analysis suggests that bunched /r/ tongue configurations induce more lip protrusion than retroflex ones, regardless of hyperarticulation. We also observed that increases in lip protrusion and retroflexion are significant hyperarticulation strategies for /r/. We conclude then that both the addition of a sublingual space and lip protrusion contribute to enhancing the discriminability of /r/. We also propose that bunchers use more lip protrusion to compensate for their negligible sublingual space, thus ensuring a stable acoustic output across all /r/ types.

Throughout this paper, we have assumed that the goal of hyperarticulation is an acoustic one. However, in our "silent speech" paradigm, speakers may be enhancing intelligibility in the visual domain rather than the acoustic one. However, enhancing visual cues cannot explain the difference in lip protrusion we observe between bunchers and retroflexers, unless retroflexers put less emphasis on the visual cue of lip protrusion than bunchers do, which is an equally interesting proposition. This leads us to question whether there is a perceptible difference between retroflex and bunched /r/ with their differing degrees of protrusion in both auditory and visual domains. We intend to work on this idea in the future.

On a methodological level, we have used techniques from deep learning to train models to learn articulatory differences from raw ultrasound and lip images. That convolutional neural networks learn their own representations from the data constitutes a promising research avenue for future phonetic studies. We have illustrated how the visualisation of activations not only makes neural networks' decisions more interpretable, but can also draw researchers' attention to potential biases in their studies. A logical extension will be to train models with whole videos rather than selected frames.

5. Acknowledgements

We would like to thank the Clinical Audiology, Speech and Language Research Centre for kindly allowing us to collect data in their facilities. We particularly wish to thank Eleanor Lawson, Jim Scobbie and Steve Cowen for their precious time and guidance. We also thank Ioana Chitoran for her invaluable input.

6. References

[1] M. Adler-Bock, B. M. Bernhardt, B. Gick, and P. Bacsfalvi, "The use of ultrasound in remediation of North American English /r/ in 2 adolescents," *American Journal of Speech-Language Pathology*, vol. 16, no. 2, pp. 128–139, 2007.

[2] P. Delattre and D. C. Freeman, "A dialect study of American r's by x-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[3] P. A. Zawadzki and D. P. Kuehn, "A cineradiographic study of static and dynamic aspects of American English /r/," *Phonetica*, vol. 37, no. 4, pp. 253–266, 1980.

[4] M. K. Tiede, S. E. Boyce, C. K. Holland, and K. A. Choe, "A new taxonomy of American English /r/ using MRI and ultrasound," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2633–2634, 2004.

[5] J. Mielke, A. Baker, and D. Archangeli, "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/," *Language*, vol. 92, no. 1, pp. 101–140, 2016.

[6] A. Gimson, *An Introduction to the Pronunciation of English*. London: Arnold, 1980.

[7] D. Dediu and S. R. Moisik, "Pushes and pulls from below: Anatomical variation, articulation and sound change," *Glossa: a journal of general linguistics*, vol. 4, no. 1, 2019.

[8] L. Magloughlin, "Accounting for variability in North American English /ɹ/: Evidence from children's articulation," *Journal of Phonetics*, vol. 54, pp. 51–67, 2016.

[9] E. Lawson, J. Stuart-Smith, and J. M. Scobbie, "The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1646–1657, 2018.

[10] J. M. Scobbie, E. Lawson, S. Nakai, J. Cleland, and J. Stuart-Smith, "Onset vs. coda asymmetry in the articulation of English /r/," in *Proceedings of the 18th International Congress of Phonetic Sciences*, University of Glasgow, Glasgow, 2015.

[11] E. Lawson, J. M. Scobbie, and J. Stuart-Smith, "The social stratification of tongue shape for postvocalic /r/ in Scottish English," *Journal of Sociolinguistics*, vol. 15, no. 2, pp. 256–268, 2011.

[12] M. Heyne, X. Wang, D. Derrick, K. Dorreen, and K. Watson, "The articulation of /ɹ/ in New Zealand English," *Journal of the International Phonetic Association*, pp. 1–23, 2018.

[13] N. Lindley and E. Lawson, "An articulatory investigation of Anglo-English prevocalic /r/," in *BAAP Colloquium*, Lancaster, 2016.

[14] C. Y. Espy-Wilson and S. Boyce, "A simple tube model for American English /r/," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999, pp. 2137–2140.

[15] G. Docherty and P. Foulkes, "Variability in (r) production-instrumental perspectives," *'R-atics: Sociolinguistic, phonetic and phonological characteristics of /r/. Etudes & Travaux 4*, pp. 173–184, 2001.

[16] J. M. Scobbie, "(R) as a variable," in *The Encyclopaedia of Language and Linguistics.*, 2nd ed., K. Brown, Ed. Oxford: Elsevier, 2006, vol. 10, pp. 337–344.

[17] D. Jones, *An Outline of English Phonetics*, 9th ed. Cambridge University Press, 1972.

[18] P. Foulkes and G. J. Docherty, "Another chapter in the story of /r/: 'Labiodental' variants in British English," *Journal of sociolinguistics*, vol. 4, no. 1, pp. 30–59, 2000.

[19] C. Y. Espy-Wilson, S. E. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of American English /r/," *The Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 343–356, 2000.

[20] A. Alwan, S. Narayanan, and K. 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*, vol. 101, no. 2, pp. 1078–1089, 1997.

[21] F. H. Guenther, C. Y. Espy-Wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, and J. S. Perkell, "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2854–2865, 1999.

[22] A. Nieto-Castanon, F. H. Guenther, J. S. Perkell, and H. D. Curtin, "A modeling investigation of articulatory variability and acoustic stability during American English /r/ production," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3196–3212, 2005.

[23] M. Tiede, S. E. Boyce, C. Y. Espy-Wilson, and V. L. Gracco, "Variability of North American English /r/ production in response to palatal perturbation," in *Speech Motor Control: New Developments in Basic and Applied Research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2011, pp. 53–67.

[24] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*. Springer, 1990, pp. 403–439.

[25] Articulate Instruments Ltd., *Articulate Assistant Advanced Ultrasound Module User Manual, Revision 2.16*. Edinburgh: Articulate Instruments Ltd., 2014.

[26] ——, *Ultrasound Stabilisation Headset Users' Manual, Revision 1.4*. Edinburgh: Articulate Instruments Ltd., 2008.

[27] D. Burnham, S. Joeffry, and L. Rice, "Computer-and human-directed speech before and after correction," *space*, vol. 6, p. 7, 2010.

[28] S. Oviatt, G. A. Levow, M. MacEachern, and K. Kuhn, "Modeling hyperarticulate speech during human-computer error resolution," in *, 4th International Conference on Spoken Language, 1996. ICSLP 96. Proceedings*, vol. 2, Oct. 1996, pp. 801–804 vol.2.

[29] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, Jun. 2009.

[30] A. J. Stent, M. K. Huffman, and S. E. Brennan, "Adapting speaking after evidence of misrecognition: Local and global hyperarticulation," *Speech Communication*, vol. 50, no. 3, pp. 163–178, 2008.

[31] J. Schertz, "Exaggeration of featural contrasts in clarifications of misheard speech in English," *Journal of Phonetics*, vol. 41, no. 3-4, pp. 249–263, 2013.

[32] E. Buz, M. K. Tanenhaus, and T. F. Jaeger, "Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations," *Journal of memory and language*, vol. 89, pp. 68–86, 2016.

[33] J. M. Scobbie, R. Punnoose, and G. Khattab, "Articulating five liquids: A single speaker ultrasound study of Malayalam," *Rhotics: New data and perspectives*, pp. 99–124, 2013.

[34] S. Bakst, "Differences in the relationship between palate shape, articulation, and acoustics of American English /r/ and /s/," *UC Berkeley Phonology Lab Annual Report*, 2016.

[35] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," Version 6.0.50, retrieved March 2019 from http://www.praat.org/, 2019.

[36] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

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

[38] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognitio," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 2921–2929.

[40] Mathworks, *MATLAB Deep Learning Toolbox R2019a*, Mathworks, Natick, MA, USA, 2019.