University of Pennsylvania Working Papers in Linguistics

Volume 22
Issue 2 *Selected Papers from New Ways of Analyzing Variation (NWAV 44)*

Article 3

12-1-2016

Automatic Detection of Sociolinguistic Variation Using Forced Alignment

George Bailey

Automatic Detection of Sociolinguistic Variation Using Forced Alignment

Abstract

Forced alignment software is now widely used in contemporary sociolinguistics, and is quickly becoming a crucial methodological tool as an increasing number of studies begin to utilise 'big data.' This study investigates the possibility of taking forced alignment one step further towards the goal of complete automation; specifically, it expands the functionality of FAVE-align to fully automate the coding of three sociolinguistic variables in British English: (th)-fronting, (td)-deletion, and (h)-dropping. This involved the expansion of pronouncing dictionaries to reflect the surface output of these variable rules; FAVE then compares the fit of competing acoustic models with the speech signal to determine the surface variant. It does so with an impressive degree of accuracy, largely comparable to inter-transcriber agreement for all variables; however, the pattern of its mistakes, which are largely false positives, suggests a difficulty in identifying the voiceless segments of (td) and (th). Although it is reassuring that inter-transcriber agreement was also lowest for these tokens, it should be noted that FAVE's accuracy decreases in faster speech rates while no comparable effect is found for agreement among human transcribers.

Automatic Detection of Sociolinguistic Variation Using Forced Alignment

George Bailey*

1 Introduction

The emergence of forced alignment and automatic vowel extraction software is arguably one of the most important methodological advances in modern-day sociolinguistics. The strong empirical and quantitative foundations of the field have led to the current trend of employing 'big data' on an unprecedented scale; datasets that fall under this description, by their very definition, are too large and complex to be adequately processed by traditional methods (Fruehwald 2015). Hand-measuring vowels in a dataset as large as the Philadelphia Neighbourhood Corpus, which consists of over 300 hours of sociolinguistic interviews, would be an enormous undertaking; with the advent of forced alignment, however, tools such as The University of Pennsylvania's Forced Alignment and Vowel Extraction suite (FAVE) can be employed to automate vowel formant measurements (Rosenfelder et al. 2011).

Forced alignment software time-aligns orthographic and phone-level transcriptions with a corresponding audio file, which facilitates a more efficient and more reliable analysis of language variation. Despite the success of FAVE in automating the vowel extraction process (see Labov et al. 2013 for an example of its application), there remains no equivalent for automating the procedure of coding consonantal variables that are categorical in nature. This study investigates the possibility of using such software to fully automate the coding of three consonantal variable rules, namely: (th)-fronting, (td)-deletion, (h)-dropping.

The accuracy rates found here are promising enough to validate this automated procedure as a reliable methodological tool; however, detailed analysis of the degree and patterning of FAVE's erroneous variant judgements highlights shortcomings that must be taken into account if automation of this kind is to be employed.

2 Background

Although the discussion to follow is based on experience of one specific aligner, FAVE-align, other aligners work in a largely similar fashion; as such, it can be taken as an overview of the alignment procedure in general, and the mechanism of expanding forced alignment to automatically code for these variable rules (as discussed in Section 3.2) should be applicable to other suites, although its exact implementation may differ.

2.1 Forced Alignment

Forced alignment typically takes two files as input: an audio file and its corresponding orthographic, word-level transcription demarcated into breath groups. It returns a Praat TextGrid where individual words, and their component phonemes, are time-aligned to the audio file (see Figure 1). This is achieved through a complex process that employs the Hidden Markov Model Toolkit (HTK), a natural language processor that performs acoustic analysis and speech recognition.

This speech recognition is made possible by a combination of two elements: a set of acoustic models for each phone in a given language variety (explored in Section 2.1.1), and a pronouncing dictionary listing every possible word in that variety with a corresponding phone-level transcription (see Section 2.1.2).

*Thanks to Laurel MacKenzie and Josef Fruehwald for their help with FAVE, as well as the audience at NWAV44 for their invaluable feedback on this paper. I would also like to gratefully acknowledge the financial support of the Economic and Social Research Council.

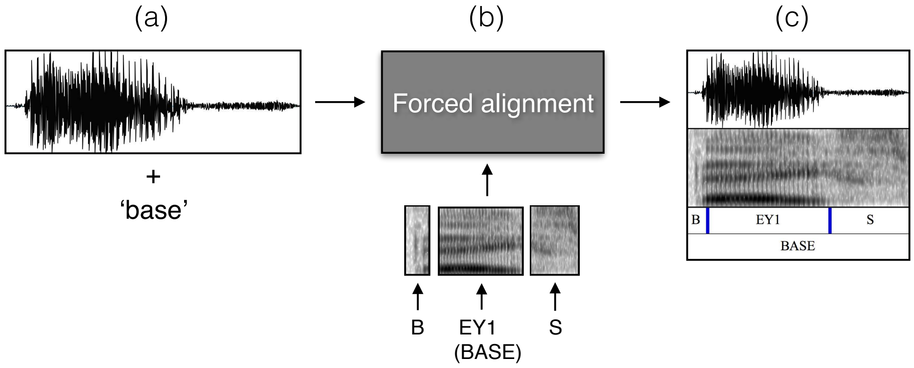

Figure 1: Diagrammatic representation of the forced alignment work-flow, including (a) the input of the audio along with an orthographic transcription, (b) the forced alignment process itself, using a pronouncing dictionary and the appropriate acoustic models, and (c) the output of a time-aligned TextGrid with phone- and word-level tiers.

2.1.1 Hidden Markov Models

Aligners require pre-trained acoustic models for each and every phone in a language's phonemic inventory; in the case of vowels, three models are required for each, specific to the levels of stress they may carry (primary, secondary, and unstressed). These acoustic descriptions take the form of Hidden Markov Models (HMMs), a statistical model consisting of 3–5 states and probability distributions representing the likelihood of outputs from, and transitions between, these states (Ghahramani 2001, Yuan et al. 2013). The exact details of how this dynamic Bayesian network operates is beyond the scope of this paper; it is enough to note for our purposes that the audio input is compared with these HMMs when performing automatic speech segmentation.

2.1.2 Pronouncing Dictionaries

Given that aligners usually take word-level transcriptions as input, they must make reference to an external pronouncing dictionary that provides a corresponding phone-level transcription. The Carnegie Mellon University dictionary is widely-used due to its extensive coverage of the lexicon (over 134,000 words) and its consistent stress-marking for vowels, the latter being a prerequisite for use with FAVE. The fact that this dictionary is based on General American for both the orthographic and phonological transcriptions somewhat restricts its applicability for studies of British varieties, though previous research has shown that this issue does not have a major impact on the reliability of aligning British English speech (MacKenzie and Turton 2013). For such purposes, one may turn to alternatives such as the British English Example Pronunciation Dictionary (BEEP), but this is inferior with respect to word coverage and stress-marking; overcoming the latter issue requires extensive pre-processing work (Agirrezabal et al. 2014).

Crucially for this study, although these pronouncing dictionaries provide broad phonemic transcriptions based on standard varieties such as General American or British RP, they *can* contain multiple entries for a single lexical item that has more than one possible realisation. This can either be due to variable stress patterns between noun-verb alternations like *present* (['pɹɛzənt] for the noun form, cf. [pɹə'zɛnt] for the verb), or allophonic variation in something like (ing).

It follows that the inclusion of separate dictionary entries reflecting different realisations of underlying forms requires the aligner to compare output probabilities from all potential HMMs before selecting the most likely pronunciation that fits the observed speech signal. This is the methodology employed by the current study (outlined in Section 3).

2.2 Available Aligners

FAVE (Rosenfelder et al. 2011) and PLA (Gorman et al. 2011) are two widely-used English language aligners. FAVE is based on the earlier p2fa suite, which has been trained on the SCOTUS corpus consisting of oral arguments from the Supreme Court of the US. Its forced alignment and automatic vowel extraction capabilities have been used on the Philadelphia Neighbourhood Corpus to great effect (Labov et al. 2013), and direct comparisons with other aligners have shown that it offers the most accurate alignment even when applied to non-American speech (MacKenzie and Turton 2013); its alignment is particularly robust even when dealing with the messiness of sociolinguistic interviews with overlapping speech and background noise. Its acoustic models, however, lack the flexibility offered by PLA, which allows users to self-train new speaker- or dialect-specific models based on given speech samples.

No matter the benefits or shortcomings of these aligners, they all share one common limitation: the need for a word-level transcription prior to alignment. Although this method of forced alignment is still more efficient than the manual segmentation of phones, an ideal system would be fully automatic. The recent Dartmouth Linguistic Automation service (DARLA) offers exactly this (Reddy and Stanford 2015); provided with only an audio file, it returns a force aligned Praat TextGrid and a detailed table of vowel measurements comparable with the output of FAVE-extract. This aligner is fully automatic, requiring no prior transcription, and combines the functionality of the FAVE software suite with an automatic speech recognition (ASR) system to almost entirely remove pre-processing time. The obvious consequence is that this extra step of automation, utilising ASR tools, introduces yet another source of error; this idea of a trade-off between accuracy and efficiency is a major point of contention in the domain of forced alignment, and is a point I will return to later.

3 Methodology

The aim of this study is to investigate the possibility of expanding the functionality of forced alignment, by testing its discriminative judgements when presented with multiple possible surface variants of a single underlying segment. Specifically, it tests FAVE's capability to discriminate between the realisations of three phonological variables in conversational data taken from an hour-long sociolinguistic interview that was conducted with a 20 year-old female speaker from Manchester in the north of England.[1] The processes of (th)-fronting and (h)-dropping have already been attested in Manchester English (Baranowski and Turton 2015), while (td)-deletion is widespread throughout varieties of English. The test stimuli yields 249 tokens of (h), 293 tokens of (td) and 364 tokens of (th).

The methodology employed here largely mirrors that used in Yuan and Liberman 2011, which investigated automatic coding of (ing), and Milne 2014, which tested forced alignment recognition of consonant cluster reduction in French.[2] Specifically, it involves the expansion of a standard pronunciation dictionary to include multiple phonemic transcriptions, each reflecting the surface output of a stochastic phonological process.

3.1 Measuring Alignment Accuracy

The accuracy of FAVE's variant discrimination will be evaluated by comparing its results to manually coded human judgements. Agreement rates between the automatic and manual coding schemes are presented in two measures: raw percentages and Cohen's kappa statistic (Cohen 1960). The latter measure corrects for expected agreement by chance and is therefore more conservative, and widely-used for measuring agreement rates among sets of categorically coded judgements. It is,

[1] The recording was made using a Sony PCM-M10 recorder and lavalier microphone, and later saved in uncompressed .wav format with a sampling rate of 44,100Hz.

[2] The methods of dictionary expansion in Yuan and Liberman 2011 do differ slightly from those described in this paper, in ways that will be explored in Section 6.1.

however, less interpretable unless an arbitrary discretization is employed (Carletta 1996). In contrast to this, raw percentage agreement potentially overestimates the rate of accuracy in automated coding, but is at least immediately interpretable.

Since the distinctions between some variant realisations can be subtle and difficult to determine even manually, a second round of human coding was carried out by a postgraduate student trained in phonetics. Inter-analyst agreement rates between these two sets of manually coded responses provide an important baseline against which the accuracy of FAVE's automatic coding can be evaluated.

3.2 Dictionary Expansion

Scripts were written in the Python programming language to carry out expansion of the pronouncing dictionary.[3] Taking the CMU dictionary as input, they identify all relevant words that fall within the envelope of variation for each of the three processes, as identified in Table 1, and perform the necessary string manipulation to create additional entries that reflect application of these variable rules. In total, this resulted in the addition of over 15,000 new dictionary entries.

Variable	**Surface form**	**Envelope of variation**	**Example word**	**Example dictionary entries**
(td)	/t, d/ ⟶ ø	Word-final /t/ or /d/ when preceded by a consonant	*just*	J AH1 S T J AH1 S
(th)	/θ, ð/ ⟶ [f, v]	Any /θ/ or /ð/	*north*	N AO1 TH N AO1 F
(h)	/h/ ⟶ ø	Word-initial, pre-vocalic /h/	*height*	H AY1 T AY1 T

Table 1: Details and example words for the three variable rules under investigation.

4 Results

This section will evaluate FAVE's success in automatically coding three categorical variable rules, and will address three separate aspects of its performance: Section 4.1 compares its accuracy across application and non-application of the three variables, Section 4.2 breaks these results down by the voicing of the target segments, while Section 4.3 investigates whether or not FAVE's variant discrimination is sensitive to speech rate.

4.1 Effect of Variable

Firstly, it should be established whether or not the success of FAVE's accuracy in automatically coding consonantal variation is dependent on the exact nature of such variation; that is, whether or not we find discrepancies in its coding accuracy between the three variables under study: (td)-deletion, (th)-fronting, and (h)-dropping. Given the possibility of biases towards identifying particular variants, it is also important to quantify its success separately for both application and non-application of each variable rule (e.g., if the aligner is particularly lenient in seeking out matches for the acoustic model for /h/, this will lead to very high success rates in identifying non-application of (h)-dropping, but consequently a very high rate of false negatives too, because it would likely 'find' /h/ when it isn't actually there).

Figure 2, and the corresponding agreement matrix in Table 2, both seem to indicate fairly impressive variant discrimination for both (h) and (th), with overall accuracy rates of 85.5% and 79.7%, respectively, but a somewhat less convincing success rate of 71% for (td)-deletion. This difficulty

[3] Along with the expanded pronouncing dictionaries themselves, these Python scripts are publicly available to download from `https://github.com/grbails`.

in automatically coding (td)-deletion is largely attributable to cases where the deletion process *has* applied, which suggests that FAVE struggles to identify presence of the /t/ or /d/. Although there is no widely-accepted convention for interpreting Cohen's Kappa, the classification suggested by Landis and Koch (1977) would describe agreement in (h) as 'substantial,' and agreement in (td) and (th) as 'moderate.' Conversely, inter-transcriber agreement is 'almost perfect' for (h) and (th), and 'substantial' for (td).

Attention should be drawn to the fact that the inter-transcriber agreement rate was also lowest for non-application of (td)-deletion, which suggests that these patterns may simply be a reflection of the difficulty in coding this variable more generally; that is, humans also show higher rates of disagreement in cases where /t/ or /d/ are supposedly present in word-final clusters.

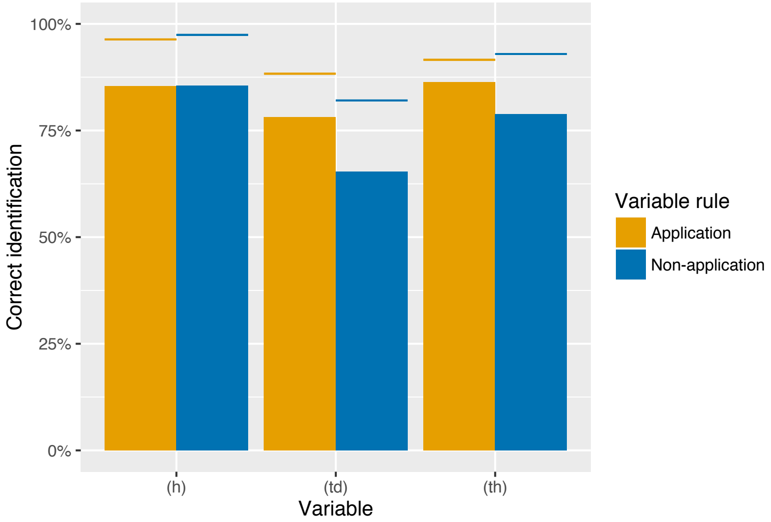

Figure 2: Accuracy of FAVE's automatic coding by variable (horizontal line indicates inter-transcriber agreement rate).

		Human								
		(h)			**(td)**			**(th)**		
			ø	[h]		ø	[t, d]		[f, v]	[θ, ð]
FAVE		ø	47 86%	28 14%	[ø]	107 78%	54 35%	[f, v]	82 86%	57 21%
		[h]	8 14%	166 86%	[t, d]	30 22%	102 65%	[θ, ð]	13 14%	212 79%
FAVE accuracy		85.54% κ = 0.63			71.33% κ = 0.43			80.77% κ = 0.57		
Inter-analyst agreement		97.19% κ = 0.91			84.98% κ = 0.70			92.58% κ = 0.81		

Table 2: Agreement matrix between FAVE and human coding. Results are split into true positives (correct identification of application; top-left cell), true negatives (correct identification of non-application; bottom-right cell), false positives (incorrect claims of application; top-right cell), and false negatives (incorrect claims of non-application; bottom-left cell).

4.2 Effect of Voicing

Given that two of these processes apply to multiple segments in a laryngeal contrast, it is also important to consider what possible effect this may have on the overall rates of variant discrimination success. Both /t/ and /d/ are of course subject to (td)-deletion, while (th)-fronting can apply to both voiced /ð/ as well as voiceless /θ/, and it is possible that acoustic differences between these segments may influence how easily they are perceived, by both automated *and* manual procedures. This would be further problematic in that the ratio of voiced to voiceless tokens is not equal in this data set, which could be skewing the overall accuracy rates (204 tokens of /t/, cf. 71 tokens of /d/; 90 tokens of /θ/, cf. 235 tokens of /ð/).

Figure 3 splits the results for (td) and (th) in this way, and this more nuanced look at the pattern of FAVE's errors reveals an interesting trend; for both variables, the lowest accuracy rates are found for non-application of the variable rule on the *voiceless* segments in each laryngeal contrast. Once again, it should be noted that there was comparably low agreement amongst human transcribers, at least for proposed cases of /θ/-presence (i.e., non-application of the fronting rule).

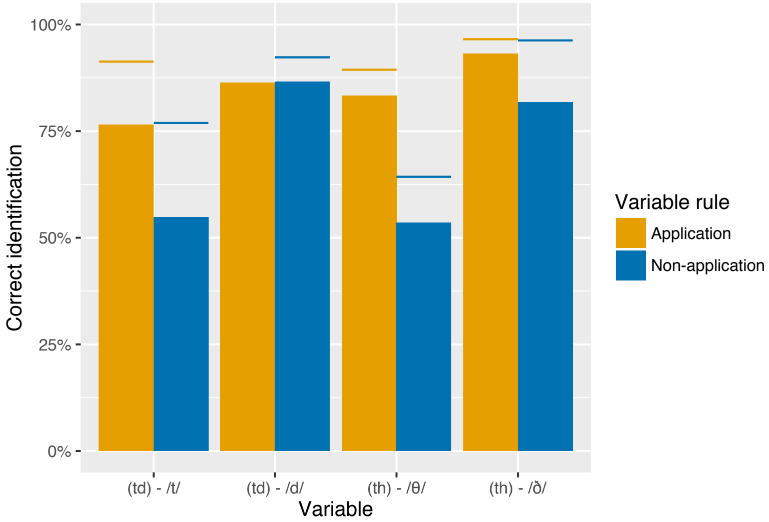

Figure 3: Accuracy of FAVE's automatic coding by voicing (horizontal line indicates inter-transcriber agreement rate).

4.3 Effect of Speech Rate

A third factor that should be considered is the effect of speech rate, a prosodic factor that has been found to influence a range of reductive phenomena, including (td)-deletion (e.g., Guy 1980, Fosler-Lussier and Morgan 1999, Jurafsky et al. 2001). Given that (h)-dropping is also a clear case of lenition, one may expect speech rate to influence automatic variant discrimination in the following ways:

(a) Increased speech rate leads to higher application of the (td) and (h) deletion rules, which could influence overall coding accuracy if there are biases towards identifying presence/absence of particular segments.

(b) Even if the segments avoid deletion, increased speech rate could lead to greater acoustic differences between their realisations and those contained within the pre-trained acoustic models; this could be increasingly problematic if these segments have undergone some gradient reduction and are somehow less acoustically distinct.

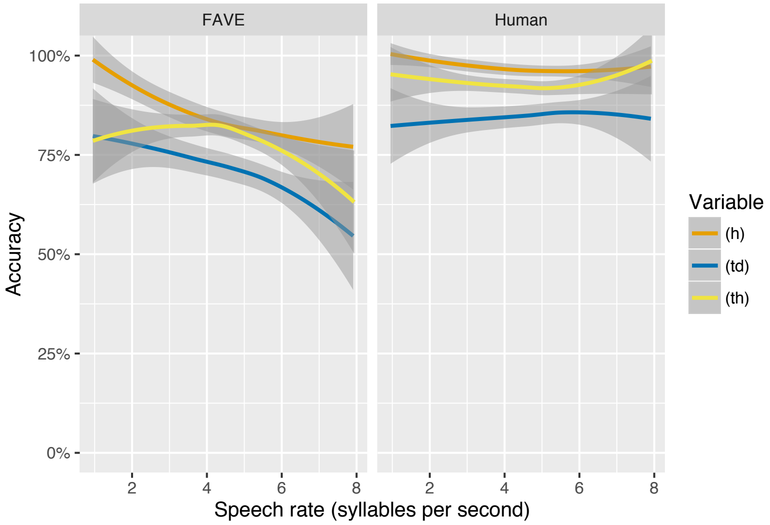

Figure 4: The effect of speech rate on FAVE's accuracy and the level of inter-transcriber agreement (using LOESS-fitted smoothing curves with 68% confidence intervals).

Figure 4 illustrates the detrimental effect increased speaking rates have on FAVE's automated variant coding, where accuracy rates are clearly lower in the faster speech rates; crucially, no such effect is evident for the agreement between human transcribers.

4.4 Logistic Regression

Logistic regression was carried out in R separately for each of the three variables under study, where FAVE's accuracy is the dependent variable and the predictors are application of the process (1 = application, 0 = non-application), voicing (1 = voiced, 0 = voiceless), and speech rate (a continuous measure in syllables per second). The results are given in Table 3, largely confirming the statistical significance of the effects discussed in this section: an inhibiting effect of speech rate for both (h) and (td), as well as a higher probability of success for (td) and (th) when the process *has* applied, particularly to voiceless segments (with voicing being the strongest predictor in both cases).

(h)	Estimate	Std. Error	z-value	*p*
(Intercept)	3.4867	0.6406	5.443	< 0.001***
application	0.4435	0.4574	0.970	0.332
sylls.per.s	-0.3196	0.1187	-2.692	0.007**
(td)	Estimate	Std. Error	z-value	*p*
(Intercept)	1.1194	0.5087	2.201	0.028*
application	1.0848	0.3024	3.587	< 0.001***
voice	1.6753	0.4543	3.688	< 0.001***
sylls.per.s	-0.2457	0.1234	-1.992	0.046*
(th)	Estimate	Std. Error	z-value	*p*
(Intercept)	0.4103	0.6055	0.678	0.498
application	1.2550	0.4898	2.562	0.01*
voice	1.3743	0.4158	3.305	0.001***
sylls.per.s	-0.0984	0.1024	-0.961	0.337

Table 3: Results of logistic regression for each variable rule.

5 Discussion

Overall the results from Section 4 appear promising. However, the aim of this paper was not simply to provide some quantification of FAVE's success in automated variable coding, but to shed light on where exactly its performance suffers, and why this might be. As such, there are three major issues to address here: the poor performance in determining non-application of voiceless (td)-deletion, and likewise for non-application of voiceless (th)-fronting, as well as the problem of speech rate.

The major disparity in performance between identifying presence of /t/ and presence of /d/ in (td) tokens suggests that there is something more acoustically salient in the latter that makes these segments easier to detect; although this would need to be corroborated with acoustic evidence, it is likely that these /t/s in word-final consonant clusters are simply too subtle for the aligner to reliably detect as being meaningful and not just noise or silence.

Cases where /θ/ has not undergone fronting, however, are different in that the distinction is between two segments and not between presence or absence of a single segment. Crucially, this distinction between /θ/ and /f/ is known to be extremely subtle; Jongman et al. (1998) investigate the English fricative system along a number of acoustic dimensions with the aim of finding some correlation with place of articulation, but find that these acoustic parameters only serve to distinguish sibilants from non-sibilants. There is evidence to suggest that, due to this acoustic similarity between /θ/ and /f/, listeners are extremely poor at distinguishing them in perceptual experiments (Heinz and Stevens 1961), which would also account for why inter-transcriber agreement was so low for this variable.

However, the asymmetry in voiceless (th)-identification should not be overlooked; Figure 3 highlighted low accuracy rates on non-fronted tokens of /θ/, but a surprisingly high level of agreement for tokens deemed to be fronted. This suggests a perceptual bias in favour of /f/, which has been reported in earlier work (see the confusion matrices in Miller and Nicely 1955). Interestingly, Donegan and Nathan (2015:446) also note that this perceptual bias is reflected by the typological observation that diachronic changes of /θ/ > /f/ are much more common cross-linguistically than changes from /f/ > /θ/.

Section 4.3 established the relationship between speech rate and the accuracy of FAVE's variant discrimination, which are found to be negatively correlated. This is particularly interesting given that two of these variables, being lenition processes, have higher rates of application in faster speech rates, and that this study finds FAVE performing better when these variable rules *have* applied. The fact that this pattern appears, then, makes the effect of speech rate all the more striking, and particularly worrying given the tendency for fast speech rates in conversational discourse and narratives of personal experience, which are so often the object of study in sociolinguistics.

Although this effect may arise due to some acoustic quality of the segments in fast speech rates, it is worth remembering that the alignment process itself introduces an extra source of potential error; if automatic speech segmentation suffers in faster dialogue (in terms of boundary placement accuracy), then it naturally follows that variant discrimination will also be less reliable.

FAVE's sensitivity to speech rate is arguably more problematic than the other issues raised here because the agreement rate between human transcribers shows no such patterning. In contrast to this, the fact that FAVE's variant discrimination is less reliable for tokens of voiceless /t/ and /θ/ that have not undergone their respective processes of deletion and fronting is somewhat *less* worrying, given that agreement among human transcribers was also at its lowest here.

6 Conclusion

This study aimed to evaluate the viability of exploiting an existing methodological tool in novel ways in order to automate token coding for future studies in the variationist paradigm. It has achieved this by quantifying the degree of error that is introduced through using forced alignment in this way and, more importantly, by locating the exact source of these errors. Reassuringly, FAVE shows the lowest accuracy rates for tokens where inter-transcriber agreement was also at its lowest; other than its sensitivity to rate of speech, FAVE's variant discrimination shows no systematic patterning of errors that is absent among manual transcribers with experience of sociophonetics. Furthermore, when the data is filtered down to tokens with complete agreement among manual transcribers, FAVE's overall accuracy increases from 80.9% to 94.2%; clearly, when the surface variant is fairly salient, FAVE is remarkably reliable.

Whether the accuracy rates reported here are high enough for this to be considered a viable methodology is largely a personal choice in terms of what constitutes an ‘acceptable’ level of noise in the data, and is likely to be dependent on the exact research questions being asked; like many processes of automation, it is a trade-off between accuracy and efficiency. However, given the exponential increase in the amount of data sociolinguistic studies are now utilising as this move towards ‘big data’ gathers pace, it may well be a trade-off worth making.

6.1 Thoughts for Future Improvements

There are two clear areas for future work here: firstly, conducting more rigorous testing of variant discrimination in forced alignment to get a clearer idea of its shortcomings, and secondly, improving the forced alignment software itself to combat some of these issues.

With regards to the first point, it would be interesting to investigate this data set, or comparable ones, along fine-grained phonetic dimensions before testing FAVE's automated coding; it could be the case that there are correlations between certain acoustic parameters and FAVE's success in identifying the surface variant. That is, do we find lower accuracy for tokens that are somehow less salient along some phonetic dimension, or for tokens that are simply poor exemplars of a particular segment?

There are a number of ways in which forced alignment software could be improved for the purposes explored in this paper. The ability to train dialect- or even speaker-specific acoustic models, already offered in alternative aligners such as PLA, could conceivably produce more reliable results because the acoustic differences between the acoustic models and the observed speech stimuli would be minimised. It should be noted again that FAVE's models are trained on General American varieties, which makes the accuracy rates reported here for British test stimuli all the more impressive.

Composite models could also be trained for certain sequences of segments, in the manner of Yuan and Liberman 2011 where (ing) is the variable under study and the models for /ɪ/, /ŋ/, and /n/ are collapsed into composite /ɪŋ/-/ɪn/ models; this was motivated by acoustic and perceptual evidence indicating that the greatest acoustic cue for the *-in*/*-ing* alternation actually lies on the vowel, and not the following nasal. The use of composite models, or triphone models that capture sequences of three segments at a time rather than independent monophone models, could account for coarticulatory behaviour and context-specific realisations in this way.

References

Agirrezabal, Manex, Jeffrey Heinz, Mans Hulden, and Bertol Arrieta. 2014. Assigning stress to out-of-vocabulary words: Three approaches. In *Proceedings of the International Conference of Artificial Intelligence (ICAI 2014)*, 105–110.

Baranowski, Maciej, and Danielle Turton. 2015. Manchester English. In *Researching Northern English*, ed. R. Hickey, 293–316. Amsterdam: John Benjamins.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22:249–254.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

Donegan, Patricia J., and Geoffrey S. Nathan. 2015. Natural phonology and sound change. In *The Oxford Handbook of Historical Phonology*, ed. P. Honeybone and J.C. Salmons, 431–449. Oxford: Oxford University Press.

Fosler-Lussier, Eric, and Nelson Morgan. 1999. Effects of speaking rate and word frequency on pronunciation in conversational speech. *Speech Communication* 29:137–158.

Fruehwald, Josef. 2015. Big data and sociolinguistics. Paper presented at the Penn Linguistics Conference 39, University of Pennsylvania.

Ghahramani, Zoubin. 2001. An introduction to Hidden Markov Models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15:9–42.

Gorman, Kyle, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39:192–193.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating Language in Time and Space*, ed. W. Labov, 1–36. New York: Academic Press.

Heinz, John M., and Kenneth N. Stevens. 1961. On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America* 33:589–596.

Jongman, Allard, Ratree Wayland, and Serena Wong. 1998. Acoustic characteristics of English fricatives: I. Static cues. *Working Papers of the Cornell Phonetics Laboratory* 12:195–205.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the Emergence of Linguistic Structure*, ed. J. Bybee and P. Hopper, 229–254. Amsterdam: John Benjamins.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89:30–65.

Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–174.

MacKenzie, Laurel, and Danielle Turton. 2013. Crossing the pond: Extending automatic alignment techniques to British English dialect data. Paper presented at NWAV 42, University of Pittsburgh/Carnegie Mellon University.

Miller, George A., and Patricia E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27:338–352.

Milne, Peter. 2014. The Variable Pronunciations of Word-Final Consonant Clusters in a Force Aligned Corpus of Spoken French. Doctoral dissertation, University of Ottawa.

Reddy, Sravana K., and James N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1:15–28.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. Available at: `http://fave.ling.upenn.edu`.

Yuan, Jiahong, and Mark Liberman. 2011. Automatic detection of "g-dropping" in American English using forced alignment. In *Proceedings of 2011 IEEE International Conference of Acoustics, Speech and Signal Processing*, 490–493.

Yuan, Jiahong, Neville Ryant, Mark Liberman, Andreas Stolcke, Vikramjit Mitra, and Wen Wang. 2013. Automatic phonetic segmentation using boundary models. In *Proceedings of Interspeech 2013*, ed. F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, 2306–2310.

Department of Linguistics and English Language
School of Arts, Languages and Cultures
University of Manchester
Manchester, UK
M13 9PL
george.bailey@manchester.ac.uk