Sound Change or Experimental Artifact?: A study on the impact of data preparation on measuring sound change.

Catherine I. Watson and Zoe E. Evans

University of Auckland, University of Otago

c.watson@auckland.ac.nz, zoe.evans@otago.ac.nz

Abstract

Forced alignment systems are commonly used to process large amounts of data in socio-phonetic studies. We compare how two of these systems perform against manual segmentation and phonetic labelling in a database of New Zealand English. The results indicate predictable variations in terms of the relative sizes of the vowel spaces, but also suggest the need to be cautious in interpreting small phonetic variations, as these may be the result of the method used to segment and label the data.
Index Terms: sound change, forced alignment, artifact, vowel

1. Introduction

Socio-phonetic studies have been employed over many years to understand the factors that result in speech variability, such as region, age, gender, culture, language exposure, and so forth. The observed variations in these studies can often be small, but significant [1]. For sound change studies it is important to ensure that these small changes are indeed indicative of actual change. To improve the robustness of these findings, large amounts of data are gathered, but the task of manually transcribing, segmenting, and labelling these data sets is a daunting one. Consequently, it has become more common in recent years to involve automated systems in the processing of data from sizeable speech corpora [2-4]. There are many automated processes available to speech researchers, performing a range of tasks from audio dictation (e.g.[5]), to forced alignment and phonetic transcription (e.g. FAVE [6], MAUS [7], DARLA[5], LaBB-CAT [8]), through to formant extraction [9,10].

In this study, we focus on the forced alignment systems that are used to automatically segment and label the speech sounds in an audio file. In order to successfully identify phonemes and their boundaries, these systems use appropriate acoustic models and pronunciation dictionaries. Here, we compare two systems: FAVE-align, and MAUS. The Forced Alignment and Vowel Extraction (FAVE) suite provides two services, one of which is an aligner (FAVE-align) which uses a U.S. English (USE) acoustic model and USE pronunciation dictionary to automatically generate a Praat textgrid containing a word and phonetic tier. Although FAVE has been developed for, and arguably performs well for analyses of USE, it has been gaining traction outside the States for use in British English [2, 11], New Zealand English (NZE) [12], Australian English (AE) [13] and Bequia [21]. Since USE acoustic models may not perform as accurately for non-US Englishes, an alternative aligner was selected for comparison. The Munich AUtomatic Segmentation (MAUS) aligner [7] is an online service of the Bavarian Archive for Speech Signals. MAUS offers acoustic models and lexicons for a range of English dialects (and other languages), and therefore may provide more accurate phonetic segmentation and labelling for non-US Englishes.

While these automated systems vastly speed up the processing of speech data, they may not always match the results provided by a more manual analysis [14]. This distinction is important as some recent studies have used these different methodologies to draw comparisons between different data sets [15]. If automatic and manual alignments do not provide comparable data, this may result in the observation of spurious phonetic variation. The importance of using similar methodologies to compare and contrast results from different data sets cannot be overstated. In a recent comparison of two AE corpora [15], substantial differences were discovered between vowel productions. Given that the speakers in the corpora were different ages, were from different locations, and that the vowel formants were extracted using different techniques, it is difficult to draw any helpful conclusions about the observed variability. Similarly, in another analysis of AE [13], it was found that using a forced aligner to segment and label the data resulted in errors that may have influenced the variability of vowel measurements, particularly for vowel durations.

As more and more socio-phonetic studies rely on automated systems for processing large speech corpora, it is important to establish that the data generated by these systems does not differ substantially from the data generated by hand labelling. In manual segmentation and labelling, researchers carefully select tokens that have minimal phonetic reduction (i.e. from lexically stressed syllables of pitch-accented words), and from environments with minimal coarticulatory effects. Given this, and the fact that forced-alignment systems are not yet capable of identifying pitch-accented words, we would expect that a phonetic analysis using forced alignment would show greater reduction along all dimensions compared with an analysis based on hand-labelled data. If forced alignment systems are to be relied on, we would also hope to see the same patterns of phonetic change, albeit within a reduced vowel space, as are observed in manually labelled data. To this end, in the current study we provide three separate analyses of a single data set. This data was originally presented in [16], and described vowel change over time in three speakers of NZE. The data was manually segmented and labelled. In the present study, we will re-present those results alongside results from the same data set where the segmentation and labelling was carried out automatically by the FAVE aligner, and MAUS. Our aim is to establish to what degree the use of automated aligners and labellers may impact on results.

2. Method

2.1. Database

The speakers were all prominent New Zealand men (5 in total) for which there were multiple radio recordings available [16]. Most of the recordings were interviews. In this study we included 2 more speakers and for each speaker looked at

recordings made mainly in the 1950s (henceforth called the 50s set) and recordings made mainly in the 1980s (henceforth called the 80s set). The recordings were digitized to a 16 bit number, at a sampling rate of 20 kHz, and stored as WAV files. The speaker and recording details are in Table 1. Only monophthongs were used in the current analysis.

Table 1: The Speakers' date of birth and recordings.

	A	B	C	D	E
Date of Birth	1919	1901	1916	1935	1935
Recordings:19xx	54,92	54,82	55,85	60,86	60,84

2.2. Hand-Labelled Data Preparation

All the data was hand labelled at the phonetic level using EMU labeller [10]. Only the vowels from prosodically accented words were selected for analysis. The formant were automatically tracked in ESPS/WAVES+ (12th order LPC analysis, cosine window, 49 ms frame size, 5 ms frame shift). All formant data was checked and corrections were made by hand if necessary. The first formant (F1) and second formant (F2) values at the vowel target were obtained. The vowel targets were identified by hand. See [17] for the guidelines and [16] for the phonetic contexts. This resulted in 3127 monophthongs for the hand-labelled corpus (henceforth called the HAND data).

2.3. System-Labelled Data Preparation

Transcripts were generated for each of the segmented WAV files. The transcripts and their associated WAVs were then passed to two automated segmentation and labelling systems: MAUS [7], and FAVE (Forced Alignment and Vowel Extraction) [6]. The MAUS and FAVE systems returned PRAAT textgrids containing phonetic tiers. The vowel formants were calculated using the same package as the HAND data, with the same analysis settings. The first and second formant values used in this study were extracted from the vowel midpoint but no checking was performed

2.3.1. MAUS and FAVE Data Preparation

The MAUS alignment system allows the user to select from a number of English dialects. In this study, the NZE option was selected, and PRAAT textgrids were generated. Three levels were identified in each textgrid; word, phoneme, and phonetic. At the phonetic level some vowels were identified as schwa. This is determined both by the lexicon and acoustic segmentation process. This resulted in 11602 monophthongs for the MAUS labelled corpus (hence forth called MAUS data).

The textgrids generated by FAVE align contained two tiers: a word and phonetic level. The FAVE phonetic transcription is based on US English pronunciation, which may result in phonetic confusions in non USE dialects. All words in the transcripts were either already within FAVE's internal lexicon, or needed to be added to a separate dictionary. FAVE-align assigns all vowels primary, secondary, or no lexical stress. Only vowel tokens marked with primary lexical stress were included in this study. This resulted in 9612 for the FAVE labelled corpus (henceforth called FAVE data).

2.4. Vowel Space Size and Sound Change Measurements

Analysis of the formant data was in R [18]. Two measures were used to compare the data preparation approaches: one for the vowel space size and one for the movement of vowels. To calculate the vowel space size for each speaker we estimated its width by finding the difference in F2 (in Bark) between FLEECE, and THOUGHT centroids (see figure 1), and estimated its height via the difference in F1 (in Bark) between FLEECE and START centroids (see figure 1).

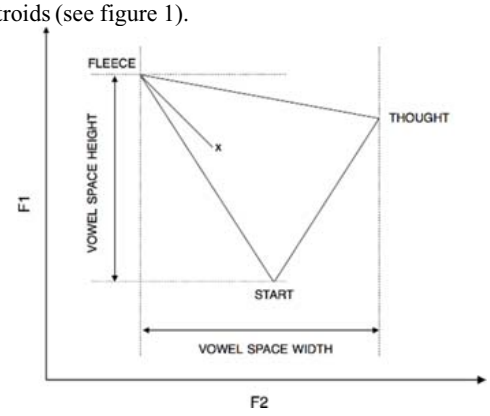

Figure 1 Stylised vowel space with point vowels

To measure the extent of the vowel shifts we used the Vowel Space Movement (VSM) measure [19]. Four vowel shifts are investigated: DRESS raising, KIT falling, KIT retracting, and GOOSE fronting. All have occurred over the timespan of the data (see e.g. [16,19]). The movements the vowel space are be measured by calculating

$$VSM(x)\frac{EU(i{:},x)}{EU(i{:},pointvowel)} \quad (1),$$

where x is the vowel of interest, , $EU(i{:},x)$ is the Euclidean distance between the FLEECE vowel and x in the F1/F2 vowel space (see Figure 1), *pointvowel* is either START or THOUGHT, depending on the direction of the movement, and $EU(i{:},pointvowel)$ is the Euclidean distance between FLEECE and *pointvowel* (see Figure 1). For movements up at the front of the vowel space the point vowel was START, and the x was DRESS or KIT. For movements along the top of the vowel space the point vowel was THOUGHT and x was either KIT or GOOSE. The Euclidean distances are calculated in Bark. The VSM results in a value between 0 and 1. The smaller the value, the closer the vowel in question (x) is to the FLEECE vowel.

3. Results

3.1. Size of the Vowel Space

Figure 2 (placed at end of the paper) contains three F1/F2 plots which have the centroids of each of the NZE monophthongs for the recordings from the 50s set and 80s set from Speaker A for each of the three different data preparation approaches. The 50s data is in the dark hue, and the 80s data is in the light hue. It can be seen the vowel spaces for all three approaches looks reasonably similar. But, as expected, the vowel space from the HAND data has a greater range in both F1 and F2 than that from the MAUS and FAVE data. Vowel shifts for each speaker between the 50s and 80s sets are apparent for all three approaches, however the extent of these changes potentially differs, for example the differences in the TRAP, NURSE and GOOSE centroids. For the FAVE data there were issues with LOT, START, and THOUGHT labels, due to differences between NZE and USE. We were able to correct some of these checking the labels of the corresponding vowels in the MAUS data.

To investigate the difference in vowel space size for all speakers we did two repeated measures ANOVA with vowel space width and height being the dependent variables, respectively and data preparation method, and year of recording being the within-subject factors. Sphericity of the within-subject factors was detected and Greenhouse–Geisser

adjustments were made to the p values. For both the height and width measures there were significant differences due to data preparation approaches (Width: $F(2,8)=44.9$ $p_{GG}<0.01$, Height: $F(2,8)=40.4$ $p_{GG}<0.01$) but nothing else. Since the year of recordings (hence age of the speaker) was not a significant factor for vowel space size combined the data from the two different years. A paired t-test (corrected for multiple comparisons) showed that the width and height from the HAND data was significantly greater at the $p<0.01$ level than that from the MAUS data (Width: $t(9)=9.5$, Height: $t(9)=4.9$), and the FAVE data (Width: $t(9)=8.7$, Height: $t(9)=9.5$). There were no significant differences in the dimensions for the MAUS and FAVE data. Table 2 is the mean width and height of the vowel space for the 5 different speakers. The vowel space size for the HAND data is greater than for the MAUS and FAVE. Next we investigated whether the extent of the vowel shift is the same in all three approaches. We used the vowel space measure (VSM) to establish whether the movement of the KIT, DRESS and GOOSE vowels is the same, regardless of the data preparation approach.

Table 2.The width and height of the vowel space in Bark by Speaker the three data preparation approaches.

	A (Bark)	**B (Bark)**	**C (Bark)**	**D (Bark)**	**E (Bark)**
HAND	6.0, 4.0	6.9, 4.0	5.9, 3.3	5.7, 2.3	4.9, 3.2
MAUS	4.7, 3.3	5.1, 3.2	4.3, 2.5	4.1, 2.0	4.1, 2.3
FAVE	4.7, 3.1	5.0, 3.2	4.6, 2.4	4.2, 1.3	4.4, 1.9

3.2. Measuring Vowel Movements

In presenting these results we first focus on the HAND data. Figure 3 shows the movement of DRESS over time, relative to the front of the vowel space. Looking at the HAND data, for all speakers except D, the VSM value for the 80s data is less than for the 50s data, indicating it is closer to the FLEECE vowel and therefore demonstrates the DRESS raising in this diachronic data. No DRESS raising was noted for Speaker D, but the low VSM 1950s value for the HAND data show his DRESS was already very raised. Figure 4 shows the movement of KIT over time relative to the top of the vowel space. For speakers B,D, and E the VSM values for the 80s data is greater than that for the 50s data, which means it is further away from FLEECE, i.e. retracted. Figure 5 shows the movement of GOOSE over time, relative to the top of the vowel space. With the HAND data it can be seen that for Speakers A and C the VSM values for the 80s data are less than that for the 50s data, indicating it is closer to FLEECE, and therefore demonstrating the well-known GOOSE fronting. Due to lack of space the plot for KIT lowering is not included.

When looking at the VSM values from the MAUS and FAVE approaches, we found the VSMs for the three different methods are all highly and significant correlated, see Table 3. However for any speaker-vowel combination the VSM measures for the vowel movements are rarely similar for the three data approaches and the magnitude of the change between the 50s and 80s data differs across the three methodologies, regardless of speaker.

Figure 6 gives the vowel shift results for each speaker, where red hue indicates the VSM for the 80s data is less than the 50s data, and a green hue indicates it is greater than. The intensity of the colour indicates the magnitude of the difference of the VSM between the 50s and 80s. The greater the difference the darker the hue. It would be expected that the cells for the DRESS rising and GOOSE fronting would all be a reddish hue (VSM being closer to 0 for the 80s data), and the cells for KIT falling and retracting to have a greenish hue (VSM being closer to 1 for the 80s data). There are 20 sound changes investigated (5 speakers X 4 vowel movements). The expected changes happened 12/20 possible times for the HAND and FAVE data, and 15/20 for the MAUS data. In addition the three measures were in complete agreement only 10/20 times (as indicated by the similar hue in all three cells for the speaker), and there were a further 5 where either FAVE or MAUS were in the same direction as the HAND data. However, even when there is agreement in the direction of the vowel change between the three measures, the varying intensity of the cell colour indicates that the three methods rarely agree on the *extent* of the sound change.

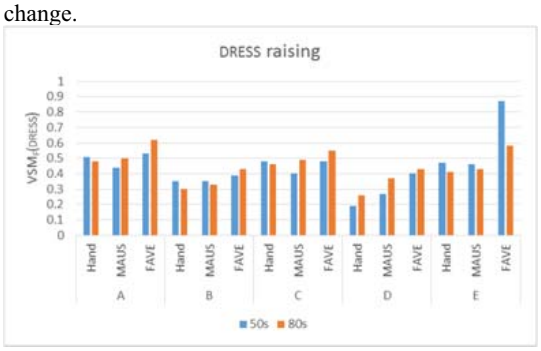

Figure 3: The VSM of DRESS raising in the front of the vowel space for the three different approaches.

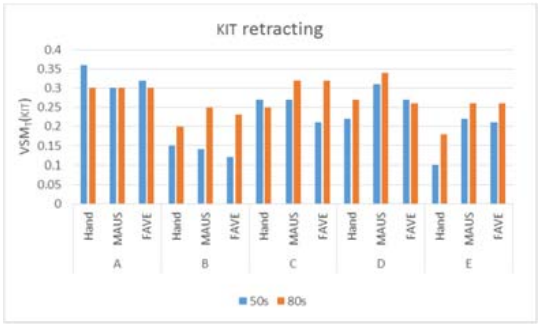

Figure 4 The VSM of KIT retracting along top of the vowel space for the three different approaches.

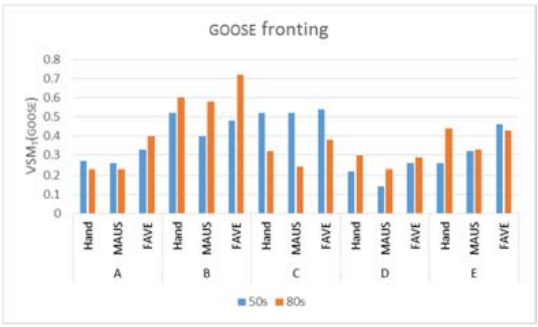

Figure 5: The VSM of GOOSE fronting along the top of the vowel space for the three different approaches.

4. Discussion and Conclusions

In this study we have re-investigated the vowel shift observed in diachronic data of 3 New Zealand English speakers, and included 2 more speakers. We compared three different data preparation approaches: hand labelling (which was used in [16]), labelling via MAUS, and labelling via FAVE.

The three data preparation methods took data from exactly the same set of recordings. However since only vowel tokens

with sentence stress (and word stress) were selected for the HAND data, this labelling approach had considerably fewer tokens for analysis compared to the other two. In those two cases only unstressed vowels were excluded from analysis. This yielded 3-4 times more tokens than in the HAND instance, but also resulted in significantly reduced vowel spaces. Without some form of natural language processing it is not possible to automatically select such tokens from MAUS and FAVE. We used the VSM to measure the extent of the vowel shifts. Our studies found that all three methods give highly significant correlated VSMs. However there is differences in the extent of the measured vowel shift. The three data approaches are only in complete agreement of the direction of movement ~50% of the time, and a further 25% of the time either FAVE or MAUS data agrees with the HAND data.

Table 4: Correlations of the VSM for data preparation approaches

	HAND VS MAUS: R,P	**HAND VS FAVE: R,P**	**MAUS VS FAVE: R,P**
50s	0.84,p<<0.01	0.72,p<<0.01	0.80,p<<0.01
80s	0.79,p<<0.01	0.89,p<<0.01	0.78,p<<0.01

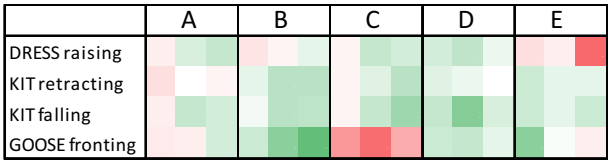

Figure 6 A heat plot of the vowel shift results for the HAND, MAUS, and FAVE data: red indicates 80s VSM < 50s VSM, and green indicates 80s VSM > 50s VSM.

This study has focused mainly on the comparison between the three data labelling approaches. However it is worthwhile to note that in terms of measuring sound change, perhaps this diachronic data is not the best data set for this. In contrast to other studies (eg. [20]) the differences in the speech of the 5 individuals over time is not overwhelming. In conclusion whilst there is a general agreement across the three methods in the direction of the sound change but there is little agreement on the extent of the sound change.

5. Acknowledgements

Thanks to Sallyanne Palethorpe for help in data preparation, thanks to Miriam Meyerhoff for feedback on the paper, and thanks to Aldo Cortesi for his python assistance.

6. References

1. Stuart-Smith, J., Jose, B., Rathcke, T., Macdonald, R. &Lawson, E., 'Changing sounds in a changing city: An acoustic phonetic investigation of real-time change across a century of Glaswegian', In E. Moore and C. Montgomery (eds), A Sense of Place, Cambridge: Cambridge University Press, (in press)
2. Cham, B.Q., "Sixty years of speech: A study of language change in adulthood", Lifespans & Styles 2(1): 17-26, 2016
3. Stanford, J., Severance, N. & Baclawski, K.,"Multiple vectors of unidirectional dialect change in eastern New England". Language Variation and Change 26:103-140, 2014.
4. Labov, W., Rosenfelder, I., & Fruehwald, J.."One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis". Language, 89(1), 30-65 , 2013.
5. Reddy, S. and Stanford, J., "Toward completely automated vowel extraction: Introducing DARLA." Linguistics Vanguard. 2015
6. Rosenfelder, I., Fruehwald J., Evanini K, & Y. Jiahong, FAVE (forced alignment and vowel extraction) program suite. http://fave.ling.upenn.edu, 2011
7. Kisler, T, Schiel, F. & Sloetjes, H., "Signal Processing via web services: the use case WebMAUS" in Proceedings Digital Humanities, Hamburg Germany, 30-34, 2012
8. Fromont, R. & Hay, J. "LaBB-CAT: An annotation store.": Australasian Language Technology Workshop (ALTA), 4-6 Dec 2012. In Proceedings 10:: 113-117, Dunedin, 2012.
9. Boersma P. & Weenink D,: Praat: doing phonetics by computer. Version 5.3.51, http://www.praat.org/ ,2013.
10. Cassidy, S. and J. Harrington, "Multi-level annotation in the Emu speech database management system", Speech Comm., 33, 61-77, 2001
11. MacKenzie, L., & Turton, D.. "Crossing the pond: Extending automatic alignment techniques to British English dialect data."New Ways of Analyzing Variation, 42, 2013.
12. Hazenburg, E. "Gendering vowels: Sociophonetics and community affiliation". NZ Ling. Soc. Conf., Dec, Dunedin. 2015
13. Nguyen, N. & Shaw, J.A., "Why the SQUARE vowel is the most variable in Sydney", in Proc. 15th Aust. Int. SST 36-39, 2014
14. Severance, N., Evanini, K. & Dinkin, A., "Examining the reliability of automated vowel analyses using FAVE." Presented at NoWPhon 2, Eugene, Ore, 2016
15. Elvin, J. & Escudero, P., "Comparing acoustic analyses of Australian English vowels from Sydney: Cox (2006) versus AusTalk" in Proc. of the 15th Aust. Int. SST Conf., 44-48, 2014
16. Watson, C., Palethorpe, S., & Harrington, J. (2004). Capturing the vowel change In New Zealand English over a thirty year period via a diachronic study. In Proc. 10th Aust. Int. SST, 201-206, 2004.
17. C. I. Watson, J. Harrington & Z. Evans "An acoustic comparison between New Zealand, and Australian English vowels", Australian Journal of Linguistics, 18 (2), 185-207, 1998
18. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2015.
19. Watson, C. I., Maclagan, M., King, J., Harlow, R., & Keegan, P. J. "Sound change in Māori and the Influence of New Zealand English". Journal of the International Phonetics Association DOI 10.1017/S0025100316000025, 2016.
20. Harrington, J., Palethorpe, S., & Watson, C. I., "Does the Queen speak the Queen's English?, Nature, 408(6815), 927-928, 2000
21. Walker, J. A & M. Meyerhoff "Mergers of the Caribbean: Low-back vowels in Bequia" Paper presented at Change and Variation in Canada 8. Queen's University, Ontario., 2014.

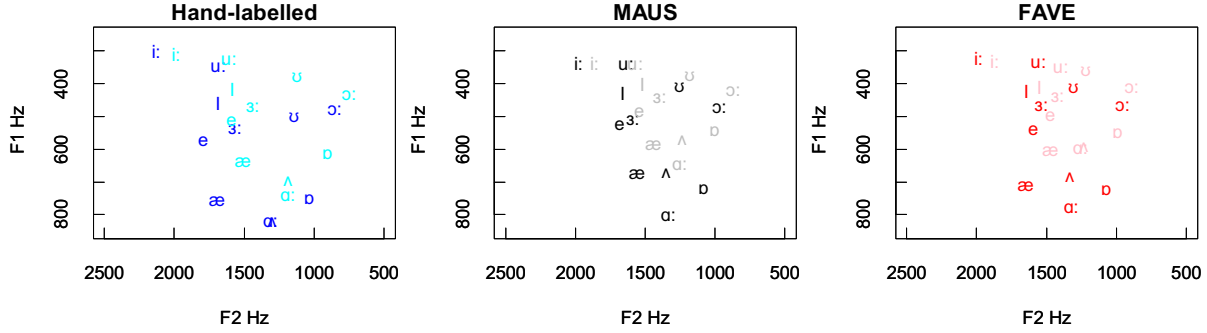

Figure 2: The vowel space of Speaker A from the 50s data (dark) and 80s data (light) for the three different data approaches