A real-time study of plosives in Glaswegian using an automatic measurement algorithm: change or age-grading?[i]

Jane Stuart-Smith, Tamara Rathcke*, Morgan Sonderegger+, and Rachel Macdonald

English Language/Glasgow University Laboratory of Phonetics (GULP), Glasgow University

**English Language and Linguistics, University of Kent*

+Linguistics, McGill University

Abstract

This paper presents a collaborative study of variation and potential change in the voicing contrast in Scottish English plosives, analysed in recordings from twelve vernacular female speakers of different generations made in the 1970s and the 2000s in Glasgow. We adapted an existing automatic measurement algorithm for predicting Voice Onset Time (VOT) originally developed for voiceless stops, for the analysis of voiced and voiceless plosives in casual sociolinguistic speech recordings of different kinds. Our semi-automatic method, which involved quick manual coding of automatically-generated positive VOT predictions, resulted in correct or close to correct measures for two-thirds of our data, and allowed us to process a very large number of tokens very quickly, especially for voiceless stops. The VOT results themselves indicate that the voicing contrast is being maintained, but suggest that a change in the phonetic realization of the stops may have been in progress since the middle of the 20th century, specifically a lengthening of aspiration for /p/ and /t/, and a trend to a longer release phase in their voiced counterparts.

1. Introduction

Shifts in the realization of the voicing contrast for stops, 'Grimm's Law', constitutes one of the most notable aspects of the Germanic language group in its development from Proto-Indo-European. The subsequent fate of plosives in the recent history of English seems to have been less eventful, at least with respect to voicing and aspiration. The contrast is now typically between phonetically partially voiced stops and voiceless aspirated stops . But variation – especially in the durational measure of Voice Onset Time (VOT) used to capture the contrast phonetically – has also been observed according to phonetic factors such as place of articulation (Lisker and Abramson 1967), regional accent (Wells 1982; Scobbie 2006), and social and individual factors, including those to do with physiological age (Allen et al 2003). The question remains as to whether gradient change may be underway in varieties of English.

This paper reports the preliminary results from a collaborative study to consider the voicing contrast in Scottish English over real time, and at the same time, to try to find a faster way of achieving this kind of analysis than by using hand measurement.

2. The voicing contrast in Scottish English

The realization of the voicing contrast in Scottish English is thought to be different from that of Southern British English, particularly in that voiceless plosives show shorter aspiration (Wells 1982). But there is also variation within Scottish English, which ranges from Scots vernacular to Scottish Standard English (SSE). Scots is described as showing unaspirated voiceless stops alongside aspirated stops in SSE (Johnston 1997: 505; Scobbie 2006: 374). There are indications that Scots may be changing, given that Johnston's statement is based on descriptions from the 1930s and 1940s, and that he observes elsewhere that SSE aspirated variants are spreading into Scots.

There are few empirical studies of the voicing contrast in either Scots or Scottish Standard English. Scobbie's (2006) brief summary of the results available to date concludes that the voicing lag of SSE is likely to be similar to that of Standard British English, with that of Scots possibly slightly less. Scobbie's results for VOT in read wordlists by 12 young adult speakers in Shetland, with Shetlandic, Scottish, or English parents, show a range of durations, from prevoiced /b/ and short-lag /p/, to short-lag /b/ and longer-lag /p/. The former pattern is more typical of speakers with Shetlandic parents. Scobbie also notes indexical and individual variation in VOT values, and concludes that 'VOT for /p/ and the rate of prevoicing for /b/ are … likely to be sociolinguistic variables' (p.386).

More recently, Docherty et al. (2011) analyzed 4662 tokens for voiced and voiceless stops, hand-segmented from wordlists, from 159 speakers, older and younger, from four locations at both ends of the English/Scottish Border. Younger speakers overall had less aspiration in voiceless stops and less prevoicing in voiced stops, than older speakers. Although this result could reflect a change in progress, the authors preferred an interpretation in terms of age-grading, pointing to cross-linguistic phonetic evidence which shows that in general, younger speakers show longer VOT than older speakers. The results also showed that Scottish speakers at the Eastern end of the Border are less aspirated that those at the Western end; the Eastern Scottish speakers also show proportionately more of other 'Scottish' features, such as rhoticity.

Docherty et al's findings show that longer-lag voiceless plosives have already spread into Scottish English at the West end of the Border. But the resistance in the East confirms that this fine-grained aspect of plosive production is also subject to the subtle kind of sociolinguistic control observed in Shetland by Scobbie.

3. Research question for this study

The voicing contrast in Scottish English, or at least the Scots end of the Scottish English continuum, may be changing. But there has been rather little empirical work on this, and nothing to our knowledge investigating evidence for change. Our main research question is therefore:

- is the voicing contrast in plosives changing in real-time in Scottish English?

Here we consider the contrast in plosives as they occur in spontaneous speech in a Scots variety, Glaswegian vernacular. In order to try to tease apart the different possible factors – physiology or real-time change – we analysed speech from speakers of different ages, recorded at different points in time. We also needed to consider as many tokens as possible, but hand-labelling phases of stops from spontaneous speech is time-consuming. Our second aim is to tackle the methodological challenge presented to variationists, by adapting an automatic VOT algorithm for sociolinguistic research.

4. The Glasgow real-time project

Glasgow dialect continues a variety of West Central Scots, and exists as part of the linguistic repertoire available mainly to working-class speakers in the city. As such it offers a good case for considering whether Johnston's observation of more aspirated stops for voiceless stops also pertains to urban Scots. The results presented here are drawn from a real-time project on Glaswegian, *Sounds of the City*. The electronic real-time spoken corpus of Glaswegian is structured so that it has recordings from the 1970s to 2000s, and from three generations of speakers, male and female, for each decade. We have gathered whatever existing recordings we could find, from oral history interviews, existing sociolinguistic interviews, to broadcast materials. The Glasgow real-time corpus uses LABB-CAT software, and is stored in the form of time-aligned orthographic transcripts with sound files, automatically-created phonemic transcriptions, and preliminary automatic segmentation, carried out using LABB-CAT's HTK routine.

4.1. Sample

We report results from twelve female speakers, elderly (aged between 67-90 years), middle-aged (between 40-55 years), and adolescent (between 10-15 years), recorded in the 1970s and the 2000s, with two per group:

70-O: born 1890s, elderly in 1970s
70-M: born 1920s, middle-aged in 1970s
70-Y: born 1960s, adolescent in 1970s
00-O: born 1920s, elderly in 2000s
00-M: born 1950s, middle-aged in 2000s
00-Y: born 1990s, adolescent in 2000s

The recordings consist of oral history interviews (the older women), sociolinguistic interviews (the middle-aged women and girls from the 1970s) and sociolinguistic recordings made without the interviewer present (the middle-aged women and girls from the 2000s).

4.2. Linguistic variable: /p t k/ and /b d g/

We considered all possible tokens of stressed, syllable-initial, voiceless and voiced stops. The voicing contrast can be characterized acoustically in a number of ways, but it is very usual to consider it in terms of positive and negative VOT (Lisker and Abramson 1967). Here we partially characterize the voicing contrast, since our analysis only considers positive VOT. Whilst positive VOT for voiceless and voiced stops allows us some insight into the contrast, our description is limited to the release phase: sometimes positive VOT will reflect only the burst of an otherwise voiced stop, sometimes it will capture burst and some aspiration. We are currently working on automatic methods of capturing closure duration and proportion of voicing during closure.

4.3. An automatic algorithm for measuring VOT

4.3.1. Algorithm description

The VOT measurement algorithm uses a set of hand-labelled VOT measures to train a classifier which then can be used to predict positive VOT measures for new files (Sonderegger and Keshet 2012). It takes as input speech segments, each containing a VOT region, delimited by the *burst onset* and the *voicing onset*, labeled t_b and t_v, together called an *onset pair*; the difference between t_b and t_v is the VOT. Conceptually, the algorithm consists of two phases: training and testing. In training, t_b and t_v are known. In testing, the algorithm predicts t_b and t_v.

Phase 1: Training: In this phase, a set of speech segments where the burst and voicing onsets have already been labelled manually and so are known, is used to train a function which takes as input a speech segment containing an (unknown) VOT, and a hypothesized onset pair. The function is a weighted sum of 62 *feature maps*, each of which corresponds to a quantity computed for a given speech segment and hypothesized onset pair. The training data is used to choose a set of weights for the 62 feature maps such that the value of the function for "good" onset pairs is maximally different from its value for "bad" onset pairs. Thus, the function can be thought of as producing a measure of "goodness" for each hypothesized onset pair. Intuitively, the quantity corresponding to each feature map should have high values when the hypothesized onsets are close to the true burst and voicing onsets, and low values otherwise. For example, one of the feature maps is the mean high-frequency energy (>3000 Hz) between t_b and t_v, minus the mean high-frequency energy up to t_b. This quantity is high when t_b and t_v delineate the boundaries of a stop release and are preceded by a closure, as in many stop productions (e.g. Fig. 1, left); it is lower as t_b and t_v are moved away from these locations, as in Fig 1, right. The feature maps quantify acoustic cues from the spectrogram and the speech signal that human annotators use when annotating VOT (for example, the presence of a period of high-frequency energy).

Phase 2: Testing: Following training, the function can be used as a *classifier* to predict VOT for any speech segment containing a VOT region. The classifier returns the VOT corresponding to the best onset pair: the one for which the function takes on its maximum value. This classifier can now be used to predict VOT (and a corresponding onset pair) for any new set of speech segments.

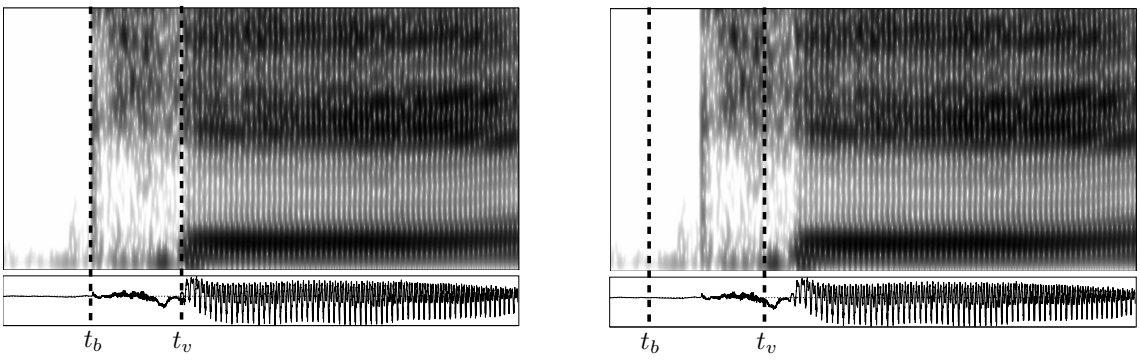

Figure 1 : Speech segment with good (left) and bad (right) hypothesized burst and voicing onsets, t_b and t_v.

4.3.2. Previous results

Sonderegger and Keshet (2012) evaluated the algorithm on word-initial voiceless stops /p t k/ in four different English audio datasets. For each, automatic measurements were assigned by training the classifier on part of the data and using it to assign automatic VOT measurements to the other part. The results were evaluated in several ways:

1. *Raw automatic/manual measurement differences*: Overall, the mean absolute difference between automatic and manual VOT measurements was less than 10 ms for 83.3-93.9% of tokens.
2. *Comparison to inter-transcriber* agreement: For the 3 datasets for which a subset of double-transcribed data was available, the agreement between automatic and manual measurements was nearly as good as the agreement between two human transcribers.
3. *Comparison of induced regression models*: It is important to show that the automatic measurements would give the same results if used in place of manual measurements in a linguistic study. Accordingly, two mixed-effects regression models were built for data from a speech production experiment, using automatically and manually-measured VOTs. The models were extremely similar in terms of the values and significances of terms in the models (correlation between predictions of the two models: $r = 0.992$).

The algorithm's excellent performance was based on a very large amount of training data (400-4000 examples) for each dataset. A further set of experiments showed that the algorithm can be successfully applied to new datasets by first hand-labelling VOTs for a subset of the data (250 examples), training a classifier, then applying it to predict VOTs for the remainder of the data.

Here we used the algorithm to predict positive VOT for both voiceless and voiced stops. There are several reasons that predicting VOTs in this setting could be a challenge. The different recordings are of variable quality, and the speech is often highly casual, showing extensive reduction in comparison to previous test datasets. Perhaps more importantly, the algorithm has not previously been extensively tested on VOT prediction for *voiced* stops, which have much smaller positive VOTs than the voiceless stops.

4.4. Procedure for this study

First, VOT was manually labelled for around 100 tokens for five different speakers to act as training data. Then the algorithm was run as a classifier on the entire recordings of the sample speakers, using the Praat Textgrids with the HTK segmentation as a guide. The algorithm used the lefthand boundary of an interval containing 'p t k b d g' as the place to start looking for VOT. The algorithm was applied once to search and predict for voiceless plosives, and then again for voiced ones.

We then manually inspected and coded the predictions: (Code 1) the prediction was correct; (Code 2) the prediction was very close, and so was corrected manually; (Codes 3-8) the prediction was completely wrong and not corrected, but the possible reason was recorded. This feedback was used to tweak the algorithm, and another round of manual correction was applied with improved performance.

In Section 5 we discuss the algorithm's performance. In Section 6, we present results from an initial analysis of VOT durations using Linear Mixed Effects (LME) modelling, reporting only those which showed $p < 0.0053$, given multiple planned comparisons. We effectively normalized for global speech rate by including 'speaker' as a random factor; we are currently developing a method of calculating estimates of local speech rate for each token.

5. Algorithm performance

5.1. Coding predictions

Figure 2 shows the distribution of codes from the manual inspection of predictions, shown as a proportion for each of the twelve speakers, arranged according to their decade of birth.

Working up from the base of the bars, Codes 1 and 2 are those tokens which were either correct (the darker shade) or easily corrected (the lighter shade). The proportion of correct tokens varies across the speakers, and does not seem to be contingent on any particular recording, despite the different kinds of recordings in our sample.

We distinguished different kinds of 'wrong' prediction. Codes 3-5 were difficult for the algorithm: where the forced alignment was incorrect (Code 3), so the lefthand boundary was wrongly placed; when speakers were overlapping with each other (Code 4); and when background noise interfered with the signal (Code 5). Code 3 showed us, incidentally, that the forced alignment was better than we had expected, and that poor alignment was not necessarily the property of any particular recording.

We could not explain some of the incorrect predictions (Code 6). It was more understandable that strong reduction of a plosive would lead to wrong prediction (Code 7). But more interesting were the instances of fricatives (for voiceless) and approximants (for voiced) stops (Code 8), which were far more common than expected, especially for /k/ (25% of the tokens for this stop were fricatives).

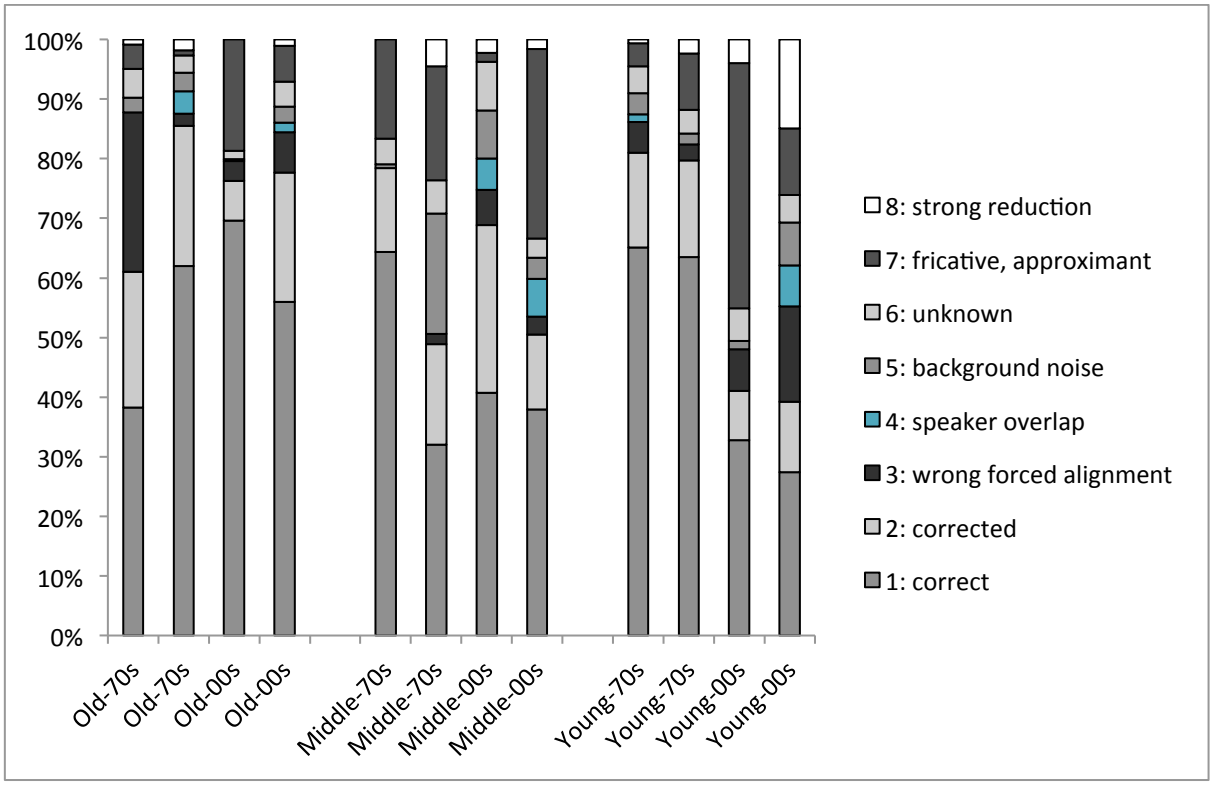

Figure 2: Distribution of codes from manual inspection of the VOT predictions for voiced and voiceless plosives (n = 4491). Each bar represents a speaker.

5.2.Overall results

For this sample, 52% of tokens were deemed correct and another 15% could easily be corrected; a third were wrong. The performance for voiceless stops (61% Code 1, 12% Code 2, 25% Codes 3-8) was better than for voiced stops (45% Code 1, 18% Code 2, 37% Codes 3-8). This was not surprising, given that the algorithm had only been recently adapted for the voiced stops, and these tokens were often fully voiced with slack and/or lenited release and very short burst.

In contrast to a manual procedure of locating each stop and manually labelling burst and voicing onset, our semi-automatic method provided an extremely fast way of obtaining reliable measures for two-thirds of our data. After the final round, it was possible to correct a TextGrid with voiceless VOT predictions for an entire speaker's recording (typically at least 20 minutes in duration) in under 30 minutes. We processed an initial count of 6125 tokens, which after removal of incorrect tokens was reduced to 4491 coded tokens. From these we gained a substantial dataset of 3012 reliable measures in a much shorter time than had we used manual segmentation.

6. VOT results

6.1.VOT and phonetic factors

The results for VOT for voiced and voiceless stops, according to place of articulation, are shown in Figure 3.

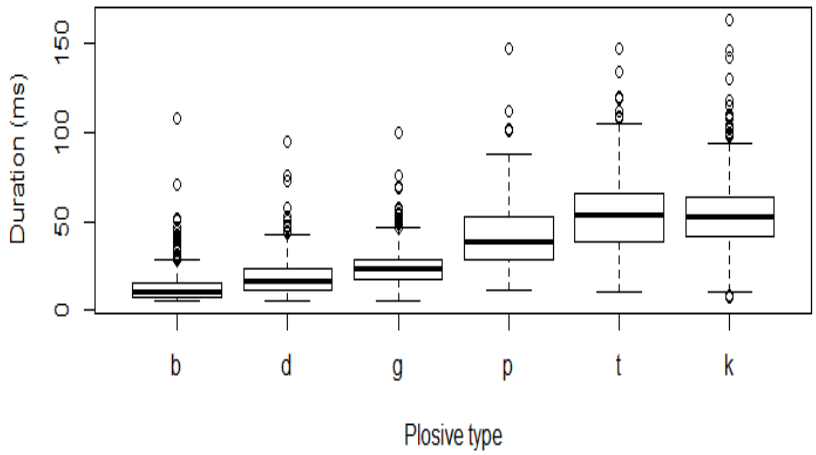

Figure 3: Boxplot of durations of VOT in milliseconds by place of articulation for voiced and voiceless stops (n = 3012).

A LME regression with fixed factors of voicing and following vowel, and random factors for speaker and word, confirms clearly longer VOT durations in voiceless stops. Asexpected, VOT patterns according to place of articulation (Cho and Ladefoged 1999), with velar stops significantly longer than bilabial stops.

6.2. VOT in voiced plosives in real and apparent time

The results for VOT in voiced plosives are shown in Figure 4. It is clear that these results reflect durations of both burst and release phase (cf Scobbie 2006), though aspiration is more apparent in the velar stop.

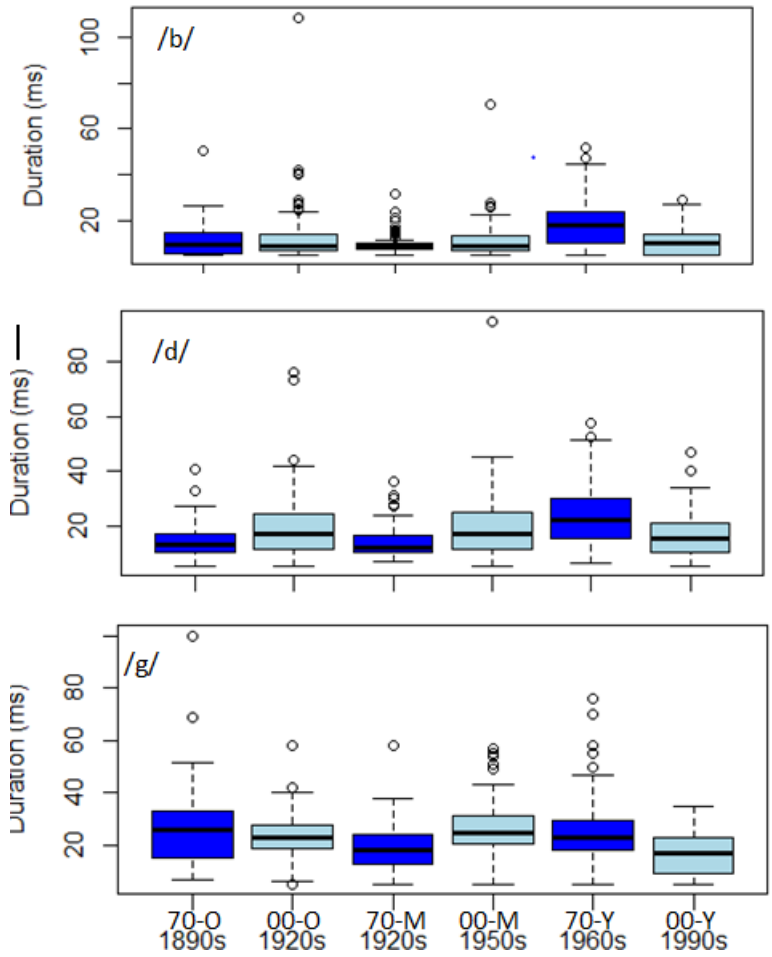

Figure 4: VOT durations in milliseconds for voiced stops according to decade of birth and date of recording (dark , 1970s; light, 2000s), two speakers per group (n = 1669).

LME models were run for the voiced stops with fixed factors of 'group', 'plosive type', and 'following vowel', and random factors of 'speaker' and 'word'. No effects were found to be significant with p-values below our threshold, though trends ($p < 0.05$) point to possible apparent time lengthening in 70s younger speakers compared to 70s elderly speakers for /b/ and /d/ (though for /b/ they are also longer than the adolescents recorded in the 2000s).

7.3. VOT in voiceless plosives in real and apparent time

The results for VOT durations for the voiceless stops are shown in Figure 5.

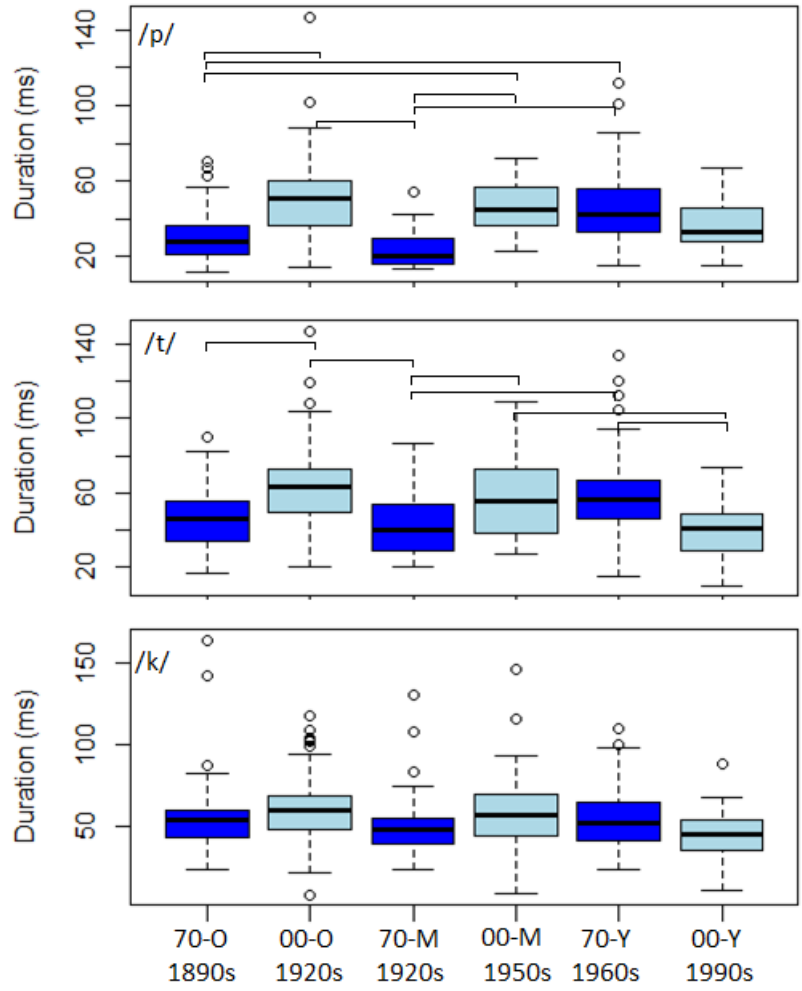

Figure 5: VOT durations in milliseconds for voiceless stops according to decade of birth and date of recording (dark , 1970s; light, 2000s), with two speakers per group (n = 1341). Comparisons between groups significant to $p < 0.0053$ are indicated with a bracket (/p/ and /t/).

LME models with the same factors as for the voiced stops were run for the voiceless stops. The results, while certainly preliminary with so few speakers, are suggestive. For /p/, in apparent time adolescents recorded in the 1970s show longer VOT durations than both middle-aged and elderly speakers. There are also real-time differences, with both elderly and middle-aged speakers recorded in the 2000s showing longer VOT durations than their 1970s counterparts. These results do not clearly pattern with physiological age. Elderly speakers can show shorter or longer VOTs; younger speakers can show longer or shorter durations. Nor are there indications that e.g. slower speech rate in older speakers leads to longer VOT durations.

/t/ shows similar results. There is an apparent-time lengthening in the 70s adolescents with respect to the middle-aged speakers recorded in the same decade; and real-time lengthening is found in the elderly and middle-aged speakers recorded in the 2000s, in comparison to the same aged speakers in the 1970s. But we also find that the adolescents recorded in the 2000s show shorter VOT durations than speakers born in the decades immediately before them (00-M and 70-Y). /k/ showed no significant effects.

7. Discussion

7.1.Semi-automatic measurement of VOT in sociolinguistic data

We had originally hoped that we could achieve a fully automatic system, but our semi-automatic coding, especially for voiceless stops, proved very good indeed, giving us 75% of our tokens as usable data in a much shorter time than could be possible using manual segmentation, or even manual correction of existing boundaries.

We were also encouraged by the fact that our coding system could be easily learned. This written paper includes a replacement speaker for one of the elderly speakers recorded in the 1970s. The predictions for this speaker were easily coded by a new annotator after very little training; including the random factor of 'Annotator' into the models presented above did not change the pattern of results.

The overall prediction of 52% stops judged as correct in spontaneous Glaswegian vernacular is still close to the previous results found by Sonderegger and Keshet (2012) for the spontaneous Switchboard and Big Brother corpora. This is all the more impressive when one acknowledges that neither of their test corpora suffered from factors such as incorrect forced-alignment, background noise, and substantial lenition of plosives.

At the same time, we are currently working to develop improved semi-automatic methods for characterizing voiced plosives. Measurement of negative VOT will be valuable, as will measures of closure duration and proportion of voicing during closure, even though these are computationally challenging for automatic analysis.

7.2.Age-grading or real-time change in Glaswegian stops?

We examined voiced and voiceless plosives from two groups of female speakers of Glaswegian vernacular, of different ages, recorded some thirty years apart. Even with our reduced view of the voicing contrast, only in terms of positive VOT durations, we found that the voicing contrast appears to be robustly maintained in these speakers; if there is change in progress, it would have to be a shift in phonetic realization. We also observe straightaway that our limited results point more to change than age-grading, since they do not align well with the findings, albeit from read speech, from the Scottish Border. Nor did we find consistency in VOT duration according to physiological age.

The clearest results for change are for /p/ and then /t/, where we find longer VOT durations in both apparent, and real time, suggesting an increase in aspiration of these voiceless plosives over time. The elderly and middle-aged speakers recorded in the 1970s were born in the 1890s and 1920s, and show much shorter VOTs than the adolescents who were born in the 1960s. But the intriguing finding is the longer VOTs in the elderly speakers recorded in the 2000s. These speakers were born around the same time as the 70s middle-aged speakers, but their much longer durations suggest a lengthening of aspiration during their lifetime. It is difficult to say from this sample when such a change may have started, but it seems likely that the period of the Second World War, with increased geographical and social mobility, as well as the major socio-spatial changes to the city of Glasgow, initiated in the 1950s, may have played a role. At the same time, the voiced stops at the same place of articulation show only a trend to a longer release phase; more information is needed about the contrast (Scobbie 2006).

This change may reflect the spread of SSE aspiration into Scots, as mooted by Johnston. Comparison of the average durations for the bilabial plosives in the Glaswegian speakers with those given in Scobbie 2006 (Table 4) is tricky given the likely differences between read and conversational speech. All of the Glaswegian speakers, irrespective of decade of birth or recording date, show average durations (24-50ms), which are closer to speakers with Shetlandic parents (22-47ms) than those of Scottish parents (61-83ms). But Scobbie's results also show gradience and individual variation , as opposed to clear-cut boundaries for particular varieties. Change in aspiration duration may be gradual, relational and flexible, with subtle shifts being triggered by, e.g. more contact with standard speakers, as opposed to the adoption of discretely more aspirated stops. It will be interesting to see the extent to which these results persist with more speakers per group, and also then to inspect individual differences.

We also note that the two adolescents recorded in the 2000s tend to show shorter durations, though not for the velar stops, suggesting that this is unlikely to be due to an effect of speech rate. The reason for the shorter VOTs is not entirely clear, but may reflect a return to more vernacular patterning in these speakers, especially if lengthened VOTs in the speakers born earlier also reflect more SSE-like patterning; analysis of more speakers will help us investigate this further.[ii]

8. Future directions

This paper marks the beginning of a collaboration to develop and apply automatic measurement methods to sociolinguistic data, in order to help increase the amount of reliable, usable datapoints for analysis, whilst at the same time reducing the huge time commitment which manual acoustic analysis of stops entails. We have two main goals for the future. We intend to improve and extend the algorithm for voiced plosives, and this work is now underway, though it is not a trivial undertaking. We also need to analyse more speakers for these time periods, and from other periods in between, in order to continue to investigate the extent to which these initial findings do represent real-time change in the voicing contrast in Scottish English.

References

Allen, J. Sean, Joanne L. Miller, and David DeSteno. "Individual Talker Differences in Voice-Onset-Time." *JASA* 113 (2003): 544-52.

Cho, Taehong, and Peter Ladefoged. "Variations and Universals in Vot: Evidence from 18 Languages.". *Journal of Phonetics* 27 (1999): 207-29.

Docherty, Gerard, Dominic Watt, Carmen Llamas, Damien Hall, and Jennifer Nycz. "Variation in Voice Onset Time Along the Scottish-English Border." Paper presented at the ICPhS XVII, Hong Kong, 2011.

Johnston, Paul. "Regional Variation." In *The Edinburgh History of the Scots Language*, edited by Charles Jones. 433-513. Edinburgh: Edinburgh University Press, 1997.

Lisker, L., and A.S Abramson. "Some Effects of Context on Voice Onset Time in English Stops." *Language and Speech* 10 (1967): 1-28.

Scobbie, James M. "Flexibility in the Face of Incompatible English Vot Systems." In *Laboratory Phonology*. 367-92. Berlin: Mouton de Gruyter, 2006.

Sonderegger, Morgan, and Joseph Keshet. "Automatic Measurement of Voice Onset Time Using Discriminative Structured Prediction." *JASA* 132 (2012): 3965–79.
Wells, John C. *Accents of English*. Cambridge: Cambridge University Press, 1982.

[i] This research was supported by grant RPG-142 from the Leverhulme Trust. We are very grateful to Cordula Klein for her help with manual coding, and to Robert Fromont for his help with LABB-CAT, and to audiences at ICLaVE7 and NWAV42 for their feedback on earlier versions of this paper. We are also grateful to Eivind Torgersen for his patience, and to two anonymous reviewers for their constructive feedback on an earlier version.
[ii] We are grateful to an anonymous reviewer for suggesting this interpretation.