Title
An Exemplar-based Approach to Automatic Burst Detection in Spontaneous Speech

Permalink
https://escholarship.org/uc/item/7f3546t9

Journal
UC Berkeley PhonLab Annual Report, 5(5)

ISSN
2768-5047

Author
Yao, Yao

Publication Date
2009

DOI
10.5070/P77f3546t9

An Exemplar-based Approach to Automatic Burst Detection in Spontaneous Speech

Yao Yao

University of California, Berkeley
1203 Dwinelle Hall, Linguistics Department, UC Berkeley, CA94720
yaoyao@berkeley.edu

Abstract

This paper introduces a novel algorithm for detecting burst in voiceless stops in spontaneous speech. This algorithm uses an exemplar-based approach for detecting aspiration noise, and avoids the normalization problem since the exemplars are inherently speaker-specific and environment-specific. The algorithm is trained and tested on 19 speakers' data. The overall error is estimated to be under 5 ms. We also show the wide range of variation in the phonetic makeup of stops in spontaneous speech and how the algorithm is improved to deal with the difficult cases.

Keywords: automatic burst detection, VOT, spontaneous speech.

1. Background

1.1 Methodological issue in research on pronunciation variation

In recent years, as many large-scale speech corpora (TIMIT, Switchboard, the Buckeye corpus, among others) are made available, quantitative analysis of these corpora has become an active new research area. In a seminal work, Keating et al. (1994) demonstrated two studies on the TIMIT corpus of read speech: a transcription study on segmental variation and an acoustic study using the audio signal. Since then, a growing body of literature has developed in the area of pronunciation variation in spontaneous speech (Byrd 1993; Keating 1997; Jurafsky et al. 1998, 2002; Gregory et al, 1999; Bell et al 2003, to appear; Raymond et al, 2006; Gahl, 2008; among others). However, the majority of these studies are limited to segmental and durational variation, such as shortening/lengthening, t/d deletion, flapping, and vowel alternation. Acoustic signal analysis is relatively rare (maybe with the only exception of vowel formants). This asymmetry in the literature is at least partly due to the fact that segmental/durational variation is easy to code as the information is already available in the transcription.

However, as the research on pronunciation variation develops in both depth and breadth, it becomes necessary to go beyond the transcription files and enter the acoustic signal. In order to do so, new methods need to be developed for extracting phonetically-important information from the speech data.

For many acoustic measures, there already exist automatic processing techniques. Nonetheless, these techniques are mostly developed for speech engineering and may not be directly applicable to the type of the research discussed here. Among other things, the techniques used in speech engineering are often designed to aid speech recognition and therefore precision (either in time domain or in frequency domain) is not of the highest concern. In pronunciation variation studies, however, precision is highly important, both because the acoustic measures are the actual objects of investigation, and that the size of the effect that researchers are looking for is often very small. For instance, it is not uncommon that certain factors are found to correlate with less than 5% of the variation in the acoustic measure, which would require the random error in the acoustic measure to be well under 5%.

1.2 The current study

In this paper, we report an attempt to address the above methodological issue, by presenting a case study on automatic burst detection in English voiceless stops. Automatic burst detection is widely used in speech engineering. The prevailing algorithm is one that detects the point of maximal energy change in high frequencies (Liu 1996; Niyogi and Ramesh, 1998; Das and Hansen, 2004). Liu (1996) reported an automatic burst detector as part of a larger landmark detecting system. The detector was trained on four speakers' read speech recordings (20 sentences per speaker) and was tested on two new speakers' recordings of 20 new sentences. In the training set, the detector had 5% deletion errors (i.e. missing real bursts) and 6% insertion errors (i.e. detecting spurious bursts) while in the test set, the rates are 10% and 2%, respectively. Liu didn't specify the temporal precision of the burst detector, but it was mentioned that of all landmarks (three different types altogether), 44% were detected within 5ms of the hand-labeled transcription, and 73% within 10ms. Though the system was not designed with high precision requirement, it still serves as a baseline model for the current study.

Our system makes use of a different approach. The general idea is that before burst (i.e. during closure phase), the spectrogram is similar to that of silence, while after burst (i.e. during aspiration phase), the spectrogram is similar to that of a fricative (see Figure 1). Therefore, the program finds the point of burst by constantly comparing the spectra of a moving time window to the spectral templates of silence and fricatives and looking for the point where silence-like-ness suddenly drops and fricative-like-ness suddenly rises. The system is trained and tested on 19 speakers' data from the Buckeye speech corpus (Pitt et al., 2007). The average temporal error is estimated to be within 5ms. The spectral template approach was first introduced in Johnson (2006), as an attempt to automatically analyze large speech corpora in a speaker-specific way. The main advantage of this approach is that it is inherently sensitive to differences among talkers and recording environments and therefore is more generalizable to new data.

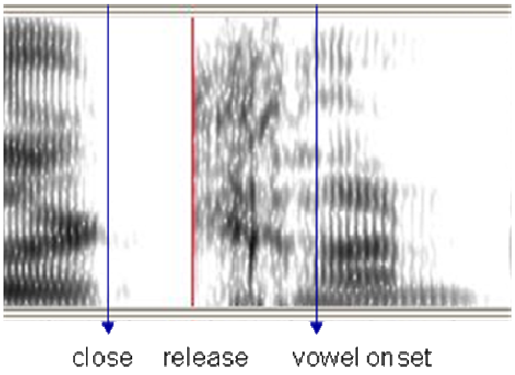


Figure 1. Spectrogram of a typical voiceless stop (the blue arrows mark the beginning and the end of the stop, while the red line marks the point of release.)

2. Data

2.1. Corpus

The Buckeye Corpus contains interview recordings of 40 speakers, all local residents of Columbus, OH. Each speaker was interviewed for about an hour with one interviewer. Only the interviewee's speech was digitally recorded. At the time of this study, 20 speakers' transcription was available, among which, one speaker's data were not used due to inconsistencies in the transcripts. The remaining 19 speakers are nearly balanced in gender and age (10 female, 9 male; 10 above 40 years, 9 under). Non-linguistic sounds, including silence, noise, laughter, and interviewer's speech, are also time-marked in the transcription. Silence in a running speech flow is not transcribed as silence, but attributed to neighboring sounds.

2.2. Target set

Since word-medial stops are often flapped in American English, we limit our target set to word-initial [p], [t] and [k], of which each speaker has from 231 to 1243 tokens (see Table 1).

Speaker	F01	F02	F03	F04	F05	F06	F07	F08	F09	F10
N	674	572	777	900	1243	490	231	449	699	412

Speaker	M01	M02	M03	M04	M05	M06	M07	M08	M09
N	514	931	624	793	657	406	541	557	628

Table 1. Count of target tokens in all speakers (top: female speakers; bottom: male speakers). N= number of tokens

3. Algorithm design

In view of the pattern in Figure 1, we build spectral templates of silence and voiceless fricatives for each speaker, and use these templates as references for evaluating how silence-like and fricative-like a certain chunk of acoustic data is.

3.1. Building spectral templates

Separate spectral templates are built for silence and voiceless fricatives of each speaker, using the following procedure. First, find all tokens of the phone in the speaker's speech data and discard the ones that are shorter than the medial duration (which would technically exclude half of the tokens). For each remaining token, calculate a 1X60 Mel frequency spectral vector using a 20 ms analysis window centered at the center of the phone and average across all tokens. The final template consists of an average Mel spectral vector, as well as the standard deviation of each dimension. Figure 2 below illustrates the spectral templates of [f] and silence of one speaker as examples.

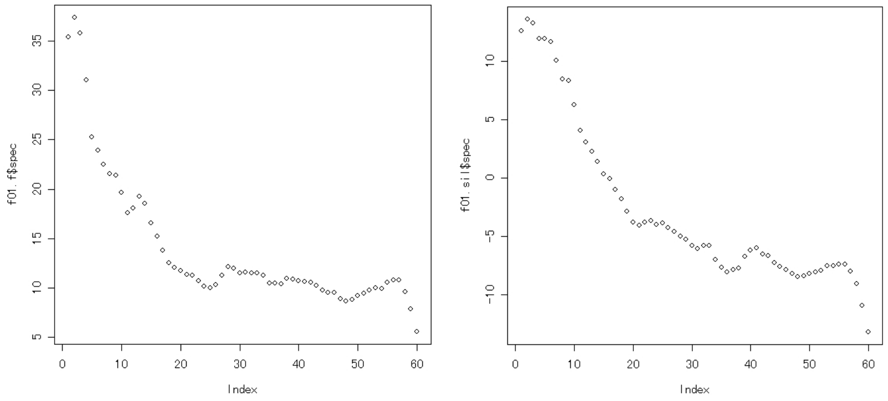


Figure 2. The Mel spectral vector in the templates for [f] (left) and silence (right) of speaker F01. The X-axis represents 60 equidistant bins on the Mel scale from 0 to 8000 and the Y-axis is the value of the corresponding dimension.

3.2. Calculating similarity scores

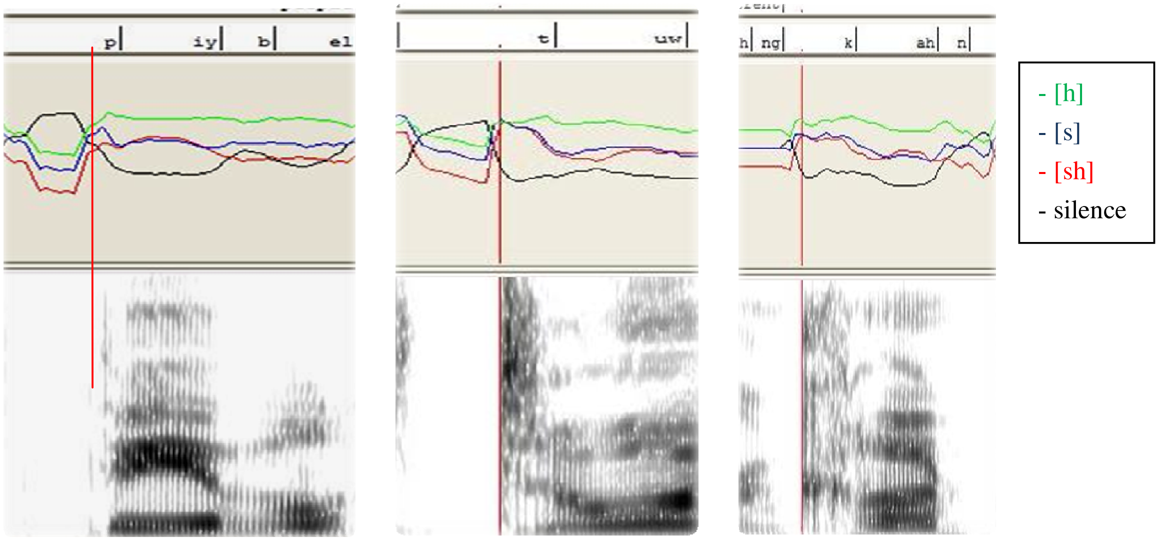
A similarity score measures how similar the acoustic data in the current window (size=20ms) is to a spectral template. It is calculated in two steps. A distance measure is first calculated between the Mel spectral vector of the current window and the average Mel spectral vector the template (see (1)), and then normalized to be the similarity score (see (2)).

$$(1) \qquad d_{x,u} = \frac{\sum_{j=1}^{60} | x_j - u_j | \frac{1}{sd(u_j)}}{60}$$

(where $d_{x,u}$ is the distance measure between the Mel spectral vector of window x and the template u; x_j is the jth coordinate in the Mel spectral vector of x, and u_j is the jth coordinate of the Mel spectral vector of template u; $sd(u_j)$ is the standard deviation of the jth coordinate in the Mel spectral vector of template u.)

(2) $S_i = e^{-0.005di}$ (where S_i is the similarity score of the current window to template i.)

The window moves with a step size equal to 5ms. Figure 3 below illustrates the similarity scores for silence and some fricatives during three example tokens of speaker F01. It can be seen that in all three tokens, the fricative similarity scores all rise around the point of release whereas the silence score drops.



[p] **[t]** **[k]**

Figure 3. Similarity scores and spectrogram of three stop tokens of F01: [p] (left), [t] (middle) and [k] (right). A red bar marks the position of first release in every token. Four similarity scores are shown: [h] (green), [s] (blue), [sh] (red) and silence (black).

3.3. Finding the point of release

As mentioned above, the general idea of the algorithm is to find the point within the stop where the silence similarity score suddenly drops and the fricative similarity score suddenly rises. There are two issues that need to be resolved here: (a) which period of rise/drop should be used, and (b) which fricative sound(s) should be used in spectral comparison. In the preliminary analysis, we found that the point of burst occurs most consistently after the point of fastest change in the similarity scores (i.e. maximal/minimal slope), and that using only one fricative score, the [sh] similarity score, is enough to capture the fricative-like-ness. Thus our baseline algorithm (see (3)) makes use of the silence similarity score (hereafter the <silence> score) and the [sh] similarity score (hereafter the <sh> score).

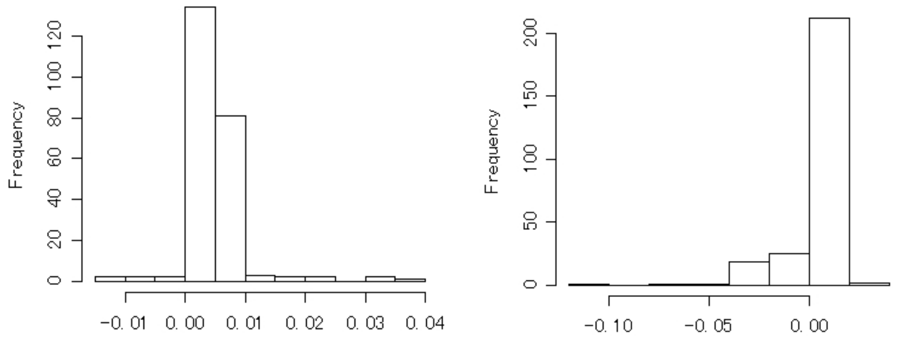
(3) Baseline algorithm

Find the end point of the period of fastest decrease in <silence> score and the end point of the period of fastest increase in <sh> score, and return the midpoint of the two as the point of release. If no decreasing period is found in <silence> score or no increasing period is found in <sh> score, exclude the token from the data set.

4. Testing and tuning

Part of speakers F07 and M08's data are used as developmental data. These two speakers are selected because they differ from each other in all available dimensions. Speaker F07 is an older female speaker, with the lowest average speaking rate (4.022 syll/s) of all 19 subjects, while speaker M08 is a young male speaker, with the highest average speaking rate (6.434 syll/s) (see Appendix I for all speaker's average speech rate). The developmental dataset consists of 231 tokens from F07 and 261 tokens from M08. Each token is hand-tagged for the point of release, judging from both the waveform and the spectrogram. If a stop token has no reliable trace for release, the beginning point of the phone is marked as the point of release, for the sake of calculating errors. If the stop has more than one release, the first release point is recorded.

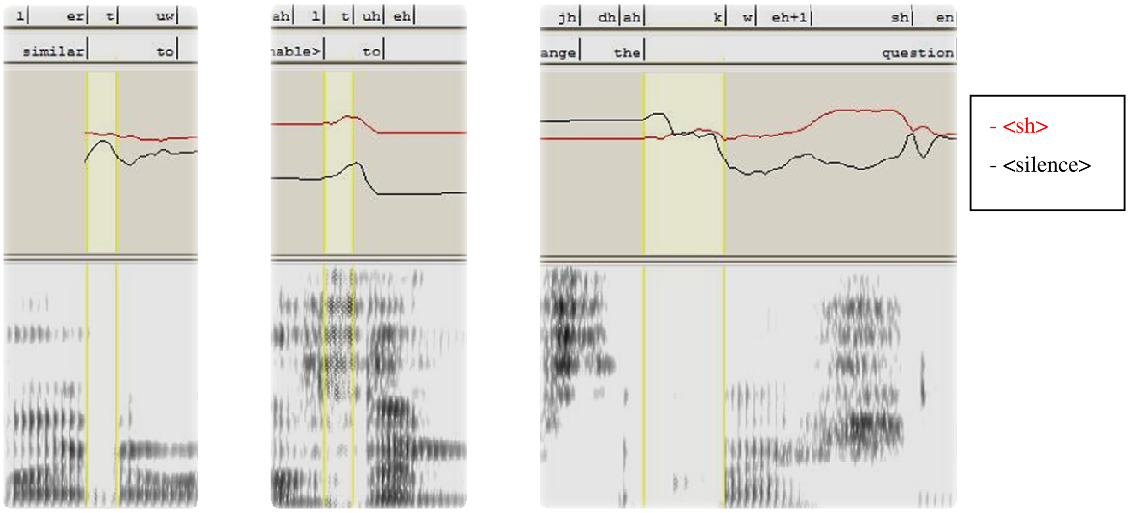
Using the baseline algorithm, the root mean square (RMS) of error (calculated as the lag between estimated point and tagged point) for F07 is 7.22ms. Moreover, errors are mostly distributed around 5ms, with a mean of 5.35ms (see Figure 4). This suggests that the estimated point is consistently earlier by about 5ms than the real point of release. If 5ms is added to all estimated values, the RMS of error is further reduced to 4.85ms.



F07 **M08**

Figure 4. Distribution of error values (in s) in F07 (left) and M08 (right). X-axis shows the error intervals in s, and Y-axis is the number of cases in the error interval. Error is calculated as the lag between the estimated point and the tagged point

When the baseline algorithm is applied to M08's training data, the RMS of error is much bigger, exceeding 13ms. The majority of errors are within 20ms, but there are a number of outliers that are more than 50ms in absolute value (see Figure 4). Similar to F07's data, most errors are positive, suggesting that the estimated point is consistently earlier than the real burst point. However, when 5ms is added to the estimation, the RMS error goes up to 14ms, probably due to the negative outliers. A closer examination of the outlier cases reveals three common types: cases with no release, cases with no closure, and cases of multiple releases (see Figure 5).



a. no release **b.** no closure **c.** multiple release

Figure 5. Illustration of three problematic cases in M08: (a) no release, (b) no closure and (c) multiple release. The duration of the target stop token is highlighted. <sh> score is shown in red and <silence> score is shown in black.

In Figure 5a, the transcribed duration of [t] is basically all blank in spectrogram. In other words, the release happens as the following vowel starts, but not during the stop. Figure 5b shows a case where the transcribed duration of the stop is all aspiration, with no closure portion. Being an extremely fast and soft talker, M08 has many tokens like these in the training set (23 out of 261). Since the points of release in these cases are hand-tagged as the starting point of the phone (for the purpose of calculating error), they greatly inflate the average error. Figure 5c shows a word-initial [k] in speaker M08. This velar stop is weakly (and doubly) released, which corresponds to two faint lines on the spectrogram around the mid point of the duration of the phone, with no noise-like distribution of energy following the release, which makes it hard for the program to recognize.

4.1 First rejection rule

In view of the first two types of problematic cases (see Figure 5a and 5b), we implemented the first rejection rule (see (4)) to reject cases that have insignificant changes in the similarity scores due to no obvious closure-burst transition.

(4) *First rejection rule*

A target word will be rejected if the most drastic changes found in scores are not drastic enough. The delta criterion is defined as a rising rate of 0.02 per step (i.e. per 5 ms) for <sh> score and a dropping rate of 0.04 per step for the <silence> score. If the <silence> score and <sh> score don’t meet the delta criterion, the case will be rejected, i.e. no release point will be estimated.

The two cutoff numbers, 0.02 and 0.04, are decided based on the observation from the training dataset. By applying the first rejection rule, 28 cases in M08's training data are rejected, 19 of which are hand-tagged for not having a reliable release point. The RMS error in M08's training data goes down to 9.27ms (see Figure 6). When 5ms is added to the estimated point, the error goes down by 0.01ms. For comparison, when applied to F07's data, the rule rejects 4 cases, and the RMS error goes down to 6.81ms. When 5ms is added to the estimated points, the RMS error further goes down to 4.22ms.

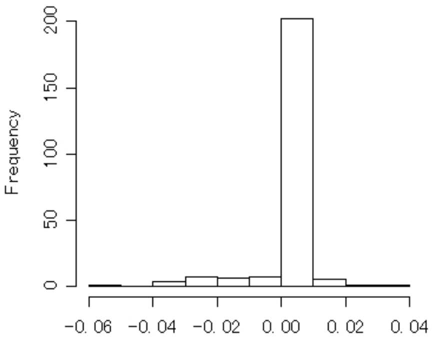


Figure 6. Error distribution in M08 after the first rejection rule is applied

4.2 Second rejection rule

The first rejection rule is designed to tackle with cases with no obvious closure-burst transition, due to missing closure or release gestures. What remains a problem are the cases with multiple releases. The program is designed to find the most significant change in similarity scores, but not necessarily the first one. This becomes a problem in multiple-release cases, since the first release is not always the most significant one. Multiple-release happens most often in velar stops. In fact, the case with the greatest error value (error = -60ms) in M08 is a multiply-released initial [k] in the word *cause* (see Figure 7). Not only is the velar stop multiply-released but also the first three (or four) releases are widely apart. Instead of finding the first release, the <silence> score tracker finds the second major release while the <sh> score tracker finds the third major release, and thus the program returns the mid point of the two, which is 60ms later than the first release.

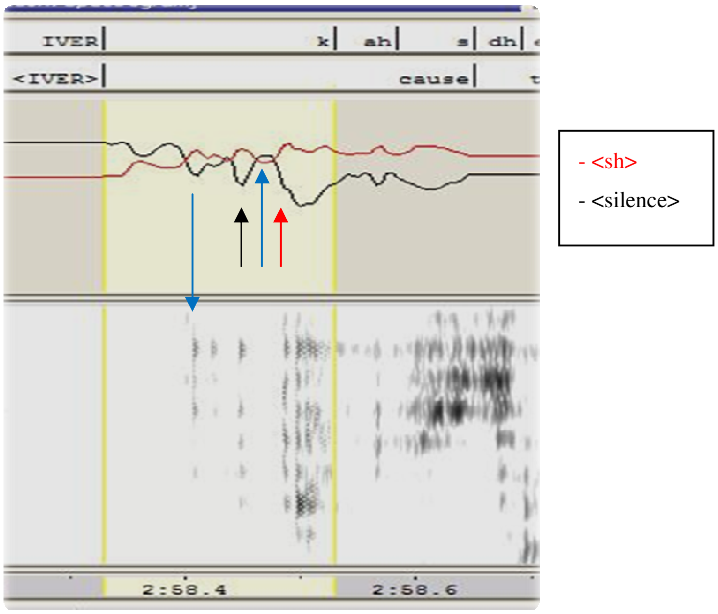


Figure 7. Multiply-released initial [k] in the word *cause* of speaker M08. First release is marked by the blue downward arrow; the candidate point found by the <silence> score is marked by the black upward arrow while the point found by <sh> score is marked by the red upward arrow; the final point of burst returned by the program is marked by the blue upward arrow in the middle

It would be ideal if the program find all points of release during the stop and return the first one. However, in practice, this is hard to do. Among other things, this would potentially interfere with the rejection of spurious releases. Therefore for the time being, we use a simple rule to reject cases of multiple releases, which partially addresses the problem. The general idea is to exclude cases where the two candidate points of release, returned by the <silence> score and the <sh> score respectively, are too far apart, which is indicative of an unusual multiple-release.

(5) Second rejection rule

If the two candidate points, one located in the <sh> score and the other one located in the <silence> score, are apart by more than 20ms, the case will be rejected, i.e. excluded from the data set.

By using the second rejection rule, the case shown in Figure 8 will be rejected because the two candidate points are apart by 40ms. It should be noted that this rule only rejects a particular type of multiple-release cases, i.e. the two candidate points (returned by the silence score and the fricative score) represent two separate releases and the two releases are more than 20ms apart. Even for this type, what the rejection rule does is simply exclude the case from the training set, without returning any release point.

Applying the second rejection rule to M08's data excludes 20 more cases and the RMS error is 5.64ms. After adding 5ms to the estimate values, the error is reduced to 3.44ms (see Figure 8). Notice that the number of outliers (i.e. residual after the rejection rules) is reduced to only 2, one on the positive side and one on the negative side. After applying the rule to F07's data, 3 more cases are rejected, and the RMS error goes down to 6.02ms. When 5ms is added to the prediction, the error is further reduced to 3.22 ms.

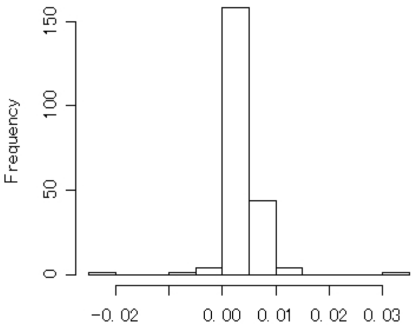


Figure 8. Error distribution in M08 after the second rejection rule is applied

4.3. Testing the algorithm on the rest of data

We have shown that the two rejection rules significantly improve the performance of the algorithm in both speakers' training data, especially in speaker M08's. Table 2 summarizes the number of cases excluded and the decrease in RMS error in both speakers after applying the two rejection rules sequentially.

	F07				**M08**			
	size	**error**	**error$_{+5}$**	**sd**	**size**	**error**	**error$_{+5}$**	**sd**
Baseline algorithm	231	7.22	4.85	4.85	261	13.11	14.00	13.17
after 1st rejection	227	6.81	4.19	4.19	233	9.27	9.26	8.94
after 2nd rejection	224	6.02	3.22	3.23	213	5.64	3.44	3.41

Table 2. Results with speaker F07's and speaker M08's developmental data. *Size* is the number of cases; *error* is the RMS error value; *error*$_{+5}$ is the RMS error after the estimates are shifted by 5ms to the right; *sd* is the standard deviation of error.

Overall, 7 of 231 cases are dropped from F07's data (rejection rate = 3.03%), and the RMS error is improved by 33.6%; in M08's data, 48 of 261 cases are dropped (rejection rate = 15.05%), and the RMS error is improved by 75.4%. For both speakers, the RMS error is further reduced when 5 ms is added to all estimated values, which suggests that the point found by the algorithm is consistently earlier than the real point of burst. Both speakers achieved a RMS error lower than 3.5ms after applying the two rejection rules. This is near-optimal, because given the step size of 5ms when calculating similarity scores, the optimal error in theory is 5/2 = 2.5ms. However, the large difference in rejection rates, 3.03% vs. 15.05%, suggests that there is a great amount of individual differences, in terms of the detectability of stop releases. Apart from gender and age, the most important difference between F07 and M08 is probably in speech style, as F07 is a relatively slow talker while M08 is extremely fast and soft (though the softness might be due to recording conditions).

We applied the baseline algorithm as well as the two rejection rules to all speakers' data, and found that the rejection rate ranges from 3.03% to 30.5%, with the average value of 13.13% and a standard deviation of 8.6%. (The details of rejection in all speakers are attached in Appendix II.) We also conducted a second test, using a random sample of 50 target tokens from all speakers, in which about half of the cases were from speakers with a high rejection rate (>20%). All 50 cases were hand-tagged for point of burst and the results were checked against the estimated values given by the program. Altogether 7 cases were rejected. In the remaining 43 cases, RMS error is within 5ms; in the 7 cases that were rejected, 4 were rejected by the first rule and 3 the second rule. Two rejected cases, one from each rule, were judged to be not strongly evidenced.

The complete algorithm, together with two rejection rules, is illustrated in the flow chart below.

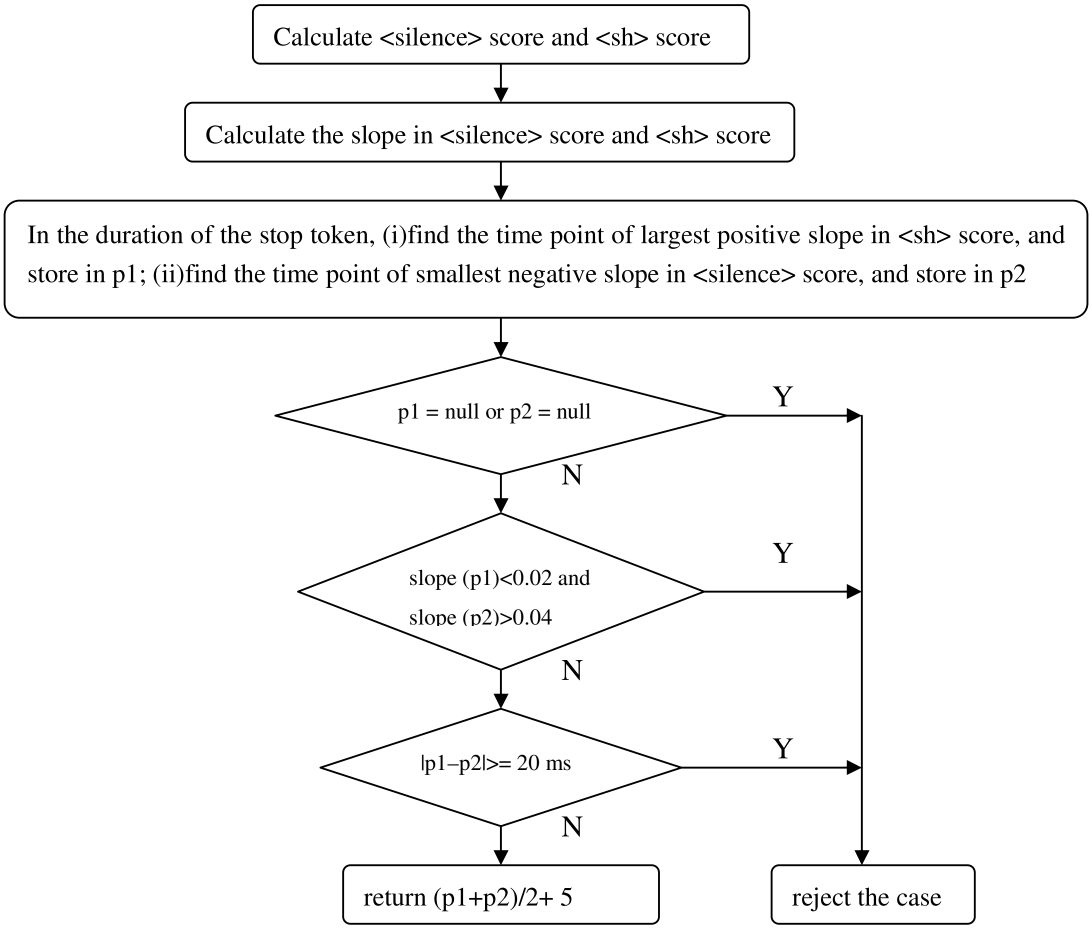

Calculate <silence> score and <sh> score
Calculate the slope in <silence> score and <sh> score
In the duration of the stop token, (i)find the time point of largest positive slope in <sh> score, and store in p1; (ii)find the time point of smallest negative slope in <silence> score, and store in p2
p1 = null or p2 = null
Y
N
slope (p1)<0.02 and slope (p2)>0.04
Y
N
|p1−p2|>= 20 ms
Y
N
return (p1+p2)/2+ 5
reject the case

Figure 9. Flow chart for finding the point of release

4. Summary of results

Table 3 lists the estimated mean values and standard deviations of closure duration, VOT, and total duration across all speakers by place of articulation.

	labial ([p])	**alveolar ([t])**	**velar ([k])**
N	2461	4142	3566
Mean(D_c)	69.5	48.9	54.9
Sd (D_c)	36.4	23.9	22.9
Mean(D_r)	48.0	51.2	57.9
Sd (D_r)	25.1	27.5	26.0
Mean(D_t)	117.6	100.2	112.9
Sd (D_t)	46.5	41.2	37.7

Table 3. Summary of duration values (in ms). N = total number of tokens; D_c = closure duration; D_r = VOT; D_t = total duration

Compared with the average durations found in Byrd (1993) for read speech in TIMIT (see Table 4), the Buckeye values are very similar, though the VOT values are a little bit longer.

	labial ([p])	**alveolar ([t])**	**velar ([k])**
Mean(D_c)	69	53	60
Sd (D_c)	24	29	26
Mean(D_r)	44	49	52
Sd (D_r)	22	24	24

Table 4. Duration values (in ms) from Byrd (1993). D_c = closure duration; D_r = VOT; D_t = total duration

5. General discussion

We present in this paper a pioneer case study in burst detection in voiceless stops in spontaneous speech. We use the exemplar-based spectral template approach, which was first proposed in Johnson (2006), and implement a burst detection program that finds the most likely point of burst within the duration of a voiceless stop in word-initial position.

A large part of the paper is devoted to the illustration of the wide range of variation in the realization of voiceless stops in spontaneous speech. Little is reported on this issue in the current literature, but we believe that it is central to the success of any automatic burst detection algorithm, especially those designed for spontaneous speech. We show in detail how different types of realization affects the performance of the algorithm, and how the algorithm can be improved to deal with the diversity. Two rejection rules are implemented to exclude cases where there is no obvious closure-burst transition and cases with more than one releases. Altogether these two rejection rules reduce the error by about 33.6% and 75.4% in two speakers' training data. The final RMS error is around 3.22ms in the training data, and is within 5ms in the 50 random test cases (for comparison, 44% of the landmarks in Liu [1996] are found within 5ms of the hand-transcription). The estimated values of closure duration and VOT in this study are similar to previous results of corpus studies. In particular, the estimated VOT values show the canonical pattern of increasing as the place of articulation moves from the lips to the velum (i.e. [p] < [t] < [k]).

Further improvement of the algorithm can be made in the following aspects. First, in the current study, we only tested two speakers' data in detail but didn't explore the full range of speaker differences. In the next step, we plan to test the algorithm more thoroughly, using all 19 speakers' data (and presumably the other 21 speakers in the corpus, whose data have been made available recently). Second, the cases of multiple releases can be investigated in more detail. The second rejection rule in the current algorithm doesn't fully address this problem – it only excludes the most extraordinary cases of multiple-release, i.e. the cases where the silence score and the fricative score find two separate releases and the two releases are apart by more than 20ms. Future work will focus on the modification of this rule by providing a way to identify all existing releases and return the earliest one. Last but not least, the current algorithm is only trained and tested on word-initial voiceless stops. It should be possible to extend the current program to stops that are word-medial or word-final, as well as voiced stops, for finding point of release and calculating VOT values.

References

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113, 1001–1024.

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Daniel Jurafsky. To appear. Predictability Effects on Durations of Content and Function Words in Conversational English. *Journal of Memory and Language*.

Byrd, Dani. 1993. 54,000 American stops. *UCLA Working Papers in Phonetics* 83, 97-116.

Das, Sharmistha & John H.L. Hansen 2004. Detection of Voice Onset Time (VOT) for unvoiced stops (/p/, /t/, /k/) Using the Teager Energy Operator (TEO) for automatic detection of Accented English. Proceedings of *the 6th Nordic Signal Processing Symposium*, 344-347.

Gahl, Susanne. 2008. Time and Thyme are not homophones: The effect of lemma frequency on word durations in a corpus of spontaneous speech. *Language* 84(3), 474-496.

Gregory, Michelle L., William D., Raymond, Alan Bell, Eric Fosler-Lussier & Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. In proceedings of *CLS* 35, 151-166.

Johnson, Keith. 2006. Acoustic attribute scoring: A preliminary report.

Jurafsky, Daniel, Alan Bell, Eric Fosler-Lussier, Cynthia Girand & William Raymond. 1998. Reductions of English function words in Switchboard. In Proceedings of *the International Congress of Speech and Language Processing* 98, 3111-3114.

Jurafsky, Daniel, Alan Bell & Cynthia Girand. 2002. The role of lemma in form variation. Carlos Gussenhoven and Natasha Warner (eds.) *Papers in Laboratory Phonology* 7, 1-34. Berlin: Mouton de Gruyter

Keating, Patricia A. 1997. Word-level phonetic variation in large speech corpora. In Berndt Pompino-Marschal (ed.) *ZAS Working Papers in Linguistics*.

Keating, Patricia A., Dani Byrd, Edward Flemming & Yuichi Todaka. 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication* 14 (1994) 131-142

Liu, Sharlene A. 1996. Landmark detection for distinctive feature-based speech recognition. *Journal of the Acoustical Society of America* 100 (5), 3417-3430.

Niyogi, Partha & Padma Ramesh. (1998) Incorporating voice onset time to improve letter recognition accuracies. Proceedings of *ICASSP '98* (1), 13-16.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume & Eric Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Raymond, William D., Robin Dautricourt & Elizabeth Hume. 2006. Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18, 55–97.

Appendix I Speakers’ average speaking rate and their relative rank in the group

	Average speaking rate	rank
F01	5.8552	3
F02	5.1846	10
F03	5.7704	4
F04	5.3442	8
F05	5.3042	9
F06	4.5032	16
F07	4.0218	19
F08	4.8831	12
F09	5.3513	7
F10	4.3584	18
M01	4.4421	17
M02	5.889	2
M03	4.8757	13
M04	4.6359	14
M05	5.6882	5
M06	4.6359	15
M07	5.6137	6
M08	6.4345	1
M09	5.1081	11
Mean	**5.152**	

Average speaking rate = total number of syllables produced / total amount of time (in s)

rank: the fastest (highest averaging speaking rate) is ranked 1, and second fastest speaker is ranked 2, and so on.

Appendix II Rejection rates in all speakers

	F01	F02	F03	F04	F05	F06	F07	F08	F09	F10
N	674	572	777	900	1243	490	231	449	699	412
Rsil	2	1	0	0	4	0	0	1	0	0
Rsh	1	1	1	0	1	0	0	0	0	0
R1	46	48	207	28	75	24	4	29	21	14
R2	12	30	29	33	48	31	3	18	15	18
Ngood	613	492	540	839	1115	435	224	401	663	380
R%	9.05	13.98	30.5	6.77	9.64	11.22	3.03	10.69	5.15	7.76

	M01	M02	M03	M04	M05	M06	M07	M08	M09
N	564	1027	784	865	724	512	636	618	718
Rsil	0	0	0	1	0	0	0	0	0
Rsh	0	1	0	0	2	0	1	0	1
R1	31	94	7	48	93	53	3	54	7
R2	12	128	39	27	98	21	12	44	20
Ngood	521	804	738	789	531	438	663	520	690
R%	7.62	21.71	5.86	8.78	26.65	14.45	2.51	15.85	3.89

N = the total number of target cases
Rsil = the number of cases where no decreasing period is found in <silence> score
Rsh = the number of cases where no increasing period is found in <sh> score
R1 = the number of cases rejected by the first rejection rule
R2 = the number of cases rejected by the second rejection rule
Ngood = the number of remaining cases after all rejection
R% = 1- Ngood /N, the rejection rate

Rejection is applied in the above sequence.