Closure duration analysis of incomplete stop consonants due to stop-stop interaction

Prasanta Kumar Ghosh and Shrikanth S. Narayanan
Department of Electrical Engineering, Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, California 90089
prasantg@usc.edu; shri@sipi.usc.edu

Abstract: An incomplete stop consonant is characterized either by an indistinguishable closure or a missing burst. If an incomplete stop happens due to a stop following another stop [stop-stop interaction (SSI)], its acoustics typically resemble that of a complete stop—one closure followed by a single burst. As a consequence, stop detectors would fail to distinguish an SSI from a complete stop. Analysis of the TIMIT corpus shows 35.04% incomplete stops (14.97% SSI). It is shown that by using automatically estimated (and hand-labeled) closure duration, complete stops can be distinguished from incomplete stops due to SSI with 69.66% (79.14%) accuracy.

PACS numbers: 43.72.Ar [DOS]
Date Received: February 21, 2009 **Date Accepted:** April 21, 2009

1. Introduction

Stop consonants in English speech have been a topic of research over the past few decades, particularly to better understand and analyze the dynamic and highly speaker- and context-dependent nature of these sounds. A stop consonant (/b/, /d/, /g/ [voiced] and /p/, /t/, /k/ [unvoiced]) is produced when there is complete closure of the articulators, stopping the airflow in the vocal tract, followed by a release or burst of air.[1] However, in conversational or even read speech, this acoustic signature of a stop is not always apparent due to intergestural overlap,[2] which sometimes results in the absence of a clear stop release or burst.[3,4] These short-dynamic acoustic variations make the state-of-the-art hidden Markov model based automatic speech recognizer (ASR) incapable of performing accurate fine phoneme distinctions for this class of sounds.[5]

To address this problem of ASR, researchers have proposed several alternative features and models to detect stop consonants. For example, to detect stop consonants, spectral and temporal features,[5,6] the optimal filter approach,[7] and the wavelet transform approach[8] have been used to capture a period of extremely low energy (corresponding to the period of closure) followed by a sharp, broadband signal (corresponding to the release). These features are in turn being used within novel automatic speech recognition frameworks such as those based on landmark detectors.[9] However, all these approaches for stop detectors implicitly assume that a stop should be a complete stop,[3] which is defined as one that should include an identifiable closure portion followed by a burst release. But corpus studies (in English) have shown that acoustic implementation of complete stops is only a fraction of the possibilities. For example, in a comprehensive study by Crystal and House,[3] complete stops accounted for only about 45% of the identified stops.

In this work, we investigate the complete and incomplete stops in the TIMIT database[10] with a specific focus on the robustness of the stop detectors in the presence of incomplete stops. In particular, we provide the detailed analysis of incomplete stop consonants[4] in the presence of stop-stop interaction (SSI). Spectro-temporally, these sounds share similar patterns with complete stops[1] motivating us to analyze their acoustic properties further. In a framework like distinctive feature landmark detection for speech recognition,[9] such an analysis would provide more insights into detecting stops more accurately. It should be noted that the formant transition is often used as an acoustic cue in stop detection,[6] which is not affected, in general, by the SSI.

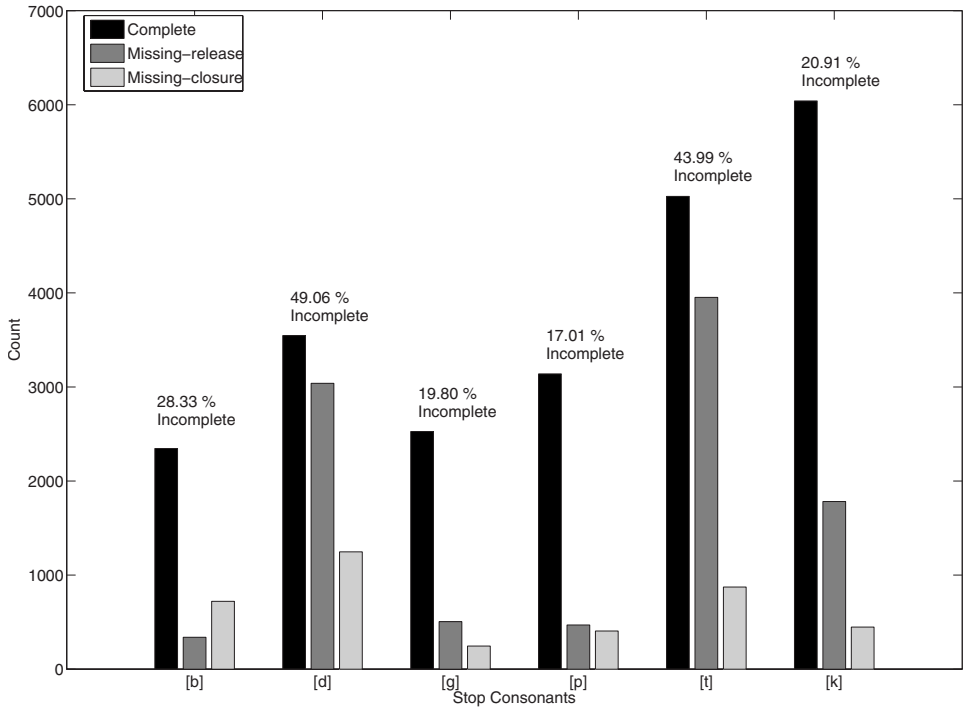

Fig. 1. Number of complete and different incomplete stops in TIMIT.

However, it is difficult to estimate formant transitions near a stop closure.[11] Hence both temporal and spectral properties of burst, closure duration, and voice onset time (VOT) are mainly used as acoustic features in stop detection.[5] But due to acoustic similarity between incomplete stops due to SSI and complete stop, stop detectors (which assume each stop is a complete stop) would miss one stop for every SSI. The analysis in this paper provides insight as to how the closure duration can be used to disambiguate SSIs from complete stops, leading to potential improvements to stop detectors. In our analysis, we found that with hand-labeled closure duration as a cue we can distinguish non-released stops due to SSI from complete stops with an accuracy of 79.14%. With automatically estimated closure durations, we can achieve a detection accuracy of 69.66%.

2. Complete and incomplete stops in TIMIT

We first study the relative frequencies of complete and incomplete stops of American English in TIMIT. The reason for choosing TIMIT is that it is a phonetically balanced database that has been well studied. In TIMIT, the phonetic transcriptions of the release of six stops are denoted by /b/, /d/, /g/, /p/, /t/, /k/ and their closures by /bcl/, /dcl/, /gcl/, /pcl/, /tcl/, /kcl/, respectively. A stop consonant [t] is counted as complete if /tcl/ is followed by /t/. On the other hand, an incomplete stop can be of two types—(1) if /tcl/ is followed by any phoneme other than /t/ (we call this a *missing-release* stop) and (2) if /t/ is preceded by any phoneme other than /tcl/ (we call this a *missing-closure* stop). This terminology applies to other stop consonants as well. *Carpet cleaners* (/kcl/k/p/ix/tcl/k/l/iy/n/..) is an example of a missing-release stop and *events* (/ix/v/eh/n/t/s/) is an example of a missing-closure stop (note the underlined portions). Figure 1 shows the number of complete, missing-release, and missing-closure stops for each of the six stop consonants in TIMIT. The percentages of the total incomplete (No. of missing-release + No. of missing-closure) stops are also shown in Fig. 1. Considering all the stop consonants together, there are 35.04% incomplete stops (missing-release, 28.96% and missing-closure, 11.31%) in the TIMIT database. It is clear from Fig. 1 that the percentage of incomplete [d] and [t] is high.

Table 1. Top 5 missing-release stops (for each stop closure); %SSI is the percentage of missing-release stops, due to SSI.

	Following phonemes					
Closures	First	Second	Third	Fourth	Fifth	%SSI
/bcl/	/d/	/t/	/jh/	/s/	/el/	28.02
/dcl/	/b/	/t/	/y/	/dh/	/z/	26.68
/gcl/	/l/	/z/	/n/	/ix/	/d/	20.19
/pcl/	/t/	/s/	/b/	/dh/	/m/	41.06
/tcl/	/s/	/dh/	/q/	/b/	/d/	16.41
/kcl/	/dh/	/m/	/t/	/s/	/d/	28.05

On average, for every two occurrences of [d] or [t], any acoustic feature based stop detector may fail to detect one of those occurrences.

Tables 1 and 2 show the top five phonemes that follow each of the stop closures and precede different stop releases in the case of missing-release and missing-closure stops, respectively. The underlined entries in both tables suggest that the release of a stop can follow a closure of another stop (we call this an SSI) and can contribute to an incomplete stop (23.54% for missing-release and 46.35% for missing-closure stops).

Browman and Goldstein[2] showed that gestural overlap can cause such deletion or assimilation leaving no acoustic evidence of the consonant burst, resulting in incomplete stops.

2.1 Incomplete stops due to SSI

From Tables 1 and 2, we see that when a stop consonant follows (interacts with) another stop consonant, it can lose its individual acoustic signature and manifest itself as an incomplete stop; we refer to them as incomplete stops due to SSI. A stop consonant interacts with another stop consonant either within a word (e.g., *subject* and *jumped*) or across words (e.g., *rapid car* and *sharp dresser*).[1] Incomplete stop due to SSI has one closure followed by a single burst and thus appears acoustically indistinguishable from complete stop.

%SSIs in Tables 1 and 2 refer to the percentages of different missing-release and missing-closure stops, which are due to SSI. We can see that the SSIs (particularly for missing-closure stops) cover a significant portion of the incomplete stops. This motivates us to investigate how an incomplete stop due to SSI can be distinguished from a complete stop using acoustic cues.

3. Closure duration of incomplete stop due to SSI vs complete stop

Although the acoustic pattern of an incomplete stop due to SSI appears similar to that of a complete stop, the mean duration of closure (as specified in the TIMIT transcription) for incom-

Table 2. Top 5 missing-closure stops (for each stop release); %SSI is the percentage of missing-closure stops, due to SSI.

	Preceding phonemes					
Release	First	Second	Third	Fourth	Fifth	%SSI
/b/	/dcl/	/tcl/	/pau/	/pcl/	/gcl/	56.31
/d/	/tcl/	/n/	/kcl/	/pau/	/bcl/	21.57
/g/	/tcl/	/ng/	/kcl/	/dcl/	/pau/	43.08
/p/	/tcl/	/dcl/	/kcl/	/gcl/	/bcl/	58.12
/t/	/kcl/	/dcl/	/pcl/	/pau/	/n/	71.93
/k/	/tcl/	/pau/	/dcl/	/pcl/	/bcl/	40.09

Table 3. Mean (SD) closure durations (in second) for different incomplete stops due to SSI, complete stops (bold entries correspond to complete stops) and stop-fricative, stop-nasal, stop-glides, stop-vowel interactions.

	Following phoneme categories									
	Stop release (SSI and complete stop)									
Closure	/b/	/d/	/g/	/p/	/t/	/k/	Fricative	Nasal	Glides	Vowel
/bcl/	**0.063**	0.099	0.133	0.111	0.098	0.094	0.063	0.056	0.049	0.059
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
/dcl/	0.086	**0.049**	0.088	0.096	0.072	0.093	0.043	0.05	0.056	0.057
	(0.02)	**(0.02)**	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
/gcl/	0.122	0.096	**0.047**	0.101	0.118	0.089	0.059	0.056	0.052	0.048
	(0.02)	(0.02)	**(0.02)**	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
/pcl/	0.109	0.122	0.125	**0.067**	0.091	0.1129	0.071	0.058	0.070	0.086
	(0.02)	(0.02)	(0.02)	**(0.02)**	(0.03)	(0.02)	(0.04)	(0.03)	(0.02)	(0.04)
/tcl/	0.097	0.086	0.098	0.093	**0.048**	0.085	0.045	0.054	0.063	0.063
	(0.03)	(0.03)	(0.03)	(0.04)	**(0.02)**	(0.03)	(0.03)	(0.02)	(0.04)	(0.04)
/kcl/	0.103	0.104	0.112	0.120	0.094	**0.054**	0.050	0.057	0.073	0.089
	(0.03)	(0.02)	(0.05)	(0.02)	(0.03)	**(0.02)**	(0.02)	(0.02)	(0.04)	(0.03)

plete stops due to SSI is consistently higher than that of complete stop consonants. The mean closure durations [with standard deviation (SD)] for all incomplete stops due to SSI and complete stops are shown in Table 3. The bold entries in this table indicate the minimum mean closure duration among all SSIs in a row. It is clear that the minimum durations also correspond to the complete stops (bold entries). This observation supports many previous studies in the literature: Olive *et al.*[1] reported smaller closure duration for single [t] than that of a geminate; Homma[12] provided a similar observation from an experiment on a set of 24 words spoken by four speakers; Manuel *et al.*[13] also observed similar differences in nasal consonant durations in "in a" and "in the." However, to the best of our knowledge, there has not been a comprehensive analysis of stop closure durations of different SSIs on a large multitalker dataset.

The total number of complete stops in TIMIT is 22624 and that of incomplete stops due to SSI is 1826 (8.07%). The normalized histograms of the closure duration of these two classes are shown in Fig. 2(a). This figure clearly shows that incomplete stops due to SSI and complete stops can be distinguished to some extent based on their closure duration.

We also found that when there is an interaction (for *missing-release*) where any phoneme other than a stop follows a stop consonant, the closure duration of this stop is not necessarily higher than that of the complete stop. Table 3 supports this observation. To compute the mean closure duration in such incomplete stops due to non-SSI, we first categorize the other interacting phonemes into fricatives, nasals, glides (and liquids), and vowels. From Table 3, it is seen that the mean closure durations in these cases are similar to those of the complete stops (as seen in bold entries of Table 3).

Also for missing-closure stops, there are cases where a stop is preceded by a phoneme other than a stop consonant. In these cases the closure duration does not exist for the stop and these are, in general, difficult to detect by acoustic cues (they are 17.33% of all incomplete stops).

4. Automatic classification of complete stop consonants and incomplete stops due to SSI

We have already seen that the mean closure durations (as mentioned in the TIMIT transcription) of a complete stop and an incomplete stop due to SSI are different even though both exhibit a similar acoustic pattern of a single closure followed by a single air burst. We perform two classification experiments to investigate the discrimination power of using closure duration as a feature. First, we use the actual closure durations transcribed in the TIMIT for an "oracle test,"

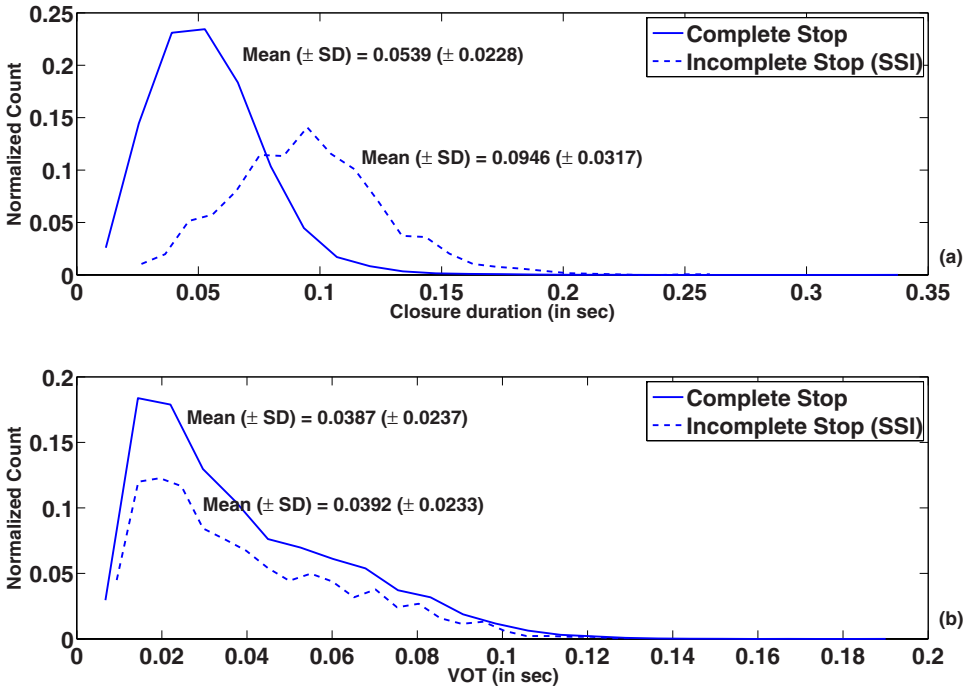

Fig. 2. (Color online) Normalized histogram (normalization is done so that the fractional counts in the histogram add up to 1) of (a) closure duration and (b) VOT.

that is, a classification experiment in which manually transcribed closure duration is used to discriminate between complete and incomplete stops due to SSI. Second, we perform the same detection experiments with automatically measured stop closure durations.

4.1 Oracle detection experiment

We randomly selected 500 complete stops and 500 incomplete stops due to SSI from TIMIT for testing and used the remaining segments for training. We trained two-component Gaussian mixture models (GMMs) for both classes using the EM algorithm (we tried other numbers of mixtures, but the best accuracy was obtained with two components). GMMs observed the closure durations that were provided in the TIMIT database. Since our test set was balanced and did not reflect the bias in the training set, we implemented maximum likelihood as opposed to maximum-a-posteriori classification scheme.

Classification accuracy was computed using 50-fold cross-validation, resulting in an average accuracy of 79.14% (SD=1.23%). Complete stops were classified with a mean accuracy of 80.08% (SD=1.92%), whereas SSIs with 78.2% (SD=1.73%). Thus class specific accuracies are not very different. We also tested the use of VOT as an additional feature, but it did not improve the accuracy significantly. Thus it can be concluded that VOT is not a useful cue for distinguishing SSI from complete stops. The histogram of VOT for these two classes supports this fact [see Fig. 2(b)].

4.2 Detection using automatically estimated closure duration

In this experiment, we designed a simple stop detector using the energy of the closure duration as a feature and consequently estimated the closure duration for a detected stop. Since stop releases are transient events, we used an analysis window of 1 ms length with no overlap. We computed the energy of the speech signal (normalized to ±1) in each analysis window. We used the TIMIT training dataset to learn the distributions of energies of the signal in the analysis

window for stop closures and for events other than stop closures. Stop closures were detected by thresholding the energy, using the equal error rate threshold (the threshold at which recall and precision rates are equal), which turned out to be 0.6974 for these data. If the energy of the signal in an analysis window was less than 0.6974, we declared that the frame belongs to a stop closure. However, in this approach, many spurious frames were detected as frames belonging to stop closure. To prevent this, we imposed another constraint. From the histogram of closure durations [Fig. 2(a)], we observe that the minimum closure duration is ~15 ms. Thus if at least 15 consecutive frame energies are below 0.6974, we declared that the respective sequence of frames corresponds to a stop closure. If N ($N \geq 15$) consecutive frame energies were consistently less than the threshold, we considered the estimated duration of the respective stop closure is N ms.

We used the TIMIT test dataset for evaluation. Using the above-mentioned simple stop closure detection algorithm, we could detect 92.39% of all stops (including both complete and SSI) in the test dataset. The estimated stop closure durations were then used to classify the stops into either complete stops or SSIs, using two-component GMMs trained on the transcribed closure durations of the training dataset. The mean classification accuracy was 69.66%. Complete stops were classified with a mean accuracy of 63.70%, whereas SSIs with 75.62%. The reduction in accuracy compared to that of the oracle test is due to the fact that the closure durations are estimated and not the actual ones as transcribed in the TIMIT database.

5. Conclusion

We showed that closure duration can be used as a feature to classify the incomplete stops due to SSI and the complete stops in read speech of the TIMIT with 69.66% (79.14%) accuracy for automatically estimated (reference) durations. We also found that the closure durations of the incomplete stops, when not due to SSI, are similar to those of complete stops.

Our analysis provides opportunities for improvement of stop consonant detectors, particularly for applications such as distinctive feature landmark detectors for speech recognition.[9] It is important to note that the problem of nonreleased stops is not speaker dependent as the data analyzed come from multiple talkers. However, a similar study on conversational speech remains to be done to fully understand the effect of variability in stop production on the performance of stop detectors.

Acknowledgments

Work supported in part by NIH and ONR-MURI.

References and links

[1]J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach* (Springer-Verlag, Berlin, 1993), pp. 227–312.
[2]C. Browman and L. Goldstein, *Tiers in Articulatory Phonology, With Some Implications for Casual Speech*, Papers in Laboratory Phonology, edited by J. Kingston and M. E. Beckman (Cambridge University Press, Cambridge, 1990), pp. 341–386.
[3]T. H. Crystal and A. S. House, "The duration of American-English stop consonants: An overview," J. Phonetics **16**, 285–294 (1988).
[4]T. Deelman and C. M. Connine, "Missing information in spoken word recognition: Nonreleased stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **27**(3), 656–663 (2001).
[5]A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," IEEE Trans. Speech Audio Process. **9**, 833–841 (2001).
[6]M. F. Dorman, M. Studdert-Kennedy, and L. J. Raphael, "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychophys. **22**(2), 109–122 (1977).
[7]P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," J. Acoust. Soc. Am. **111**, 1063–1076 (2002).
[8]F. Malbos, M. Baudry, and S. Montresor, "Detection of stop consonants with the wavelet transform," in Proceedings of IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (1994), pp. 612–615.
[9]A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detector for speech recognition," J. Acoust. Soc. Am. **124**, 1739–1758 (2008).
[10]J. S. Garofolo, "TIMIT acoustic-phonetic continuous speech corpus," LDC, Philadelphia (1993).

[11]Y. Zheng, "Acoustic modeling and feature selection for speech recognition," Ph.D. thesis, University of Illinois at Urbana-Champaign (2005).
[12]Y. Homma, "Durational relationship between japanese stops and vowels," J. Phonetics **9**, 273–281 (1981).
[13]S. J. Manuel, S. S. Hufnagel, M. Huffman, K. N. Stevens, R. Carlson, and S. Hunnicutt, "Studies of vowel and consonant reduction," in Proceedings of ICSLP (1992), pp. 943–946.