54,000 American Stops

Dani Byrd
Department of Linguistics, UCLA
405 Hilgard Avenue; Los Angeles, CA 90024-1543 USA

Abstract

An analysis of oral and nasal stops, affricates, oral and nasal flaps, and glottal stops was conducted using data from the TIMIT database. The results offer descriptions of the frequency of these segments in the large TIMIT corpus, mean segmental durations, voice onset times, and certain effects of voicing, place, word position, and speaker sex. Because of the quantity and diversity of speech data included, this study provides a characterization of American English stops which represents an overview of American speakers' production of these consonants in read materials. The results obtained are considered with respect to standard descriptions in the field which have generally used datasets which are smaller both with respect to context variability and speaker pool.

Introduction

All the world's languages have stop consonants (Maddieson, 1984). Stops can be considered to be composed of three phases: onset, closure, and offset. They can occur at many places of articulation, with many variations in glottal state and airstream mechanism (see Henton, Ladefoged, and Maddieson, 1992). English utilizes only a small subset of these possibilities. Accounts of the nature of stop consonants in English include Keating (1984), Fox and Terbeek, (1977), Crystal and House (1988a,b,c), Lisker and Abramson (1964), Zue and Laferriere (1979) and many others. This paper will augment these efforts with a report on some characteristics of a large and geographically comprehensive sample of American English stops. The TIMIT database of read American English offers an enormous quantity of data produced by many different speakers.

The TIMIT database was designed jointly by the Massachusetts Institute of Technology, Texas Instruments, and SRI International under sponsorship from the Defense Advanced Research Projects Agency-Information Science and Technology Office (DARPA-ISTO) for the development and evaluation of automatic speech recognition systems (Lamel, Kassel, and Seneff, 1986). It is described by Zue, Seneff and Glass (1990) and Pallett (1990). It was intended that TIMIT incorporate sufficient variability to examine the acoustic realization of phonetic segments as affected by canonical characteristics of the phoneme, contextual dependencies, syntactic effects, and speaker-specific factors of age, dialect, sex and education (Lamel et al., 1986). TIMIT includes 2342 different sentences read by 630 speakers (ten sentences per speaker). There were also two "dialect calibration" sentences read by all 630 speakers. All of the sentences are segmented and labeled as outlined by Seneff and Zue (1988). The validity of the results reported here depends on the correctness and consistency of the phonetic transcriptions. See Keating, Blankenship, Byrd, Flemming, and Todaka (1992) and Byrd (1992a) for a description of UCLA's database implementation of TIMIT and its uses in linguistic phonetic research.

The segments, or "phones," included in this study are the six phonemic oral stop consonants, three consonantal and three syllabic nasal stops, the two affricates, the glottal stop, and the oral and nasal alveolar flaps. Included in this sample are 54,384 stops, affricates, and flaps. The breakdown of the quantity of the data is shown in Table 1.

stops	**24,414** oral stop closures
	18,101 nasal stop closures
affricates	**2,055**
flaps	**3,649** oral flaps
	1,331 nasal flaps
glottal stops	**4,834**

Table 1—data included in this study of the TIMIT database

This then is by far the most comprehensive study in terms of quantity, and probably diversity, of data on American stops offered to date. Labeling in the TIMIT corpus includes two phases for oral stop consonants: closure and offset (release); and similarly for the affricate. However, nasals, flaps, and glottal stops are segmented as closures only. Below distributional frequency and the durational characteristics of these consonants in TIMIT are described. Some discussion of stop release frequency is also offered. Details of the behavior of these stops in assimilatory processes or their acoustic nature are not addressed. Rather TIMIT is exploited for the "big picture" it offers for the general description of American stops.

I. Oral Stops

Randolph (1989) in his dissertation used three databases, one of which was TIMIT. He noted that transcription errors were infrequent and of a limited number of types. He does note that weak stop releases may go undetected and that /t/'s might be more likely than /p/'s and /k/'s to be transcribed as glottal stop in the presences of pitch irregularities.

A. Closures

A search of the TIMIT database was made for oral stop closures and releases. The dialect calibration sentences were excluded so that the frequency counts would not be biased by the repetition of these sentences by all speakers. In the TIMIT transcription, 24,414 oral stop closures and 21,847 oral stop releases were found. This suggests a ratio of releases to closures of about .895. However, note that in sequences of stops in which the first stop is unreleased, the closure is transcribed as having the place of articulation of the first stop and the release of the second stop. Thus, the ratio of release frequency given for each stop should be considered only rough due to frequent stop assimilation in English. The following table records the number of stop closures and releases of each type produced in a search of the database.

stop	closures (n)	releases (n)	rel (n)/ clo (n)
p	3599	3545	.985
b	2651	3067	1.16
t	6736	5326	.791
d	4179	3275	.784
k	5472	4998	.913
g	1777	1643	.925

Table 2—number of oral stop closures and releases

Note that the /t/ and /d/ categories here and their corresponding durations below exclude flaps, which are labeled separately in TIMIT.

The durations of stop closures are given in the table below and shown graphically in Figure 1.

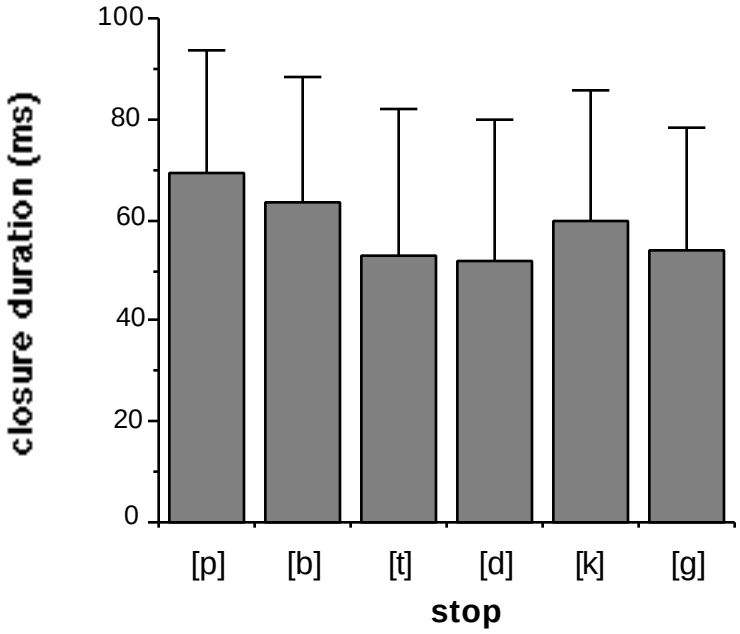

Figure — oral stop closure duration by stop identity; [p] 69 (s.d. 24), [b] 64 (s.d. 25), [t] 53 (s.d. 29), [d] 52 (s.d. 28), [k] 60 (s.d. 26), [g] 54 (s.d. 24); all means and standard deviations here and below are reported to the nearest millisecond

Analysis of variance determined there to be a significant effect of stop identity on the duration of closure ($F(5,24408)=253.211$, $p=.0001$). A post-hoc Scheffé's S-test showed all pairwise comparisons of closure durations to be significantly ($p<.05$) different except [d] from [t] and [g], and [t] from [g]. Keating (1984) has found fronter closures to have longer durations. This data supports this for the comparison of the labials and velars; however, the intermediate alveolars

have the shortest durations. Zue's (1976) finding, not supported in Crystal and House (1988a), of longer closure duration for [p] than for [t] and [k] is supported here.

A further examination of closure durations was conducted with respect to the factors of voicing and place. The voiced closures had a mean duration of 56ms (s.d.=27), and the voiceless closures had a mean duration of 59ms (s.d.=28ms). Figure 2 shows closure duration by place.

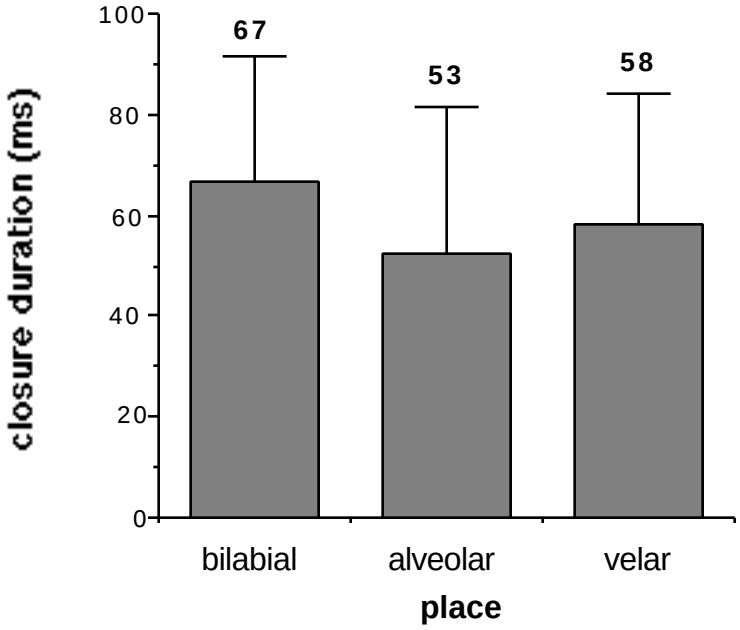

Figure — oral stop closure duration by place of articulation; means are shown above the error bars in ms, standard deviations range from 25 - 29ms.

ANOVA determines there to be a significant effect of both voicing ($F(1,24405)=122.697$, $p=.0001$) and place ($F(2,24405)=525.019$, $p=.0001$). There was also a significant interaction of voicing and place ($F(2,24405)=18.365$, $p=.0001$). A post-hoc Scheffé's S-test on place shows all pairwise comparisons to be significantly different ($p=.0001$). The interaction of place and voicing was significant at the $p=.0001$ level. The pattern of duration decreasing from bilabial to velar to alveolar is maintained in both voiced and voiceless stop closures. While the mean voiced alveolar stop closure was shorter than the voiceless in accordance with the overall pattern, this difference was on the order of 1ms, whereas the difference at the bilabial and velar places was approximately 6ms. Recall from the above analysis of the effect of stop identity, that the [d] closure duration was not significantly different from that of [t] as shown in the post-hoc test.

These results are in agreement with Crystal and House (1988a,c and 1982) in the general pattern displayed for closure duration as a function of place, with the alveolar closures shorter than those at other places. However, the differences of place found by Crystal and House were slight (1988a). Also in contrast with Crystal and House (1988a,c), these results support Luce and Charles-Luce's (1985) finding that closure duration gets progressively shorter from bilabials to velars to alveolars. Crystal and House do not even find the [p] to be longer than [t] and [k], as found by Zue (1976) and others. The difference in closure duration between the bilabials and

other places found here is somewhat smaller than that reported by Fisher-Jørgensen (1964). However, the difference between [b] and [d] is close to that found by Smith (1978). Subtelney, Worth and Sukuda (1966) found that dental stop closures were longer than the labials in their data set, a finding clearly not supported here. Finding labial closures to be longer than those at other places has been suggested to be context free (MacNeilage, 1972). The alveolar closures are shorter presumably because the lesser mass of the tongue tip permits more rapid movement (cf. Kuehn and Moll, 1976). Similar observations concerning the differences between bilabial and velar articulations have to take into account the fact that it seems that the aerodynamic conditions favor a longer closure for bilabials in the case of the stops, perhaps because oral pressure build less rapidly, although Ohala (1983) suggests that this difference is negligible. He suggests that it can be increased through passive and active oral cavity enlargement.

Crystal and House (1988c) found no effect of voicing on closure duration, in contrast to our results at the velar and bilabial places. Note that the studies of Chen (1970) using citation forms and Luce and Charles-Luce (1985) using words in a frame found that voiceless closure duration were longer than voiced. The statistical results reported above for the TIMIT read sentences also generally support this finding. Crystal and House (1988c) remark that it is "sobering to note" that they did not find voiceless stops to have longer closures than voiced stops in their connected speech data as this was found by Meyer (cited in Madebrink, 1955), Medebrink (1955), and Suen and Beddoes (1974) in early work. See also Lisker (1957), Lisker (1972), Umeda (1977) (connected speech), Malécot (1968), Subtelney, Worth and Sukuda (1966), Stathopoulos and Weismer (1983), Slis (1970, 1971), and Prosek and House (1975). Many of these investigators found this difference to be limited to particular word positions such as word-initial, pre-stressed. However, the results above for the large quantity of connected speech data in TIMIT do find voiceless closures to be longer than voiced, although the difference for [t] and [d] did not reach significance in the post-hoc test. Note that Subtelney et al.(1966) found a large difference in the dentals, larger than the bilabials. The difference between the voiceless and voiced bilabials and velars is slightly smaller than that found by Umeda (1977) and Stathopoulos and Weismer (1983) in connected speech; it is substantially smaller than the differences found in the non-connected speech experiments noted above.

B. Release Duration

When followed by a vowel or sonorant, the duration of the labeled stop release phone in TIMIT is equivalent to the phonetic measure of voice onset time. "This period begins with the location of the stop burst and ends at the first sign of periodicity in the waveform" (Randolph, 1989). If a released stop is followed by an obstruent, its release duration will be marked as ending when the characteristic acoustic effect of the next phone begins, i.e. frication for a fricative (Randolph, 1989). Not surprisingly, ANOVA (including both voiced and voiceless stops) finds a significant effect of stop on release duration or VOT ($F(5,21841)=1941.821$, $p=.0001$). The release durations are given in Table 3 below. Note the relatively small standard deviations for the voiced stops.

stop	release (ms)	s. d.
p	44	22
b	18	7
t	49	24
d	24	14
k	52	24
g	27	11

Table 3—stop release durations and standard deviations in milliseconds

A post-hoc Scheffé's S-test shows all pairwise comparisons to be significantly different at the p=.0001 level. In comparison with Lisker's (1964) VOT values for four speakers producing stops in sentences, the TIMIT values are somewhat longer, although they are not as long as Lisker and Abramson's (1964) data in isolated words. Lisker (1957) also notes several cases of prevoicing in his sentence data (see also Lisker and Abramson, 1964; Keating, et al., 1983). The version of TIMIT which is distributed was recorded with a close-talking microphone which does not generally preserves this low amplitude prevoicing voicing, and prevoicing is not segmented/labeled in TIMIT. The TIMIT release duration values for [p, b, d, g] are generally in close agreement with Klatt's (1975) VOT values for word initial single stop consonants spoken in monosyllables by three speakers in a frame sentence, although we might expect the values in that condition to be higher than those for the mixture of read sentences here. However, the VOT values found here for [t] and [k] are much shorter than in Klatt's results.

A two-factor ANOVA testing the effect of place and voicing, and their interaction, on release duration shows there to be significant effects of both place (F(2,21841)=300.759, p=.0001) and, of course, voicing (F(1,21841)=7695.412, p=.0001). There is no significant interaction. The mean release duration for voiced stops is 22ms and for voiceless stops is 49ms.

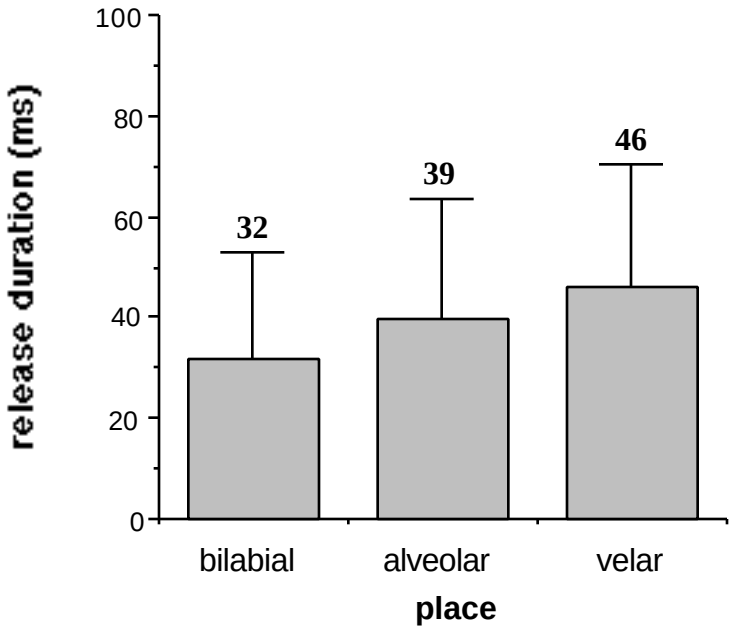

Figure 3—voice onset time by place of articulation

A post-hoc Scheffé's S-test on place show all pairwise comparisons to be significantly different (p=.0001). Henton, Ladefoged, and Maddieson (1992), Crystal and House (1988a), Zue (1976), and others have noted that VOT increases on average as the place of articulation moves from bilabial to alveolar to velar. The TIMIT data is in accordance with this finding. Randolph, using three databases of read sentences, found velars to have the longest release duration at 31ms and labials the shortest at 18ms (Randolph, 1989). This follows the same pattern shown above, but reports generally shorter times.

The differences between the mean voiced and voiceless release duration here are smaller than those reported by Carlson and Granström (1986) but slightly greater than those reported by Crystal and House (1988c). The TIMIT mean release duration for both voiced and voiceless stops is greater than Lisker's (1967) values for 3 speakers in a sentence condition. The TIMIT voiced VOT mean is in close agreement with Klatt's (1975) value for voiced stops before sonorants which he found to be slightly longer than the value before vowels. However, the TIMIT voiceless release duration is much shorter than that found by Klatt (1975) with three speakers recording 25 monosyllabic words beginning with one to three consonants in frame sentences.

Randolph's 1989 dissertation is a major source of segment data from read speech databases. reports voiceless release durations as being twice as long as voiced (49ms vs. 24ms.) in his three database study. Randolph's reported release durations and standard deviations are almost in exact agreement with those found here for the TIMIT database alone. Randolph goes on to note that voiceless stops in syllable onset position have approximately 1.5 times as long a release duration as stops in other syllable positions. He comments that voiceless stops in non-(syllable) initial positions have longer release durations when they precede nasals, glides, and vowels than when preceding affricates, other stops, and fricatives. He notes that stops in a syllable onset have shorter release durations when following obstruents (48ms) then when following nasals, glides, and vowels (Randolph, 1989). Of the onset stops preceded by obstruents, those in clusters with /s/ have the shortest release duration (32ms) (Randolph, 1989). Randolph finds that stops in a non-falling stress environment have the longest release duration of non-onset stops. He comments that this is perhaps an indication of these stops being resyllabified as onset stops.

The two-way interaction of place and voicing is statistically significant at the p=.0040 level. The alveolar and velar stops both have approximately 25ms difference in voiced and voiceless release duration. The bilabial place has approximately a 27ms difference. However, when mean total (closure+release) duration is considered, bilabial and velar releases both have a 31ms difference between voiced and voiceless, and alveolars a 26ms difference between voiced and voiceless. The ratio of voiced total mean duration to voiceless total mean duration is approximately the same for all three places (from .723 for velars to .745 for alveolars.

The relationship between mean closure and release duration for each stop can be seen in Figure 4.

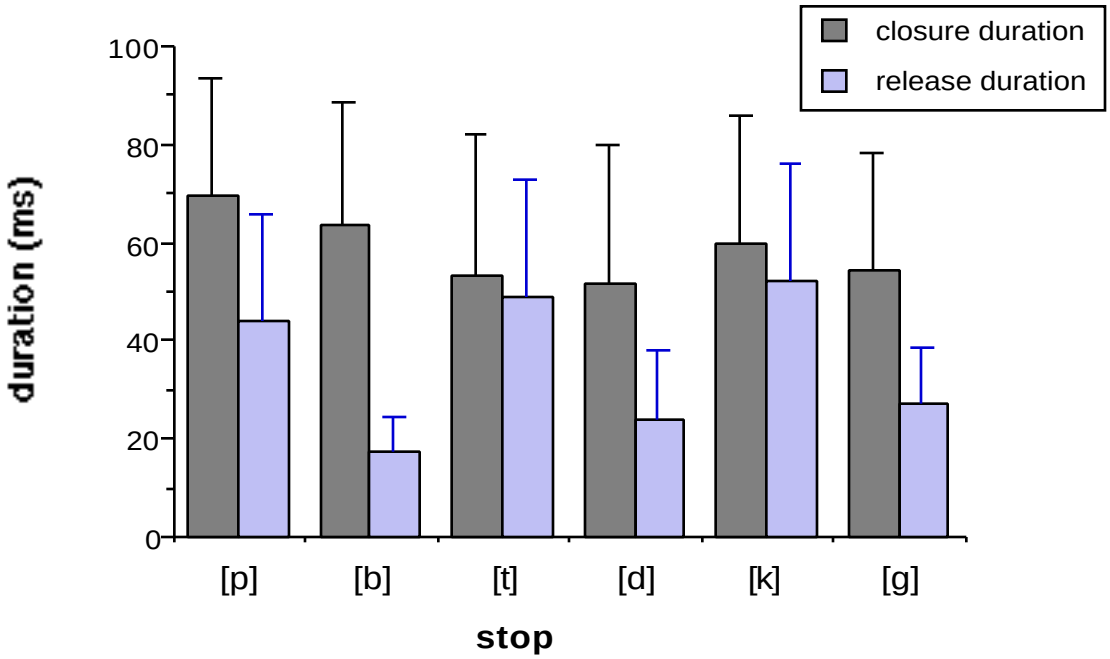

Figure — closure versus release duration by stop consonant identity.

Crystal and House (1988a) describe total stop (closure plus release) duration to be similar for alveolars and labials (about 80ms) and longer for velars (100ms). If we add the mean values for closure and release durations calculated from all the stops in TIMIT, regardless of their context, we find that the alveolar place at 92ms is still shorter than the bilabial and velar places. The labial place is the next longest with a value of 99ms, while the velar is longest at 104ms. While the value for the velar place is close to Crystal and House's (1988a) report, the bilabial and alveolar places are somewhat longer and do not follow the pattern suggested by Crystal and House of yielding approximately the same total duration.

C. Release Frequency in Sentence-Final Stops

In order to provide a standard environment in which the release of oral stops could be examined systematically using database searches, the transcription given for the sentence-final position of all (non-calibration) sentences was examined for the occurrence of released and unreleased oral stops.

1130 sentence-final stops were located in the search of the database. 9.1% were bilabial, 77.8% were alveolar, and 13.1% were velar. 37.8% of the stops were voiced, 62.2% were voiceless. The distribution of all six stops is as follows: [k], 11.5%; [t], 43%; [p], 7.7%; [g], 1.2%; [d], 34.8%; [b], 1.8%. A released stop occurred in 59.7% of the cases and an unreleased in 40.3% of the cases.

The place of articulation had a significant effect on whether a release occurred or not as determined by a contingency table analysis (χ^2=40.829, p=.0001). Bilabial stops were released 49.5% of the time, alveolar stops 57% of the time, and velar stops 83.11% of the time. Values for release frequency are presented graphically in Figure 5.

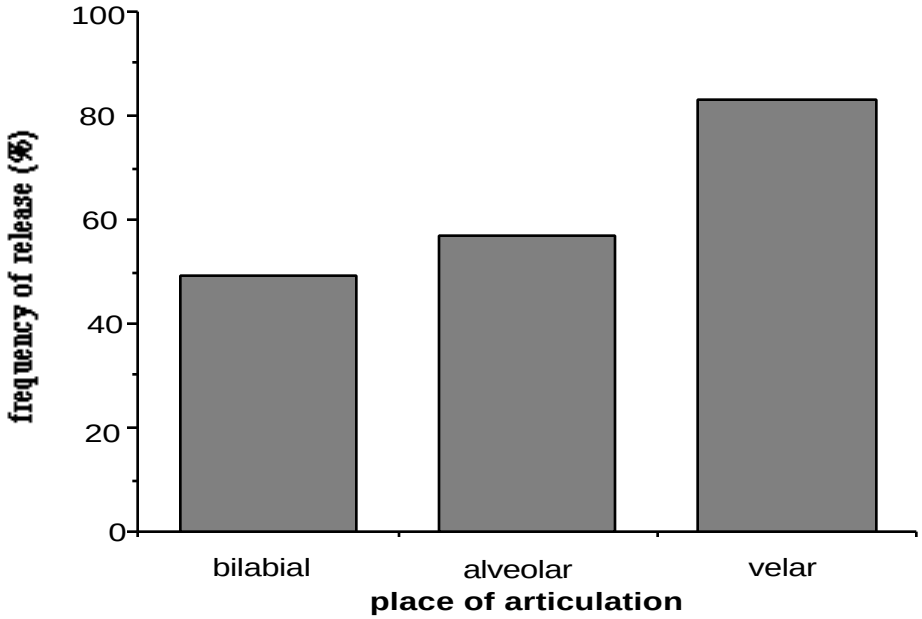

Figure 5—percent frequency of release by place of articulation for sentence-final stops; bilabial 49.5% of the time, alveolar 57% of the time, and velar stops 83.11% of the time.

There are presumably aerodynamic causes for these differences as the velar stops with a small volume of air behind occlusion might produce a more audible release than the alveolar and labial which have increasingly large volumes of air to absorb pressure produced by the pulmonic airstream. Recall however that bilabials also have a longer closure than velars during which pressure may build up, although cavity expansion might mitigate this.

Voicing, however, did not have an overall effect on whether a release occurred, despite the fact that voiced stop bursts are reported as having a lower amplitude than voiceless stop bursts (Halle, Hughes, and Radley, 1957). Voiced stops were released 59.5% of the time and voiceless stops 59.9% of the time. This is in close accordance with Crystal and House's (1988c) report of an overall release frequency of 59% for stop consonant across *all* sentence positions. However, in word-final position, Crystal and House (1988c) found only a 33% release frequency, but they note a problem with this figure due to inadequate diversity in their sample. Crystal and House (1988a) report a tendency for voiceless stops to include a release more often than voiced stops. As an overall effect, this was not evident in the data considered here. Crystal and House (1988c) also find a difference across all contexts in the release frequency between voiced and voiceless stops (65% vs. 33%). This difference is not supported in our data for sentence-final position. Crystal and House (1988a:1557) find "a tendency for voiceless stops to be completed a higher percentage of the time than voiced stops, particularly in word final position." When each place of articulation is tested independently in a contingency table analysis, the bilabial and alveolar place show no effect of voicing on whether a released or unreleased stop occurred. For the velar stops, however, voicing did have a significant effect on whether a released or unreleased stop occurred (χ=3.902, p=.0482). The voiced velars were

released in 64.3% of the cases while the voiceless velars were released 85.1% of the time. This fact supports the aerodynamic explanation offered above for the place effects as in American English the voiceless velars are the most likely to have the oral pressure build-up favoring audible stop releases.

Randolph (1989) provides information about stop release frequency as a function of syllable position, information not present in the commercially available TIMIT database. Randolph finds that "stops in the outer onset position are practically always released (97% of the time), whereas in the coda position, they are mostly unreleased (43% vs. 31% released). One also sees that large percentages of stops are released when they are followed by vowels and glides (a necessary but not sufficient condition for the stop belonging to the onset)." (Randolph, 1989, p. 115). Randolph's sub-study of stop realization across three databases included 12,161 tokens which were realized as follows: 7855 released, 2303 unreleased, 1052 flapped, 702 deleted, and 259 glottalized (Randolph, 1989).

Crystal and House (1988a) report a tendency in their data set for the velar consonants to be released more often than bilabial or alveolar consonants and state that this tendency is attributable to the behavior of the voiceless velar consonant. The results reported here for sentence-final stops are in accordance with that finding. A contingency table analysis of the effect of place on the occurrence of a release in which the voiceless velars are excluded yields no significant effect, although the direction of the trend remains unchanged from that shown in Figure 1. As only 18 voiced velar tokens remain in this analysis, a larger sample might raise the trend to significance even without voiceless velars. Another Crystal and House result reported in 1988(b), that "labials, particularly in unstressed syllables, [tend] to be completed [ie. released] more frequently than alveolars and velars" (p. 1580), was not supported by the TIMIT data considered here.

The sex of the talker had a significant effect on release frequency, with women releasing their sentence-final stops more often than men (χ^2=49.146, p=.0001). For a fuller description of these results, see Byrd (1992b). One of the calibration sentences which every speaker read ends in the common stop-final word "that." When the sentence-final position of this sentence is examined, it is found that a released [t] occurs 24% of the time, an unreleased [t] 67% of the time, and a glottal stop 9% of the time. As for sentence-final stops in the rest of the database, when a stop occurred, women released the stop significantly more often than men (χ^2=5.57, p=.0183).[1]

II. Affricates

The database (excluding calibration sentences) was searched for sequences of an alveolar stop followed by a label for an affricate release (which is transcribed differently than the phonetically parallel post-alveolar fricative). 952 voiceless affricates and 1103 voiced affricates were found. Mean closure and frication durations are shown in Table 4 below.

[1]Note that this result differs from that reported in the footnote of a preliminary study in Byrd (1992b); a more comprehensive study on this topic can be found in Byrd (1992a).

affricate	closure (ms)	s.d. (ms)	release (ms)	s. d. (ms)
{tʃ}	43	18	86	28
{dʒ}	43	19	62	28

Table 4—affricate closure and release durations and standard deviations in ms

ANOVA determines there to be a significant effect of voicing on release duration. No effect of voicing on closure duration is found, in accordance with the findings for [t] and [d] closures. When we compare the closures for the two affricates with the two stop closures at the alveolar place, we find that the affricates have shorter closure durations. This can be seen in Figure 6.

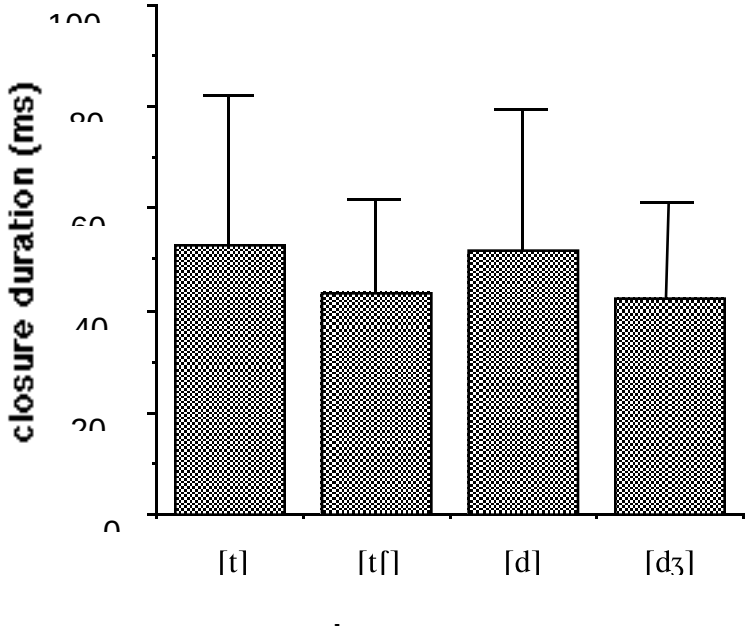

Figure 6: alveolar stop and affricate closure durations

ANOVA testing the four-level factor of closure shows there to be a significant effect ($F(3,15064)=62.389$, $p=.0001$) of this factor. A post-hoc Scheffé's-S test shows all pairwise comparisons except that of the two affricate closures to be significantly different at the $p<.001$ level. If we consider the total (closure+release) of the mean affricate and alveolar stop durations, we find that the affricates are still considerably longer than the stops, approximately 26ms.

III. Nasals

The TIMIT database includes labels for both syllabic and non-syllabic nasals at the bilabial, alveolar, and velar places of articulation. Syllabic consonants are the result of complete reduction of the vocalic syllable nucleus. A total of 18,101 nasals occur; 16,913 of them are non-syllabic. The frequency and durations of each of the nasals are shown in Table 5. (Note that nasal flaps are transcribed separately and not included below.) As the reader may be interested in the distributional frequency of these nasals without the addition of tokens from the two

calibration sentences which were read by all speakers, these values are given in parenthesis where different.

nasal	**number** (w/o calib. sen.)		**duration (ms)**	**s. d. (ms)**
m	5509	(4881)	62	26
n	9660	(8496)	55	23
ŋ	1744	(1568)	63	24
syllabic m	171		79	31
syllabic n	974	(846)	79	30
syllabic ŋ	43	(30)	81	24

Table 5— non-syllabic and syllabic nasal durations and standard deviations in milliseconds, and number of tokens; calibrations sentences are excluded from the number shown in parentheses

As reported in Byrd (1992a), a chi-square test determined there to be no effect of sex on the distributional frequency of syllabic [m̩] and [ŋ̩]. However, the sex of the speaker did have a significant effect on the frequency of [n̩] (χ^2=12.632, p=.0004). Women had significantly fewer [n̩]'s than the men. As reduction in the environment of alveolar consonants is a particularly common process, it is important to note that men and women appeared to produce this and only this syllabic consonant with different frequency.

The syllabic nasals were on average 21ms longer than the non-syllabic nasals. Based on the start and end time of each label, word positions were calculated for each phone as initial, medial, final or unaffiliated. Unaffiliated denotes a phone occurring outside the temporal extent of any word labels, such as might occur in a filled pause, or a phone which cannot be determined to belong to one of two abutting words. An ANOVA of the effects of nasal identity and word position, and their interaction, was calculated for the duration values for nasals in all TIMIT sentences (calibration sentences included). Nasal identity (F(5,18080)=8.636), p=.0001) and word position (F(3,18080)=4.070, p=.0067) both have a significant effect on duration. Post-hoc Scheffé's S-tests show the [n] to be significantly different in duration from all the other nasals, while [m] is different in duration from all but the non-syllabic velar nasal. The shorter duration of [n] parallels the findings for alveolar oral stop closures where it was suggested that the tongue tip was the most rapid articulator of the three, but this correspondence is not found in the syllabic case where there seems to be a strong tendency for these syllable nuclei to be realized with a consistent duration. None of the syllabic nasals were significantly different in duration from one another. The shorter duration of the alveolar relative to the other consonantal nasals does not carry over to the syllabic counterparts. The mean durations of the nasal stops in each word position are in Table 6 below. (Ten unaffiliated nasals are not included.)

word position	**number**	**duration (ms)**	**s.d. (ms)**
initial	3767	61	28
medial	7236	54	24
final	7088	64	24

Table 6—mean durations in milliseconds for nasal stops by word position; (ten unaffiliated nasals are not included.)

Post-hoc tests showed duration to be significantly different in each word position. (Unaffiliated nasals were not significantly different from any other position.) However, these values appear to vary quite a bit depending upon the phone in question. ANOVA also showed a significant

interaction of nasal identity with word position. Table 7 shows values for nasal duration at each word position for both the consonantal and syllabic nasals. (Note that initial syllabic velar nasals, of which there were only 15, were found in circumstances when a word phonemically beginning with a velar stop was nasalized initially due to a preceding nasal.)

nasal/position	word initial	word medial	word final
m	59	59	73
n	59	49	59
ŋ	88 (n=2)	54	65
syllabic m	80	77	81
syllabic n	84	75	80
syllabic ŋ	87	74 (n=3)	78

Table 7—nasal duration at each word position in milliseconds for both the consonantal and syllabic nasals

Initial and medial [m] are indistinct in duration, and relatively small differences exist between [m̩] in all positions.

IV. Flaps

Both oral and nasal flaps are transcribed in TIMIT. A total of 4980 flaps occurred; 3649 oral flaps and 1331 nasal flaps. Of these, 1557 were in one of the calibration sentences that included at least two potential flap sites--the word *water* and the phrase *suit in*. The analysis and results below were calculated both including and excluding the flaps from the calibration sentences. The results did not differ so data from all the flaps, including those in the calibration sentences, are presented below. The mean duration of flaps in the database is 29ms (s.d.=8ms). ANOVA testing for differences between oral and nasal flaps and between word positions shows there to be no effect of flap identity and a small but significant effect of word position ($F(2,4974)=7.238$, $p=.0007$). There is no interaction. Flaps increased in duration in 1ms steps from medial to final to initial. Zue and Laferriere state that "[a]s a phonetically defined group, flaps vary in duration from 10 to 40ms." (p. 1044) Their study with six speakers yielding 1484 flaps found mean durations of 26 to 27ms for oral flaps, with a range of 10 to 40ms. The range in the TIMIT data for flap duration was 9 to 73ms for oral flaps and 8 to 68ms for nasal flaps. Rimac and Smith found a mean flap duration of 36ms. The TIMIT mean is in close agreement with Crystal and House's (1988c) reported duration of 29ms and other like values reported in Fisher and Hirsch (1976) and Sharf (1962). Fox and Terbeek (1977) present mean durations of flaps in 40 words which appear from their graphical display to range from 21 to 33ms, but no overall mean is reported. Comparison of ranges should be done with caution as there is a high error rate in measuring the duration of flaps from spectrograms as noted by Fisher and Hirsch (1976). They state that flap duration is "on the order of 25 to 35 milliseconds and spectrographic measurements of voiced durations typically cannot be made with an accuracy greater than one period, or about ±5 milliseconds." (p. 191, also Klatt, 1971)

In the calibration sentences 99% of the speakers had a flap in *water* while only 19% had one at the word-final site. These two flaps were significantly different in duration ($F(1,743)=72.185$, $p=.0001$). The word-final flap had a mean duration of 33ms (s.d.=8ms), and the word-internal flap had a mean duration of 27ms (s.d.=6ms).

Effects of sex on the frequency of flaps are reported in Byrd (1992a) and Zue and Laferriere (1979). Women were found to have fewer flaps than men in both studies. In Byrd (1992a) chi-square tests found a significant effect of sex on the frequency of both oral and nasal flaps in the TIMIT data (χ^2=55.341, p=.0001 and χ^2=11.41, p=.0007, respectively). The women produced significantly fewer flaps than the men. A three-factor ANOVA with all two-way interactions testing for effects of speaker sex, flap identity and position on duration show there to be a no significant tendency for women to have shorter flaps than men (F(1,4970)=3.489, p=.0618). Women had a mean nasal flap duration of 28ms as compared to 29ms for the men. (The difference in oral flap durations was smaller (.6ms).) This is in spite of the fact that women spoke more slowly in the TIMIT calibration sentences (Byrd, 1992a). This may contrast with the finding by Zue and Laferriere (1979) of longer medial [t] and [d] durations for women in "seven out of eight cases" (p. 1047); they are not explicit about whether this includes the flaps. There was also a significant interaction of sex with word position (F(2,4970)=3.163, p=.0424). Any difference between male and female flap durations appears to occur in word final flaps which are approximately 2ms different. Flaps at the other positions have approximately the same duration.

V. Glottal Stops

While not phonemic in English, the glottal stop does occur in allophonic and stylistic variations, although little if any prior study has been done of its overall usage. 4834 glottal stops occur in TIMIT, including 1222 glottal stops from the calibration sentences. Ignoring occurrences in calibration sentences, the non-phonemic glottal stop of English occurred more frequently than the oral stop closures [p], [b], and [g]. Henton, Ladefoged, and Maddieson (1992) report that (phonemic) glottal stops generally have closure durations at least as long as other stops' closures. The glottal stops in TIMIT have a mean duration of 65ms (s.d.=32) making them longer than all the other oral stop closures except [p]. The glottal stop closure durations have a greater standard deviation than the oral stop closures as well. A post-hoc Scheffe's S-test determines that the glottal stop closure durations are significantly different from all the oral stop closure durations except those of [b].

When the word position of the glottal stop is considered we find that initial glottal stops made up 49% of the total, medial glottal stops 6% of the total, final glottal stops 16% of the total, and unaffiliated glottal stops 29% of the total. Note that when a glottal stop occurred between two vowels at a word boundary, it was considered unaffiliated. A two-factor ANOVA testing the effects of word position and speaker sex on glottal stop duration finds a significant effect of word position (F(3,4826)=84.745, p=.0001), no effect of sex, and a significant interaction of the two factors (F(3,4826)=4.554, p=.0034. In Figure 7 below, glottal stop duration is shown according to word position. The unaffiliated glottal stops are the longest, followed by the final glottal stops. The initial and medial glottal stops are the shortest and are the only pair not significantly different in a Scheffé's post-hoc test.

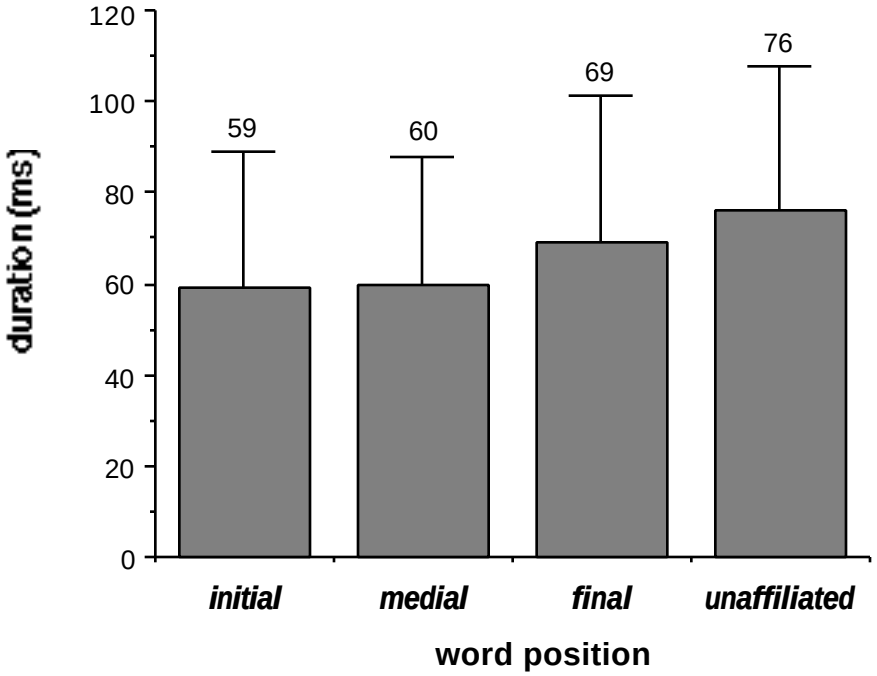

Figure 7—glottal stop duration by word position

The glottal stops produced by women had a longer mean duration by 2 to 4ms than those by the men in every position except medial where they were shorter by almost 7ms.

The effects of sex on the frequency distribution of glottal stop in TIMIT are reported in Byrd (1992a). In summary, a chi-square test shows that the women use significantly more glottal stops than the men, in every word position ($p \leq .0103$). When we consider the small set of 57 glottal stops produced in place of the sentence-final [t] of the word "that" in one calibration sentence, we find with a chi-square test that the production of a glottal stop in this position is not significantly influenced by sex, although the distribution favors the direction demonstrated above. It is somewhat unexpected to find this relationship of sex to the frequency of glottal stop. In fact, women's voices are often characterized as more breathy than men's, and glottal closure is related to creakiness in the voice quality. It may be that the glottal stop is used as a devoicing mechanism more often by women or that it participates in allophonic patterns which are less productive for the men.

VI. Conclusion

The use of a large, commercially available database in the study of segmental duration has been described. The material includes 630 American speakers from a range of geographical locations reading 2342 different sentences. The corpus yields approximately 54,000 stops making this description of stops very comprehensive in terms of quantity and context and speaker diversity. TIMIT is not as useful for the investigation of effects of specific sentential context or the variability found in the productions by a single speaker. The data presented above is intended to provide a broad picture of durational characteristics of American stops. It has also

offered results on the structure of stops, their distributional frequency, and certain effects of talker sex. Comparisons are made with findings previously discussed in the literature. A comparison of these findings with those reported in earlier studies is important as much of this work has used isolated words or words in carrier phrases and small groups of speakers. Such studies may not reflect the variability found in a larger population of the language's speakers and the limitation to carefully controlled test items may focus a speaker's attention on contrasts, thereby exaggerating them. The similarities and differences found in the TIMIT results inform us as to the extent to which we can generalize from experimental studies to corpuses including a great deal of variation in reading material and speakers. Of course an even broader scope of study would be provided by comparison of large read with non-read corpuses, something not now possible. As a large and diverse collection of labeled speech, TIMIT provides an interesting testing ground for the linguist to assess the accuracy of generalizations based on previous laboratory experimentation. In addition to being of academic interest, linguistic knowledge of reliable differences in segmental characteristics may help aid in the phonetic classification and speech recognition goals which TIMIT was designed to serve. The study presented above outlines many reliable segmental characteristics of American stops, nasals, affricates, flaps and glottal stops.

Acknowledgments

This research was supported by the National Science Foundation and the UCLA Department of Linguistics. Many thanks are due to Edward Flemming for designing and implementing the structure of a relational Macintosh database with which TIMIT is used at UCLA. I am grateful to Patricia Keating, Peter Ladefoged, and Ian Maddieson for their insightful comments on an earlier version of this paper.

References

Byrd, D. (1992a) Sex, dialects, and reduction. *Proceedings of the International Conference on Spoken Language Processing*.

Byrd, D. (1992b) Preliminary results on speaker-dependent variation in the TIMIT database. *JASA*, 92:593-596.

Chen, M. (1970) Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22, 129-159.

Crystal, T.H.; House, A.S. (1982) Segmental durations in connected-speech signals: Preliminary results. *JASA* 72(3), 705-716.

Crystal, T.H.; House, A.S. (1988a) Segmental durations in connected-speech signals: Current results. *JASA* 83(4), 1553-1573.

Crystal, T.H.; House, A.S. (1988b) Segmental durations in connected-speech signals: Syllabic stress. *JASA* 83(4), 1574-1585.

Crystal, T.H.; House, A.S. (1988c) The duration of American-English stop consonants: an overview. *J. Phon.* 16(3), 285-294.

Fischer-Jørgensen, E. (1964) Sound duration and place of articulation. *Z. Phonet. Sprachwiss. Kommunikationsforsch.* 17, 175-234.

Fisher, W.M.; Doddington, G.R.; Goudie-Marshall, K.M. (1986) The DARPA speech recognition research database: specifications and status. *Proceedings DARPA Speech Recognition Workshop*, 93-99, 1986.

Fisher, W.M.; Hirsch, I.J. (1976) Intervocalic flapping in English. *Chicago Linguist. Soc.* 12, 183-198.

Fox, R.A.; Terbeek, D. (1977) Dental flaps, vowel duration and rule ordering in American English. *J. Phonetics* 5, 27-34.

Halle, M., Hughes, G.W., and Radley, J.-P.A. (1957) Acoustic properties of stop consonants. *JASA* 29(1), 107-116.

Henton, C.; Ladefoged, P.; Maddieson, I. (1992) Stops in the World's Languages. *Phonetica* 49, 65-101.

Keating, P.A. (1984) Phonetic and phonological representations of consonant voicing. *Language* 60, 286-319.

Keating, P.A.; Blankenship, B.; Byrd, D.; Flemming, E.; Todaka, Y. (1992) Phonetic analyses of the TIMIT corpus of American English. to appear in the *Proceedings of the International Conference on Spoken Language Processing*.

Klatt, D.H. (1975) Voice onset time, frication, and aspiration in word-initial consonant clusters. *J. Speech Hearing Res.*, 18, 686-706.

Kuehn, D.P. Moll, K. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics* 4, 303-320.

Lamel, L.F., Kassel, R.H.; Seneff, S. (1986) Speech database development: design and analysis of the acoustic-phonetic corpus. *Proceedings DARPA Speech Recognition Workshop*, 100-109.

Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. *Language, 33*, 42-49.

Lisker, L.; Abramson, A.S. (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 394-422.

Lisker, L. (1972) Stop duration and voicing in English. In *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. (A. Valdman, ed.), Paris:Mouton, pp. 339-343.

Luce, P.A.; Charles-Luce, J. (1985) Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *JASA* , 78, 1949-1957.

Madebrink, R. (1955) The duration of the stops in the speech of deaf-mutes. *Folia Phoniatrica*, 7, 44-55.

Maddieson, I. (1984) *Patterns of Sounds.* Cambridge:Cambridge University Press.

Malécot, A. (1968) The force of articulation of American stops and fricatives as a function of position. *Phonetica*, 18, 95-102.

Ohala, J. (1983) The origin of sound patterns in vocal tract constraints. *The Production of Speech*. P. MacNeilage, ed., 189-216.

Pallet, D. (1990) Speech corpora and performance assessment in the DARPA SLS program. *Proceedings of the International Conference on Spoken Language Processing.*

Prosek, R.A.; House, A.S. (1975) Intraoral air pressure as a feedback cue in consonant production. *Journal of Speech and Hearing Research*, 18, 133-147.

Randolph, Mark. (1989) *Syllable-based Constraints on Properties of English Sounds.* MIT dissertation.

Scharf, D.J. (1962) Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech*, 23(3), 297-307.

Slis, I.H. (1970) Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica* 21,193-210.

Slis, I.H. (1971) Articulatory effort and its durational and electromyographic correlates. *Phonetica* 23, 171-188.

Smith, B.L. (1978) Temporal aspects of English speech production: a developmental perspective. *Journal of Phonetics* 6.

Subtelny, J.D.; Worth J.H.; Sakuda, M. (1966) Intraoral pressure and rate of flow during speech. *Journal of Speech and Hearing Research* 16, 397-420.

Suen, C.Y.; Beddoes, M.P. (1974) The silent interval of stop consonants. *Language and Speech*, 17, 126-134.

Umeda, N. (1977) Consonant duration in American English. *JASA* 61, 846-858.

Zue, V.W. (1976) *Acoustic Characteristics of Stop Consonants: A Controlled Study*. Sc.D. Thesis (MIT, Cambridge, MA).

Zue, V.W.; Laferriere, M. (1979) Acoustic study of medial /t,d/ in American English. *JASA*, 66, 1039-1050.

Zue, V.W.; Seneff, S. (1988) Transcription and alignment of the TIMIT database. *Proceedings of the Second Meeting on Advanced Man-Machine Interface through Spoken Language*.

Zue, V.W.; Seneff, S.; Glass, J. (1990) Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9, 351-356, 1990.