Stop! Don't stop! Epenthesis and Assimilation in Alveolar Clusters

James Dembowski
State University of New York at New Paltz
New Paltz, NY, USA
dembowsj@newpaltz.edu

ABSTRACT

Acoustic and articulatory variability in connected speech arise in part from phonetic processes such as epenthesis and assimilation. Some of these variations produce contrary effects. For example, speakers may introduce an epenthetic stop in nasal-fricative clusters such as /ns/, but when a stop is explicitly represented in a cluster such as /nts/, speakers may delete the stop through assimilation. The present study explores acoustic and kinematic evidence of stops in both /ns/ and /nts/ clusters. In contrast to some previous work, results fail to show an acoustic distinction between stops produced epenthetically (in /ns/ clusters) and stops that reflect an underlying linguistic representation (in /nts/ clusters). Neither the frequency of occurrence of acoustic features indicating stop production, nor the duration of closure, distinguished stops in these two contexts. Tongue blade movements for production of these clusters were highly variable across both speakers and contexts, indicating that speakers show remarkable flexibility in how they produce linguistically identical clusters in connected speech.

1. INTRODUCTION

One of the continuing challenges of speech production research is determining the relationship between linguistic units and the physical phenomena – acoustic, aerodynamic, kinematic – through which speakers manifest those units. The relationship is complicated by the fact that linguistic units are discrete, serially-ordered events, and the physical signals arising from speech production are not. Furthermore, even when certain physical features commonly characterize linguistic units (say, the transient aperiodic release of stop consonants), these features are not acoustically or kinematically invariant The relationship between linguistic units and physical speech production phenomena is still further complicated by the fact that even at the linguistic level alone, a variety of allophonic variations are permissible, contextual variations that often take the form of place, manner, or voicing assimilations. Despite these complications, listeners generally recover underlying phonemic "canonical" forms from contextually variable acoustic manifestations. Oddly, some of these allowable linguistic variations produce contrary phonetic effects. For example, in the homorganic consonant cluster /nts/, as in the word "prints," speakers commonly reduce or delete the /t/ through assimilation, leaving listeners to recover the underlying stop from a production that sounds more like /prɪns/ ("prince"). Conversely, in nasal-fricative clusters such as /ns/, as in the word "sense," speakers commonly introduce an epenthetic stop, producing an [nts] cluster, as in the word "cents." When variations such as the deletion or insertion of a stop occur in contrast to the presumed underlying linguistic intention, physical evidence of the underlying representation may remain. For example, a deleted final consonant may be indicated through the length of a preceding vowel. However, the literature is not clear on the extent to which underlying and epenthetic stops may be distinguished in the physical speech signal. Fourakis & Port [1] found that closure durations for [t] were longer for an underlying stop than for an epenthetic stop, but Blankenship [2] failed to replicate that in data from the TIMIT database. However, she did find that acoustic features for stops (closures or transient bursts) occurred more frequently when the stop was explicitly represented in /nts/ clusters than in the (epenthetic) context of /ns/ clusters. Warner & Weber [3] argue that Blankenship's failure to replicate the closure duration distinctions of Fourakis & Port resulted from the large variability in the TIMIT corpus.

The present study revisits these issues using acoustic and kinematic data from the University of Wisconsin X-Ray Microbeam (XRMB) Speech Production Database (SPD) [4]. Its goal is to explore acoustic and kinematic correlates of [t] in two types of alveolar consonant clusters: /nts/ clusters, where the /t/ is explicitly represented but is frequently assimilated, and /ns/ clusters, where /t/ is not explicitly represented but where [t] may be produced epenthetically. The study examines the likely occurrence of acoustic features for [t] in the two types of clusters, the spectro-temporal characteristics of those features, and tongue-blade movement for the cluster production. Results potentially provide insight into the degree to which linguistic representation influences the salience of physical features for stops.

2. METHODS

Data were drawn from five speakers randomly selected from the XRMB-SPD. The XRMB-SPD contains acoustic and kinematic data from 57 normal young adult speakers of

ISBN 1-876346-48-5

American English. Details of speaker characteristics, instrumentation, data collection procedures and data processing may be obtained from Westbury [4]. The acoustic signal was digitally recorded at 21.739 kHz with appropriate filtering. Kinematic signal streams represent the motions of small gold pellets glued to speech articulators: four along the midline of the tongue, one each at the midline vermilion border of each lip, and two on the jaw. Pellet positions were expressed relative to cranial axes defined by the speaker's maxillary occlusal plane (MaxOP) and central maxillary incisors (CMI). The only pellet movement analyzed for the current study was the ventral-most tongue pellet, located approximately 10 mm back from the extended tongue tip. For this analysis, the tongue movement was represented relative to the cranial coordinate system, and thus actually represents the combined movement of tongue and jaw. The decision to examine the tongue coupled to the jaw, rather than decoupled, was deliberate since the movement of interest was tongue movement relative to the palate (for production of alveolar clusters). (Decoupling the tongue pellets produces trajectories represented relative to a mandible-based coordinate system).

Analysis focused on /ns/ clusters in which speakers might produce an epenthetic [t], and /nts/ clusters in which /t/ is explicitly represented but might be assimilated in production. The /ns/ clusters came from the word "sense" produced five times by each speaker as an isolated word, and produced ten times in sentence contexts, for a total of 15 productions (in ideal circumstances wherein complete data were available from each speaker). Each speaker also produced the /ns/ cluster five times across a word boundary, in the phrase "...ofte**n s**pecial..." extracted from a complete sentence. The /nts/ clusters came from five repetitions of the word "ingredients," extracted from a sentence context, and supplemented by one production each of the words "prevents" and "haunts," also extracted from sentences. Ideally then, each speaker produced 20 repetitions of /ns/ clusters, and 7 repetitions of /nts/ clusters. Only one speaker (s61) had any missing data. This speaker lacked one production of "ingredients," and thus provided six /nts/ repetitions rather than seven.

Segment boundaries were derived from the acoustic signal, using the Windows-based TF32 version of CSpeech [5]. Criteria for segmentation were based on Olive, Greenwood & Coleman's [6] descriptions of acoustic landmarks of nasal-obstruent interactions. The acoustic waveform and spectrogram were displayed simultaneously. Cursors were placed at (1) the onset of the nasal, (2) the point of nasal offset and stop closure onset, (3) the point of closure offset and onset of aperiodic noise for the [t] and/or [s], and (4) the offset of [s]. From this segmentation, durations of the nasal, closure, and voiceless obstruent segments could be determined. At the time the segments were marked, the onset of obstruent noise following closure was judged as a "transient spike" or not, based on visual examination of the waveform and spectrogram. Specifically, if there was evidence of an abrupt amplitude spike in the waveform, or there was a clear, reasonably sharp vertical line in the spectrogram that appeared to extend across the frequency range, the obstruent noise onset was judged "transient." The presence of a transient implies the production of [t], while the lack of transient implies the onset of [s] without an initial [t]; however, we tried to avoid thinking in these "linguistic" terms. Instead, we tried to approach the phenomena under investigation purely as a set of acoustic landmarks. Thus, the interval between the nasal and the onset of obstruent noise could be considered to show one of four characteristics: (1) no closure or transient (i.e, an immediate transition from nasal to fricative); (2) closure but no transient; (3) transient but no closure; and (4) both closure and transient. In addition, spectral moments (mean, standard deviation, skewness and kurtosis of the frequency distribution) were determined at three points in time: (1) 25 ms prior to end of the nasal, (2) immediately following the onset of obstruent noise, and (3) 25 ms following onset of obstruent noise (i.e., at a point presumed to be well into [s] production). An acoustic aperiodic transient would be expected to have a broad, flat frequency spectrum, while a voiceless alveolar fricative would have high amplitude frequencies concentrated in the upper range of the spectrum. Therefore, obstruent onsets with transients might be expected to differ in their spectral moments from onsets without transients. Those with transients might be expected to show a lower mean, a wider standard deviation, and a more platykurtic distribution than those without. The acoustic signal, in summary, allowed visual judgments of the presence of closure and transient spikes which indicate oral stop production, duration measurements of nasal, closure and fricative segments, and spectral distribution characteristics of nasal and obstruent segments. Primary comparisons were between /ns/ clusters and /nts/ clusters, with the expectation that occurrence of closure and transient landmarks, duration of closure, and frequency distribution characteristics, might differ between those contexts in which [t] is only produced epenthetically, and those in which [t] is underlyingly represented. Finally, the trajectory of the ventral tongue blade pellet (presumed proximal to that part of the tongue which forms alveolar constrictions) was extracted and examined with a view toward determining whether it showed evidence of differences in behavior between the two types of clusters, as well as determining a preliminary look at intra- and inter-speaker kinematic variability in alveolar cluster production.

3. RESULTS

Frequency of occurrence of stop consonant features:

Table 1 summarizes the occurrence of acoustic features (no closure or transient, closure alone, transient alone, both closure and transient) as proportions of those features occurring out of the total number of contexts in which they were possible. The "combined" column shows the proportion of *all* stop consonant acoustic features combined (closure, transient, and closure plus transient, together). The table compares these occurrences in within-word /ns/ clusters, across-word /ns/ clusters, all /ns/

ISBN 1-876346-48-5

clusters (within- and across- word combined) and /nts/ clusters. The "range" row shows the range of proportions as they occurred across the five speakers. The table shows no substantial differences in the frequency of occurrence of acoustic features of stops in nasal-fricative clusters, regardless of whether the stop is explicitly underlying or not. Overall, 94% of all /ns/ clusters showed some evidence of stop production, and 97% of /nts/ clusters showed some evidence of stop production. This evidence took the form of closure alone, or closure plus transient. Transient bursts rarely occurred alone. This is unsurprising since the occurrence of a burst implies the release of orally impounded pressure, and the impounding of pressure implies closure. That is, a speaker can hardly produce a burst without closure, though the converse is possible. These data generally fail to agree with Blankenship's suggestion that closures and transients occur more frequently in /nts/ clusters than /ns/ clusters. These data similarly fail to agree with Blankenship's suggestion that acoustic features of stops occur more frequently when /ns/ clusters occur within words than when they occur across word boundaries. Olive et al. suggest that word boundaries are unlikely to have any effect on the occurrence of epenthetic stop features in normal connected speech, though they offer only illustrative examples rather than distributional evidence.

Contexts	Frequency of occurrence of closure & transient features				
	none	clos only	trans only	clos + trans	comb
w/i word /ns/	.08	.40	0	.52	.92
x- wrd /ns/	0	.56	0	.44	1.0
all /ns/	.06	.44	0	.50	.94
/nts/	.03	.32	.12	.53	.97
Range	.04 - .23	.15 - .78	.12 - .15	.22 - .85	.77 - 1.0

Table 1. Occurrence of stop consonant acoustic features.

Closure duration:

Table 2 below shows closure durations across all /ns/ contexts and all /nts/ contexts. Closure durations were similar in the two contexts, and statistical tests showed no significant difference. These results are similar to the closure duration results of Blankenship, who failed to find a difference in closure duration between the two contexts, and contrast with the results of Fourakis & Port, who did.

Closure duration (ms)	/ns/ contexts	/nts/ contexts
Mean	30.53	26.42
standard deviation	23.03	19.35
Range	0 - 129.35	0 - 85.05

Table 2. Closure durations in two nasal-obstruent contexts.

Spectral moments:

None of the moments of the frequency distributions varied notably across the two phonetic contexts of interest, and there's no reason why they should, considering that the frequency of occurrence of consonant acoustic features was approximately the same in both /ns/ and /nts/ contexts. However, the data were divided into utterance tokens that showed visual evidence of a transient burst at the onset of obstruent noise, and tokens that did not, in the expectation that these distinctions might be reflected in spectral moments. Space limitations preclude a detailed presentation of these results here. In general, the mean and kurtosis values varied in the expected directions. Burst distributions tended to show lower mean frequency values, and more platykurtic (flatter) distributions. The average first moment (mean frequency) of utterances with bursts was about 4.2 kHz, while the average mean frequency of utterances without bursts was about 5.8 kHz. These quantitative distinctions would appear to validate our visually-based judgments concerning the presence or absence of bursts.

Tongue blade movement:

Figure 1 below exemplifies tongue-blade pellet movement for three speakers each producing one token of the word "sense." The horizontal line along the top represents the distance along the palate from the central maxillary incisors, as it would be if the palate were flattened. The front of the face is toward the right. The trajectories below the horizontal line show the movements of the tongue pellets as they approach and retreat from the palate from the onset of /n/ to the conclusion of /s/. The /n/-onset is marked with the letter "N" and the /s/-offset is marked with the letter "S." The closed circle on each trajectory represents the acoustically marked onset of closure for each speaker, and the open circle represents the acoustically marked end of closure (onset of obstruent noise). At this stage of analysis no obvious kinematic marker for a [+stop] (vs. [-stop]) feature has emerged, however the kinematic analysis is very much preliminary at the time of this writing. Even so, certain characteristics in the trajectories are worth noting. All three speakers approach the palate over the course of the nasal, and the point of acoustically evident closure approximately corresponds to a peak in the trajectory (a point where the tongue pellet is nearest the palate). The tongue pellet descends and at the point of acoustically marked "release" (noise onset), the pellet tends to be further from the palate than at the point of closure. The tongue continues to descend over the course of the /s/ production, as much as 3 - 6 mm. Two of the speakers shown produce

ISBN 1-876346-48-5

roughly similar trajectories over the course of the /ns/ cluster in that the tongue first moves back and up to form a closure, and then moves down and forward to produce the fricative. One speaker moves in the opposite direction, however: *forward* and up to closure, and then down for /s/. Our data to date suggest that the backward-curving trajectory may be more common among speakers than the forward trajectory, but the sample at present is far too small to draw any generalizations. Within speakers, trajectories tend to be roughly similar within contexts; that is, each speaker seems to have a distinctive "habit" of articulating clusters that is fairly consistent, though it may be quite different from the habits of fellow speakers. However, speakers may produce clusters rather differently in different sentence contexts (due to the effects of coarticulation and/or speech rate?), and in rare instances may even atypically depart from his (or her) own habitual pattern for a particular cluster in a particular context. If there is one lesson to be drawn from the data to date, it's that speakers seem to have remarkable flexibility in how they choose to articulate linguistically identical consonant clusters in connected speech, and they certainly exploit that flexibility.

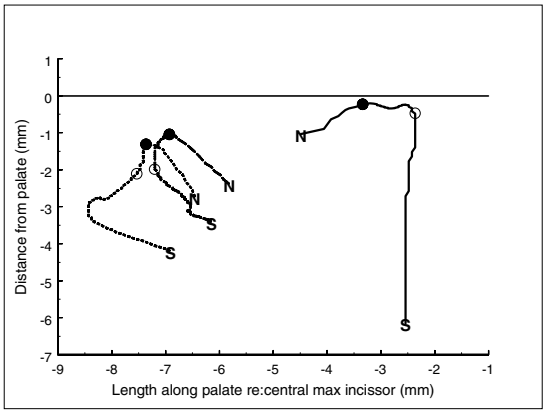

Figure 1. Tongue blade trajectories of three speakers producing /ns/ clusters.

4. CONCLUSIONS

In summary, the present investigation contrasts with at least some previous work in suggesting that there is no acoustic (and possibly no articulatory) distinction between stops produced epenthetically and stops that reflect an underlying linguistic representation. Also, the frequency of acoustically evident stops in /ns/ clusters is approximately the same whether the stop is linguistically represented within the cluster or not. Additionally, in connected speech, the presence of a word boundary within a consonant cluster does not appear to influence stop production. These observations suggest that the phenomena of stop production in homorganic nasal-stop-fricative clusters is chiefly physiologically driven, rather than linguistically driven, regardless of whether the stop is epenthetic or underlyingly represented. That is, the explicit linguistic representation of a stop in nasal-fricative clusters appears to have no substantial effect on its frequency of occurrence or its acoustic features. Acoustic and articulatory phenomena associated with stop production in the connected speech contexts studied here are highly variable, and the number of speakers examined to date is small. It's possible that distinctions between contexts in which stops are explicitly represented or not, if they exist, would emerge more obviously in better controlled "laboratory speech" utterances, or that such distinctions, if they exist in natural connected speech, require an unusually large number of speakers to be seen, because of the variability of connected speech. This study is ongoing; we hope to substantially increase the number of speakers analyzed, and look at kinematic signals in substantially greater detail.

ACKNOWLEDGMENTS

Ms. Gini Miller deserves recognition for days upon days spent efficiently, conscientiously, and cheerfully working at data extraction and segmentation for this project.

REFERENCES

[1] M. Fourakis and R. Port, "Stop epenthesis in English," *Journal of Phonetics*, vol. 14, pp. 197 – 221, 1986.

[2] B. Blankenship, "What TIMIT can tell us about epenthesis," *UCLA Working Papers in Phonetics*, vol. 81, pp. 17 – 25, 1992.

[3] N. Warner and A. Weber, "Perception of epenthetic stops," *Journal of Phonetics*, vol. 29, pp. 53 – 87, 2001.

[4] J. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook*, Madison, WI: U.W. X-Ray Microbeam Laboratory, 1994.

[5] P. Milenkovic and C. Read, *CspeechSP User's Manual*, Madison, WI, 2000.

[6] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach*, New York: Springer-Verlag, 1993.