ACOUSTICS CHARACTERISTICS OF OPEN TRANSITION IN NONNATIVE CONSONANT CLUSTER PRODUCTION

Colin Wilson[†] & Lisa Davidson[‡]

[†] Department of Cognitive Science, Johns Hopkins University, USA

[‡] Department of Linguistics, New York University, USA
colin@cogsci.jhu.edu, lisa.davidson@nyu.edu

ABSTRACT

In cross-language speech production, nonnative consonant clusters are often modified by epenthetic or transitional vocoids. Focusing on English speakers' productions of nonnative stop-initial clusters (e.g., /bn/, /bd/), the present study identified several acoustic characteristics that distinguish cases of epenthesis from accurate cluster realizations.

Both modified and accurate productions contained an open transition between the initial stop and following consonant, but epenthesis transitions had longer durations and other properties indicating greater vocal tract opening. Time course analyses revealed that the acoustic markers of epenthesis were present throughout the transition but became more pronounced following stop burst/frication.

The acoustic measures studied here are simple, largely automatic, and provide a means to supplement or replace hand-coding of cluster production accuracy with statistical classification. The combination of semi-automatic measurement and machine learning makes the phonetic study of nonnative cluster production more objective and scalable, and could be extended to the investigation of transitional vocoids more generally.

Keywords: cross-language production, consonant cluster, open transition, automatic phonetic analysis.

1. INTRODUCTION

Languages differ with respect to whether they allow word-initial consonant clusters consisting of an oral stop followed by another stop (e.g., /bn/, /bd/) [18]. In languages that have such clusters, such as Russian and Georgian, an open transition is typically found between the two consonants (e.g., [b°n], [b°d]). Open transitions result from gestural separation in cluster articulation, and contain acoustic-phonetic cues that are beneficial for accurate perception of initial stops [8, 12, 21].

When speakers of languages that lack initial stop-nasal (SN) and stop-stop (SS) clusters attempt to produce them, a variety of modifications of the target structures are observed [11, 19]. In this study, we focused on the modification type made most frequently by English speakers: epenthesis of a transitional vocoid between the two consonants. Previous analysis has established that such vocoids are distinct from intended schwas ([ə] or [ɨ]) in the same consonantal environments [11] (see also acoustic and articulatory [9] studies of epenthesis in nonnative fricative-initial clusters). However, the critical distinction between open transitions that do and do not contain vocoids has so far been made with qualitative criteria only. Moreover, it has been suggest that epenthesis in nonnative productions consists entirely of gestural separation [11]. If both epenthetic and target realizations of stop-initial clusters involve limited overlap of the two consonant constriction gestures, how can epenthesis modifications or 'errors' be differentiated from accurate productions?

We identified four quantitative and easily-measured acoustic properties of open transitions that covary with previous qualitative judgments of epenthesis vs. accurate cluster productions: duration, zero-crossing rate, pitch pulse count, and intensity. Our findings are consistent with claims that epenthesis involves an interval of greater vocal tract opening relative to target realizations [11], and provide an objective basis for classifying differing species of open transition. More generally, this research contributes to the study of transitional elements in native and nonnative consonant clusters, and to the application of (semi-)automatic coding and statistical classification in the analysis of speech production.

2. CLUSTER PRODUCTION STUDY

Data was taken from a previous study of nonnative consonant cluster production [31]. English speakers (N=24) heard and repeated isolated [C°CáCV] nonce words recorded by a native Russian speaker. The initial consonant clusters of interest were composed of stops (/p t k b d g/) followed by het-

erorganic nasals and stops (/m n p t b d/). Both voiceless and voiced stops were followed by nasals; stop-stop clusters agreed in voicing. The Russian clusters were digitally manipulated to systematically vary transition duration, transition intensity, and (for voiced stops only) intense prevoicing at initial consonant onset (see [31] for details). For purposes of this paper, we collapsed across these manipulations in order to give a maximally general characterization of the acoustic-phonetic properties of nonnative open transitions.

Waveforms and spectrograms of the English productions were inspected by several coders to determine whether any modifications were made relative to the Russian target clusters, following established coding guidelines [11] (see also [1]). In particular, responses were coded as containing epenthesis if the stop burst was followed by vocalic material that had visible first and second formants and was higher in intensity than the following closure. Epenthesis was by far the most common type of modification, occurring in 34% of all responses (voiceless SN: 27%, SS: 16%; voiced SN: 48%, SS: 45%).

2.1. Measurements

As noted in the introduction, four measures were investigated as potentially covarying with the qualitative response coding of [31]. One simple hypothesis is that, while both epenthesis and accurate productions involve gestural separation, the degree of separation is greater in the former. This would likely result in epenthesis responses having longer average **transition duration**. Another, compatible hypothesis is that epenthesis responses contain an interval of greater vocal tract opening (i.e., greater channel area at the point of smallest constriction) [11]. If the degree of opening exceeds the area that is required to maintain turbulence (cf. frication/aspiration intervals), epenthesis responses should contain relatively less aperiodic energy. **Zero-crossing rate** (ZCR) is an easily-calculated index of aperiodic content that has been widely used in automatic systems to identify frication [2] (and for voiceless vs. voiced classification [13]). Lower ZCR values in epenthesis responses, relative to accurate cluster productions, would be consistent with greater vocal tract opening. A sufficiently open vocal tract could also encourage spontaneous voicing, resulting in more detectable **pitch pulses** [16], and could also lead to higher **intensity** (as in the well-known relationship between aperture and intensity in full vowels, [23]).

Measurements were taken over the entire transition (burst, frication/aspiration, and following vocoid if present), beginning at the release of the word-initial stop and ending at the onset of the following consonant closure, using standard functions of Praat [3]. This method depends on prior demarcation of the surround consonant closures, and therefore is not fully automatic. However, it avoids the need for boundary placement or other hand-coding within the transition itself.

2.2. Results

Table 1 provides means (and standard deviations) for total transition duration (ms), ZCR (crossings / ms), number of detected pitch pulses, and average intensity (dB) across all of the production responses that were coded as containing a transitional vocoid (epenthesis) or as having no modification within the consonant transition (accurate).

Table 1: Means (and standard deviations) for all productions separated by voicing of the initial stop (vcl = voiceless, vcd = voiced) and response type.

measure	S voice	epenthesis	accurate
duration	vcl	62.96 (17.54)	44.57 (23.07)
	vcd	57.74 (22.29)	25.69 (15.95)
ZCR	vcl	2.34 (1.40)	4.39 (2.88)
	vcd	1.31 (0.83)	2.77 (2.15)
pulses	vcl	6.06 (3.23)	0.78 (2.11)
	vcd	6.62 (3.93)	0.90 (1.60)
intensity	vcl	56.94 (4.35)	42.47 (6.99)
	vcd	59.45 (4.85)	44.98 (8.52)

Differences between epenthesis and accurate responses on all measures, apparent in Table 1, were supported by separate mixed-effects linear regressions. The binary fixed factors of response type (epenthesis vs. accurate) and target voicing (voiceless vs. voiced) were scaled to have means of 0 and unit standard deviations [17], and random effect structures for participants and items were maximal. The main result was that, relative to responses coded as accurate, epenthesis responses had longer transition duration (β = 22.38, *se* = 1.93, *t* = 11.58), lower ZCR (β = -1.55, *se* = 0.20, *t* = -7.87), more pulses (β = 4.85, *se* = 0.48, *t* = 10.07), and greater intensity (β = 13.94, *se* = 0.61, *t* = 22.91). This pattern is consistent with the general claim of [31] (and other studies of nonnative cluster production) that different types of open transition can be distinguished acoustically, and with the more specific claim that epenthesis responses involve greater vocal tract opening.

The stop voicing factor also contributed significantly to the regression, with voiceless stops having longer duration (β = 8.74, *se* = 1.96, *t* = 4.46), higher

ZCR (β = 1.22, *se* = 0.23, *t* = 5.36), fewer pulses (β = -0.70, *se* = 0.15, *t* = -4.54), and lower intensity (β = -2.55, *se* = 0.53, *t* = -4.78). These effects (except perhaps for lower intensity) would be expected from the native allophonic pattern of English, according to which word-initial voiceless stops are aspirated.

Response type and voicing participated in two significant interactions: the duration difference between accurate and epenthesis responses was smaller for clusters beginning with voiceless stops (β = -15.47, *se* = 2.81, *t* = -5.51), and similarly the difference in pulses count was also smaller (β = -0.87, *se* = 0.27, *t* = -3.20). These interactions suggest a trade-off between aspiration and transitional vocoid parallel to that found between aspiration and schwa in native forms such as *potato* [10]. A post-hoc analysis of the voiceless stop-initial clusters confirmed that the response effect on duration was significant for this subset of the data in spite of the interaction (β = 15.57, *se* = 2.61, *t* = 5.98).

3. CLASSIFICATION OF OPEN TRANSITIONS

The findings of Section 2 suggest that a constellation of simple quantitative acoustic properties may suffice to identify epenthesis in nonnative productions of SN and SS clusters. We explored this possibility with two additional analyses, both of which involve statistical classification of response type on the basis of transition acoustics alone.

3.1. Random forest model

The first analysis employed random forests [4, 6], a sampling-based method that estimates how much each potential predictor contributes to a classification decision and how successfully the learned classifier can be expected to generalize beyond the training data (here, to hypothetical new instances of SN and SS cluster production by English speakers). The training cases consisted of the same productions analyzed above. Each case was labeled with its response category (accurate vs. epenthesis) as coded in [31]. The random forest model attempted to predict this binary categorization with the values for transition duration, ZCR, pulse count, and intensity.

All four acoustic measures contributed substantially to classification, with estimated importance — calculated by comparing classification accuracy with original and permuted predictor values — being highest for the intensity measure (intensity: 140.24, pulses: 88.81, duration: 63.22, ZCR: 19.92; cf. a random numerical predictor with no known a priori relation to response type had an importance of only 0.13). Most significantly, the estimated error on new data was very low (4.34% total: 3.42% for epenthesis cases, 5.54% for accurate responses). According to this estimate, the classifier would agree with human coding on more than 95% of the data in future experiments on English stop-initial cluster production.

3.2. Logistic regression and cross-validation

The second classification analysis used binary logistic regression and assessed generalizability with cross-validation [5]. We were particularly interested in generalization to new English speakers, and therefore adopted the following approach. For each of the 24 participants, a binary logistic regression was first fit with that participant's data excluded ('held out') and then evaluated for classification accuracy on the held-out data. Each regression predicted the binary response category (accurate vs. epenthesis) using only fixed effects of duration and the other acoustic measures.

Consistent with the random forest analysis, error rates on held-out participants were low for the most part (mean: 5.74%, range: 0.0% – 22.43%). Prediction error exceeded 10% for only three participants and exceed 20% for one. The chosen acoustic properties may thus suffice to identify epenthesis in nonnative cluster productions of most, but perhaps not all, English speakers. Relatively poor performance on a few participants may indicate individual differences in the articulation of nonnative clusters, a possibility worth investigating in future work.

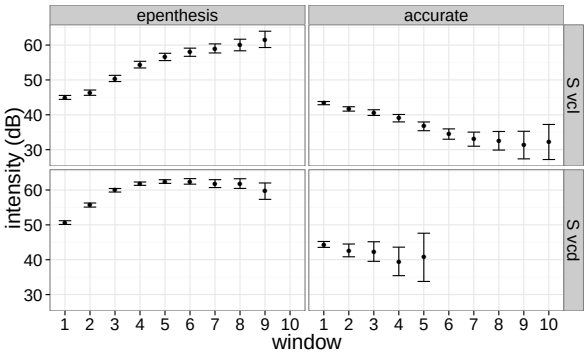
4. TIME COURSE ANALYSIS

If epenthesis responses contain an interval of greater vocal tract opening, this interval should occur later in the transition (i.e., after the stop burst and any following frication/aspiration). However, the preceding results could not provide information about the time course of the difference between epenthesis and accurate responses, as acoustic measurements were taken over the entire open transition. Therefore, we conducted a further analysis in which all of the measures other than duration were calculated in 20 ms analysis windows (10 ms frame shift) across each transition. Similar results were obtained with more fine-grained temporal analyses using windows of 10 and 5 ms (with half-window frame shifts).

Figure 1 shows that intensity generally rose over the course of the open transition in epenthesis responses while remaining flat or even falling during the transition of accurate responses. Fewer windows were plotted for accurate productions of clus-

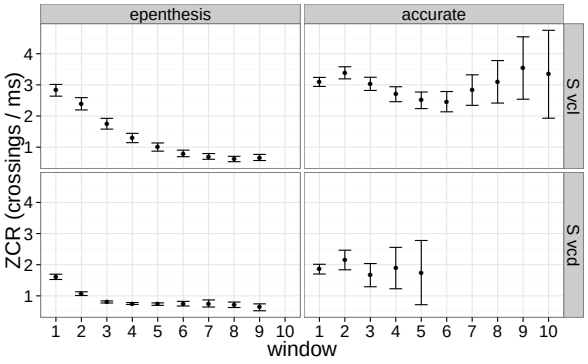
ters beginning with voiced stops because the relevant open transitions are typically short (see Table 1); the paucity of tokens made estimation of acoustic characteristics unreliable in later windows.

Figure 1: Means and confidence intervals ($\pm$ 2*se*) for intensity over time in open transitions.



As expected, the ZCR measure showed the opposite temporal pattern, with zero-crossings becoming less frequent in the later stages of epenthesis open transitions (Figure 2). The U-shaped pattern observed in accurate productions of voiceless stop-initial clusters could reflect a partially devoiced nasal in SN clusters, a possibility that could be investigated by disaggregate the time course data according to cluster type (SN vs. SS).

Figure 2: Means and confidence intervals ($\pm$ 2*se*) for ZCR over time in open transitions.



Separate mixed-effects linear regressions of intensity and ZCR in three regions of the open transition (early: windows 1–3, middle: windows 4–6, late: windows 7–9) confirmed strong separation of accurate and epenthesis responses near the end of the transition. Due to the lack of a late region in most of the accurate voiced stop productions, these analyses were limited to clusters beginning with voiceless stops. Epenthesis responses had higher intensity than accurate responses (β = 18.15, *se* = 1.08, *t* = 16.82), and crucially this difference became greater in the late region (β = 10.70, *se* = 0.78, *t* = 13.93). Responses containing epenthesis had lower ZCR (β = -1.41, *se* = 0.23, *t* = -6.10), and again there was a significant interaction between response type and region at the end of the transition (β = -0.65, *se* = 0.28, *t* = -2.29) indicating a relatively reduced ZCR in the last window of epenthesis transitions.

No clear time course pattern emerged for the pitch pulse measure. Given that detected pitch pulses are rare in the transitions overall, a longer temporal integration window may be required to observe strong differences on this measure.

5. CONCLUSION

This paper has made two main contributions to the study of nonnative consonant cluster production. First, it identified several acoustic properties that differentiate productions of stop-initial clusters with and without transitional vocoids. These properties are easily measured and, taken together, are consistent with the claim that epenthesis involves (i) greater separation of the consonant constriction gestures and (ii) an interval of greater vocal tract opening near the end of the transition. Second, the classification results reported here suggest that semi-automatic measurement of acoustic properties can supplement or possibly replace qualitative and time-consuming human coding, fostering more objective analyses and more rapid research progress.

The present findings suggest several directions for further study. Our claims about the gestural organization of accurate vs. epenthesis productions of nonnative stop-initial clusters could be evaluated articulatorily, complementing previous studies of fricative-initial sequences [9]. It will be interesting to investigate whether similar acoustic properties distinguish accurate and epenthesis responses in productions of other nonnative word-initial [7, 11, 14] and word-medial [15, 20, 25] consonant clusters. The characterization of native transitional vocoids [16, 26, 27] may also benefit from the measurement and statistical methods employed here. More generally, automatic phonetic analysis is becoming increasingly common [22, 24, 29, 30], and studies such as this one motivate further development of machine segmentation and classification techniques for phonetic analysis [28, 32, 33].

6. ACKNOWLEDGMENTS

We would like to thank Eleanor Chodroff and Sean Martin for helpful questions and comments on the

material presented here. This research was partially supported by NSF grants BCS-1052784 to Colin Wilson and BCS-1052855 to Lisa Davidson.

7. REFERENCES

[1] Beckman, M. 1996. When is a syllable not a syllable. In: Otake, T., Cutler, A., (eds), *Phonological structure and language processing: Cross-linguistic studies*. New York: Walter de Gruyter 95–123.

[2] Bitar, N. N., Espy-Wilson, C. Y. 1996. Knowledge-based parameters for HMM speech recognition. *Proceeedings of ICASSP* 29–32.

[3] Boersma, P., Weenink, D. 2013. Praat: doing phonetics by computer [Computer Program]. Version 5.3.53, retrieved 2 July 2014 from http://www.praat.org/.

[4] Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5–32.

[5] Browne, M. W. 2000. Cross-validation methods. *Journal of Mathematical Psychology* 44(1), 108–132.

[6] Bürki, A., Gaskell, M. G. 2012. Lexical representation of schwa words: Two mackerels, but only one salami. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38(3), 617–631.

[7] Carlisle, R. S. 1998. The acquisition of onsets in a markedness relationship. *Studies in Second Language Acquisition* 20(02), 245–260.

[8] Chitoran, I., Goldstein, L., Byrd, D. 2002. Gestural overlap and recoverability: Articulatory evidence from Georgian. In: Gussenhoven, C., Warner, N., (eds), *Laboratory Phonology VII*. Berlin: Walter de Gruyter 419–448.

[9] Davidson, L. 2005. Addressing phonological questions with ultrasound. *Clinical Linguistics & Phonetics* 19(6-7), 619–633.

[10] Davidson, L. 2006. Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica* 63(2-3), 79–112.

[11] Davidson, L. 2010. Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics* 38(2), 272–288.

[12] Davidson, L., Shaw, J. A. 2012. Sources of illusion in consonant cluster perception. *Journal of Phonetics* 40(2), 234–248.

[13] Deng, L., O'Shaughnessy, D. 2003. *Speech processing: a dynamic and optimization-oriented approach*. New York: Marcel Dekker.

[14] Eckman, F. R., Iverson, G. K. 1993. Sonority and markedness among onset clusters in the interlanguage of ESL learners. *Second Language Research* 9(3), 234–252.

[15] Funatsu, S., Fujimoto, M. 2012. Mechanisms of vowel epenthesis in consonant clusters: an EMA study. *Proceedings of the Acoustics 2012 Nantes Conference* 341–346.

[16] Gafos, A. I. 2002. A grammar of gestural coordination. *Natural Language & Linguistic Theory* 20(2), 269–337.

[17] Gelman, A., Hill, J. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

[18] Greenberg, J. H. 1965. Some generalizations concerning initial and final consonant sequences. *Linguistics* 3(18), 5–34.

[19] Haunz, C. 2007. *Factors in on-line loanword adaptation*. PhD thesis University of Edinburgh.

[20] Hwang, J. 2011. *Non-native Perception and Production of Foreign Sequences*. PhD thesis Stony Brook University.

[21] Kochetov, A., Goldstein, L. 2001. Competing recoverability factors and inter-gestural phasing in Russian stop clusters. *Poster presented at the Annual Meeting of the Linguistic Society of America*. January 4–7, Washington, D.C.

[22] Labov, W., Rosenfelder, I., Fruehwald, J. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1), 30–65.

[23] Lehiste, I. 1976. Suprasegmental features of speech. In: Lass, N. J., (ed), *Contemporary issues in experimental phonetics*. New York: Academic Press 225–239.

[24] Milne, P. M. 2011. Finding schwa: Comparing the results of an automatic aligner with human judgments when identifying schwa in a corpus of spoken French. *Canadian Acoustics* 39(3), 190–191.

[25] Nogita, A., Fan, Y. 2012. Not vowel epenthesis: Mandarin and Japanese ESL learners' production of English consonant clusters. *Working Papers of the Linguistics Circle* 22(1), 1–26.

[26] Ridouane, R. 2008. Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* 25(02), 321–359.

[27] Ridouane, R., Fougeron, C. 2011. Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology* 2(2), 275–300.

[28] Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., Yuan, J. 2014. FAVE (Forced Alignment and Vowel Extraction) [Computer Program]. Version 1.2.

[29] Ryant, N., Yuan, J., Liberman, M. 2013. Automating phonetic measurement: The case of voice onset time. *Proceedings of Meetings on Acoustics* volume 19 1–9.

[30] Sonderegger, M., Keshet, J. 2012. Automatic measurement of voice onset time using discriminative structured prediction. *The Journal of the Acoustical Society of America* 132(6), 3965–3979.

[31] Wilson, C., Davidson, L., Martin, S. 2014. Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language* 77, 1–24.

[32] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics 2008* 5687–5690.

[33] Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., Wang, W. 2013. Automatic phonetic segmentation using boundary models. *Proceedings of INTERSPEECH* 2306–2310.