Unpredicted vowel landmarks in TIMIT

Suyeon Yun
Chungnam National University

1 Introduction

This study examines the unexpected insertion of vowels that occurs in a read speech corpus of American English, which has not been reported in the literature. Vowel epenthesis, or vowel insertion, is a cross-linguistically common phonological process. In many cases, a vowel is epenthesized to meet a structural requirement by repairing a marked phonological structure in the language (Hall 2011). In English plural formation, for example, a vowel is inserted before the plural suffix /-z/ when the preceding noun ends in a sibilant consonant, to avoid a sequence of two sibilants, e.g., *buses:* /bʌs-z/ → [bʌsəz]. While this type of morphophonological vowel epenthesis in English is well-known, another type of vowel insertion not derived from a structural motivation is also reported for some dialects of English. Wright (1905) provides several examples of schwa insertion in dialects of English, which takes place between two consonants in the cluster, whose first member is usually *r* or *l*. For example, in those dialects, *arm* and *film* may be pronounced as [arəm] and [filəm], respectively, with a schwa inserted between the two consonants at the end of the word. This type of vowel insertion differs from the structurally-driven vowel insertion such as the one in the English plural formation, because the word-final clusters /rm/ and /lm/ in these example words are legal in English phonology and the inserted vowel does not fix any illegal structure. The inserted schwa in these dialects should rather be considered an excrescent (Levin 1987) or intrusive (Hall 2006) vowel. Excrescent vowels tend to be analyzed in the literature as a phonetic transition between consonant articulations, not a full vowel inserted by a phonological process. From a gestural perspective, the excrescent vowel is part of the gesture of the neighboring lexical vowel but appears like a separate vowel as a result of retiming of the adjacent consonant gestures (Steriade 1990, Browman and Goldstein 1992). The canonical features of excrescent vowels include that they (i) are subject to variation, (ii) have centralized vowel quality, (iii) are short in duration, (iv) occur in heterorganic clusters, and (v) are ignored in phonological processes such as stress assignment (Hall 2006). To my knowledge, no example of excrescent vowels in English has been reported in the literature other than Wright (1905). However, as excrescent vowels might not be recognized by the native speakers, it might be the case that they actually occur more prevalently than expected in spoken utterances of English.

This study is a part of a larger research project on the excrescent vowels in American English, which plans to cover multiple speech corpora. The current research addresses the following questions:

If there are excrescent/intrusive vowels in American English,
(i) In which environments do they occur?
(ii) What are their phonetic characteristics?
(iii) Do they differ from lexically epenthesized vowels, phonetically and phonologically, and how?
(iv) What phonological or pragmatic functions do they perform?
(v) Are there any sociolinguistic factors that affect their occurrences?

As a starting point, this paper investigates the cases of vowel intrusion in one read speech corpus of American English.

2 Methods

2.1 *Database* The speech corpus used is the Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus of read speech (Garofolo et al. 1993). The whole corpus contains 6,300 utterances produced by 630 American English speakers from eight different dialect regions. The corpus also provides time-aligned orthographic, phonetic and word transcriptions along with the sound files. The current study employs part of the

* I am grateful to Stefanie Shattuck-Hufnagel and Jeung-Yoon Choi for their helpful feedback on this project. I also thank audience at the 6th Asian Junior Linguistics Conference for their valuable questions and comments.

TIMIT corpus used in Yun et al. (2020), which consists of part of the Training Set (1 male and 1 female speaker selected from each dialect, each reading ten sentences) and the Core Test Set (2 male and 1 female speakers from each dialect, each reading eight sentences). In total, 352 utterances from 40 speakers (24 male and 16 female) are analyzed.

2.2 *Landmark annotations* The part of the TIMIT corpus was annotated based on the feature-cue-based framework developed by MIT Speech Communication Group (Huilgol et al. 2019), as illustrated in Figure 1 below. The framework labels speech with acoustic cues to distinctive features, which are divided into *landmarks* and *other acoustic cues*. A landmark refers to an abrupt acoustic change in speech signal, from which the information of distinctive features can be accessed (Stevens 2002). There are eight types of landmarks used in this framework: V (vowel), G (glide), Nc (nasal consonant closure), Nr (nasal consonant release), Fc (fricative consonant closure), Fr (fricative consonant release), Sc (stop consonant closure), and Sr (stop consonant release). Figure 1 shows a spectrogram of the word *modeling*, landmark-labeled in the Praat TextGrid (Boersma and Weenink 2022). The first and second tiers display the word and phone transcriptions provided by TIMIT, which are only roughly aligned with the speech signal. The third tier is practically the first tier of the feature-cue-based framework; this first tier, LM, contains the landmark labels. For instance, the vowel landmark V, circled in red, is located in the LM tier at the amplitude maximum of each vowel. Not all landmarks are realized as expected, however, and any modifications to the landmarks, deletion or insertion in particular, are notated in the next tier, LMmods. If an expected landmark is deleted, it is labeled with '-x', and if an unexpected landmark appears, it is labeled with '-+' in the LMmods tier. In Figure 1, we see that the phoneme /d/ and the following unstressed vowel are not produced, and the landmarks for those phonemes disappear in the LM tier but are labeled with '-x' in the LMmods tier, as circled in blue. There are four additional tiers for other acoustic cues below LMmods: vgplace, cplace, nasal and glottal. For the detailed information about the landmark labeling conventions, see Huilgol et al. (2019).

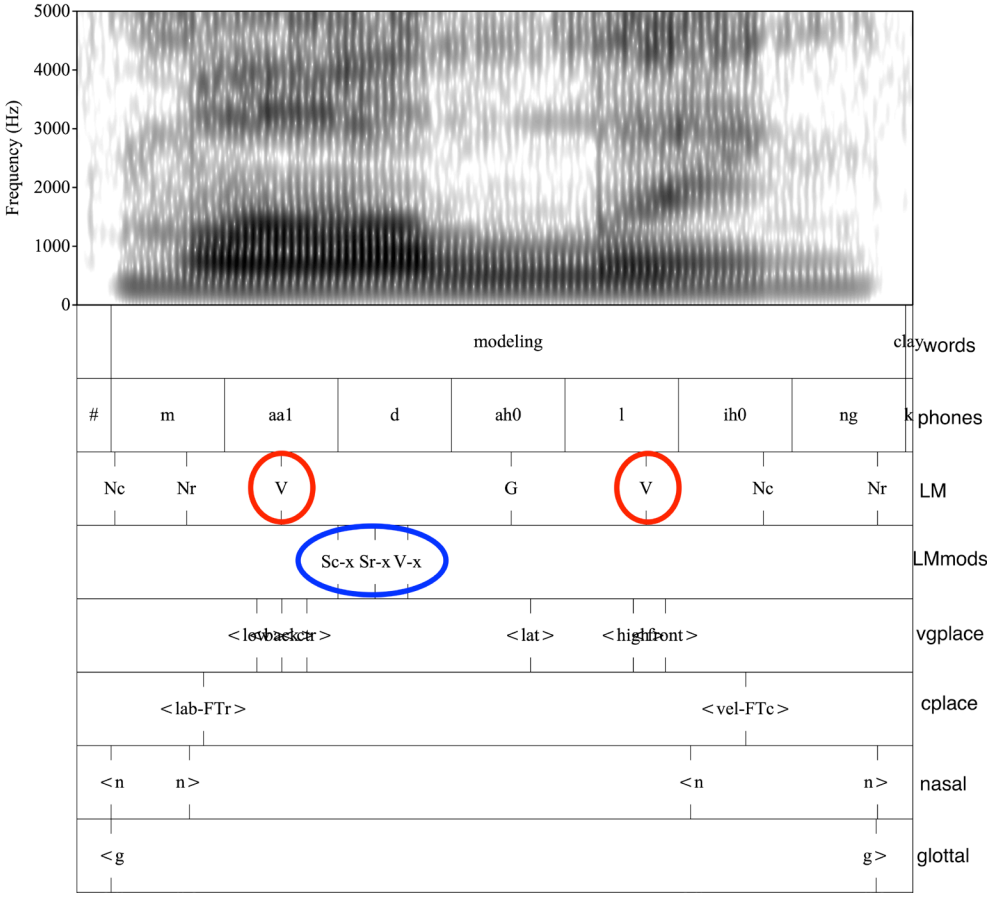

Figure 1: Labeled example of the spoken word *modeling* in TIMIT

The labels of predicted landmarks and other acoustic cues were automatically generated from the phonemic transcriptions. Student RA labelers and the author placed each label at exact position in the speech signal and added any modifications to the LMmods tier by inspecting the sounds auditorily and visually using Praat. All the

labeled TextGrids were reviewed and corrected by the author.

Of interest here is the vowels unexpectedly inserted in the spoken utterances. In landmark-based terms, a landmark for vowel, i.e., V, is inserted, which is labeled with '-+' indicating insertion, i.e., V-+, in the LMmods tier. For the current purpose, all labels in the LMmods tier are extracted using a Praat script, and the TextGrids including a V-+ are examined with the corresponding sound files.

3 Results

Only 7 unpredicted vowel landmarks are observed out of 352 utterances in the database. Table 1 lists the read sentences where each unpredicted vowel landmark appears along with the dialect and gender information of the speaker of each utterance. The vowels are observed in five dialect regions, and no speakers produce more than one inserted vowel.

Table 1. Contexts of unpredicted vowel landmarks in TIMIT (The words that contain the vowel landmarks are in bold, and the superscript ᵊ indicates the inserted vowel landmark)

Type	Dialect[1]	Gender	Sentence
1	DR2	M	Rob sat by the **pond**ᵊ and sketched the stray geese.
	DR4	M	Planned **parenthood**ᵊ organizations promote birth control.
	DR7	F	They own a **big**ᵊ house in the remote countryside.
	DR7	M	The nearest **synagogue**ᵊ may not be within walking distance.
2	DR2	M	But in this one section we **welcomed**ᵊ auditors.
	DR5	F	In tradition and in poetry the marriage bed is a place of unity **and**ᵊ harmony.
3	DR6	M	Before Thursday's **ex**ᵊ**am** review every formula.

Although small in number, the observed unpredicted vowel landmarks can be divided into three types. The inserted vowel landmarks of Type 1 appear after a word-final voiced stop. In the four examples in Table 1, the environment where the vowel landmark is inserted is after a word-final /d/ (in *pond* and *parenthood*) or /g/ (in *big* and *synagogue*). Figure 2 illustrates a spectrogram and the matching words and labeling tiers, LM and LMmods, for part of the phrase *parenthood organizations*, in which we see a vowel between the two words, circled in red. This is marked with V-+ in the LMmods tier.

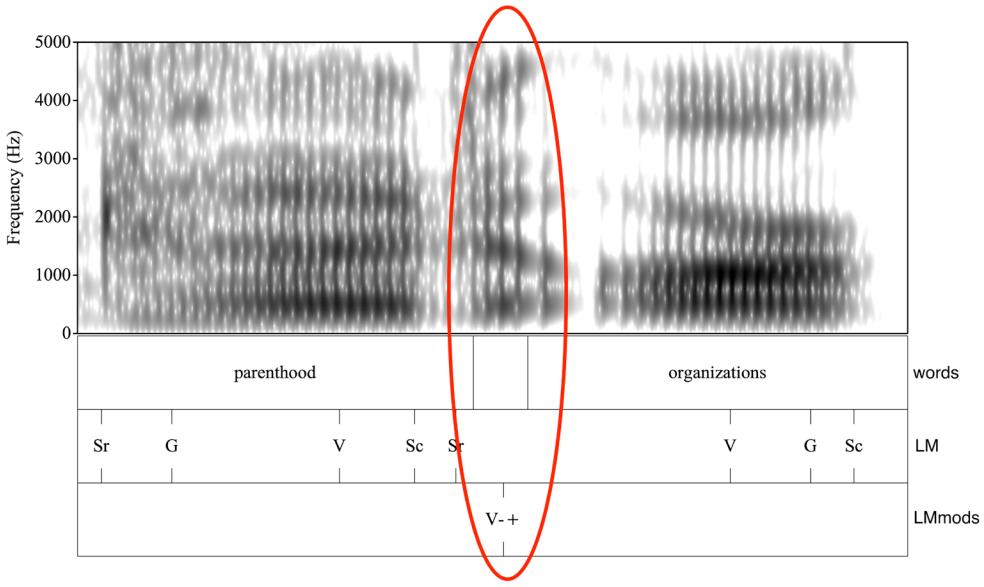

Figure 2: Part of landmark annotations for the spoken phrase *parenthood organization*[2]

Type 2 is similar to Type 1 in that it also involves the vowel landmark labeled with V-+ at the end of the word. Unlike Type 1, however, there is a possibility for Type 2 that the inserted vowel landmark represents an

[1] Dialect regions (DR) in TIMIT are classified as follows. DR1: New England, DR2: Northern, DR3: North Midland, DR4: South Midland, DR5: Southern, DR6: New York City, DR7: Western, and DR8 Army Brat (moved around).

[2] Only the words, LM and LMmods tiers are displayed in the figures hereafter for the sake of convenience.

article *a* or *the* that the speakers may have unconsciously produced, although it was not part of the text they were supposed to read. There are two cases of Type 2 vowel landmarks in the database. An example is demonstrated in Figure 3.

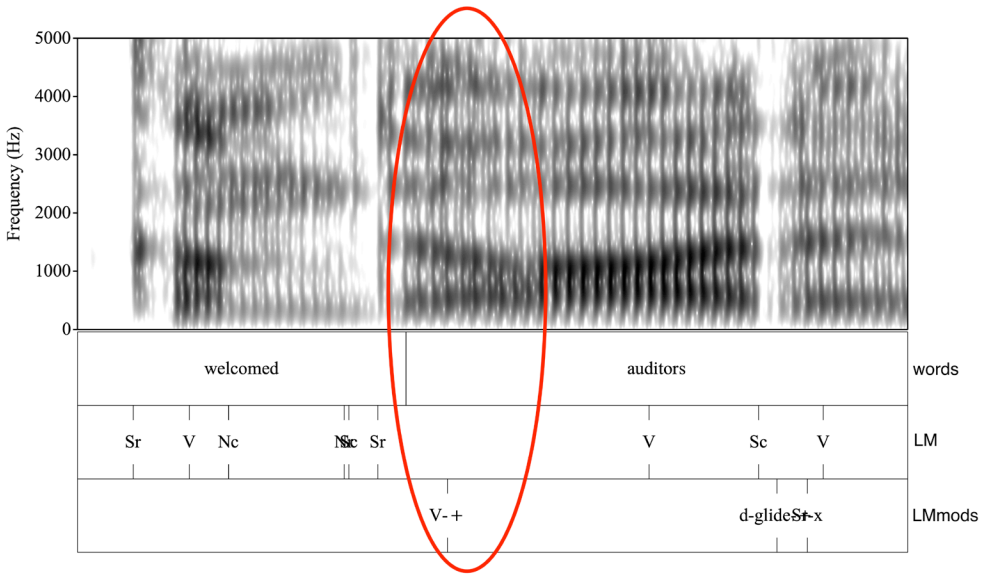

Figure 3: Part of landmark annotations for the spoken phrase *welcomed auditors*

Here we see a vowel after /d/ in *welcomed* and before the next word *auditors* starts, circled in red. This is a separate vowel, not part of the following vowel /ɔ/ in *auditors*, as they differ in the formant structures as well as in overall amplitude as shown in the spectrogram. This unexpectedly inserted vowel may represent the definite article *the*,[3] as it is located between a transitive verb and its object. The same explanation may apply to the other example, in which the unpredicted vowel landmark appears between the conjunction *and* and the following bare noun *harmony*.

While both Type 1 and 2 involve the inserted vowel across the word boundary, there is also a case where the unpredicted vowel landmark occurs within a word, which I call Type 3. As illustrated in Figure 4, a short and weak vowel appears between [g] and [z] in the word *exam*. This is the only within-word, within-cluster case of vowel insertion observed in the database, indicating that Type 3 occurs much less frequently than Type 1 and 2.

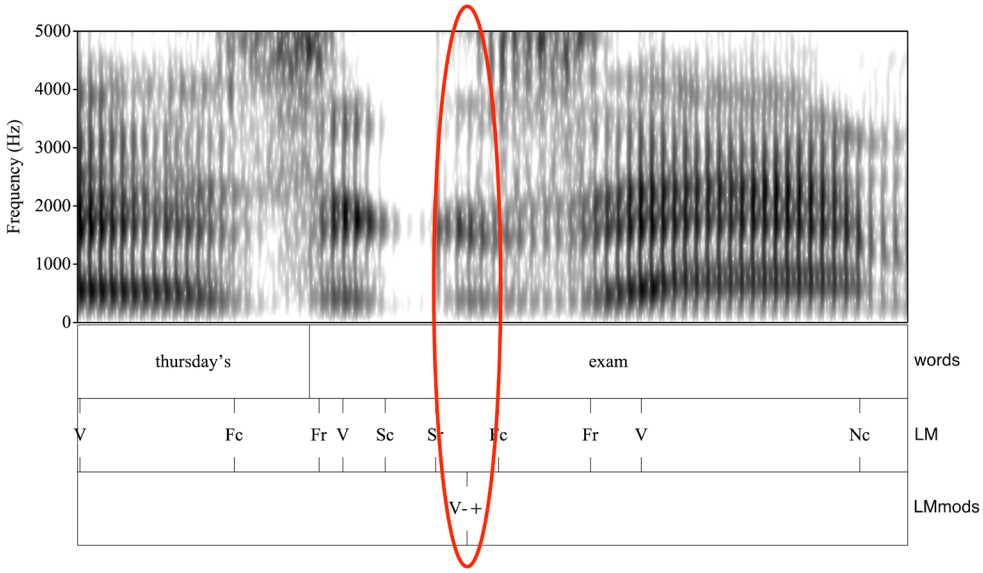

Figure 4: Part of landmark annotations for the spoken phrase *Thursday's exam*

[3] I thank Jeung-Yoon Choi for pointing out this possibility.

4 Discussion

4.1 *Phonetic characteristics of the inserted vowels* This study finds out that in the read speech, American English speakers may insert a vowel landmark where there is no underlying vowel, although not frequent. The vowels marked with an unpredicted vowel landmark display periodic patterns in waveforms and clear formant structures like lexical vowels. While these are inserted vowels, they differ from phonologically epenthesized vowels, like the ones inserted before the plural suffix /-z/; these inserted vowels are totally optional and do not repair any marked structures. In this sense, these unexpectedly inserted vowels should be described as excrescent or intrusive vowels.

Excrescent vowels are short in duration, as mentioned earlier, and so are the inserted vowels in the current database. Table 2 shows the duration of each inserted vowel. The mean duration is 44 ms, slightly shorter than the mean duration of unstressed word-final short vowels in American English, 49 ms, reported in Crystal and House (1988).

Type	Dialect	Gender	Word	Duration
1	DR2	M	pondᵊ	48
	DR4	M	parenthoodᵊ	45
	DR7	F	bigᵊ	39
	DR7	M	synagogueᵊ	28
2	DR2	M	welcomedᵊ	85
	DR5	F	andᵊ	32
3	DR6	M	exᵊam	32

Table 2: Duration of the inserted vowels (ms)

The inserted vowels in the current database are mid central vowels, like typical excrescent vowels. Table 3 lists the F1 and F2 values measured at the midpoint of each inserted vowel. The mean F1 is 502 Hz, and the mean F2 is 1616 Hz.

Type	Dialect	Gender	Word	F1	F2
1	DR2	M	pondᵊ	459	1513
	DR4	M	parenthoodᵊ	450	1350
	DR7	F	bigᵊ	429	2546
	DR7	M	synagogueᵊ	583	1501
2	DR2	M	welcomedᵊ	562	1184
	DR5	F	andᵊ	638	1705
3	DR6	M	exᵊam	395	1510

Table 3: F1 and F2 frequencies (Hz) measured at the midpoint of the inserted vowels

It may not be appropriate to say, however, that all inserted vowels observed in the present study are excrescent vowels. For the vowels of Type 2, it may possibly be a function word *a* or *the*, not just a meaningless vowel, that is inserted, as mentioned in the previous section. This possibility might be supported by the fact that the Type 2 vowel inserted after *welcomed* is 85 ms in duration, much longer than the others (see Table 2). If this is the case, Type 2 vowels are not excrescent vowels but part of the function word produced inadvertently. There is no concrete evidence, however, that they represent part of the inserted article. It should be mentioned that part of a vowel may be perceived as a separate function word depending on the speech rate (Dilley and Pitt 2010), and therefore even if the vowel is excrescent in the current examples, it is possible to be perceived as the article based on the syntactic structure.

4.2 *Phonological conditions for the inserted vowels* The excrescent vowels in American English are found to occur after a voiced stop, regardless of whether it is word-final or word-medial. One possible exception is the case of vowel inserted at the end of *and* (Type 2); the word-final /d/ is deleted and the vowel appears after the /n/, which becomes word-final.

The fact that the vowel appears only after a voiced stop indicates that it may derive from a vocalic release of the stop. Pennington (2013) describes that in emphatic speech of English, the release of word-final voiced stops may be realized with a slight vowel. The inserted vowels in the current database could be the case of vocalic release of the voiced stop, as TIMIT is a read speech corpus and the speakers read short written texts in the lab setting, which may have led them to articulate in a more careful and clear way than usual as in emphatic speech.

One might wonder whether the vowel is inserted only in the vicinity of a strong prosodic domain such as Intonational Phrase. It does not seem, however, that the prosodic domain influences the vowel insertion, at least

in the read speech. Most of the cases are embedded in the middle of a prosodic phrase, and only two that appear at the end of the words *pond* and *synagogue* are posited at the boundary of intermediate phrase.

On the other hand, stress is a possible precondition for the Type 1 vowel intrusion in American English. The only syllable in *pond* and *big* and the final syllable in *parenthood* and *synagogue*, which allow the inserted vowel at the end, bear a primary or secondary stress. It is hard to conclude though that the vowel can be inserted only after a stressed syllable, due to the limited number of examples. Future research is needed to find out the exact conditions where vowel intrusion takes place in American English.

4.3 *Functions of the inserted vowels* The last question addressed in this paper is why the vowels are inserted at all without a structural reason. The vowel insertion may happen either intentionally or unintentionally. The intended function of the inserted vowel would be to help listeners better perceive the consonant in a perceptually weak position. The stop consonant at the end of a word or before another consonant, where the vowel appears unexpectedly in the current study, are not necessarily released audibly in American English. And without audible release, it is harder for listeners to recognize whether there is a stop or what kind of stop it is, compared to when it is clearly released. Having a short vowel is the strongest form of the release of a voiced stop, and the inserted vowel facilitates the perception of the preceding voiced stop having it audibly released with formant transitions into the vowel. Increasing perceptibility of the adjacent consonants has also been pointed out as a function of excrescent vowels in the other languages (Hall 2006 and references cited therein).

On the other hand, it may also be possible that a vowel appears unintentionally without purpose, as a result of gestural retiming. As described earlier, vowel intrusion has commonly been explained as part of the existing vowel gesture exposed by the retiming of the adjacent consonant gestures within the Articulatory Phonology framework. It might be just the case that the vowel gesture is expanded over the following consonant gesture (Type 1 & 2) or the gestures of the two consonants in a cluster are loosely coordinated, resulting in a period of open vocal tract between the two consonant gestures (Type 3), which is recognized as an inserted vowel.

5 Conclusion

To summarize, a close investigation of the inserted vowel landmarks in part of the TIMIT database shows that in American English read speech, an excrescent vowel may occur with a low frequency. The vowel may appear (i) after a word-final voiced stop (Type 1), (ii) after a word-final voiced stop possibly as part of a function word (Type 2), or (iii) between voiced obstruents (Type 3). Like typical excrescent vowels, the current ones are mid central vowels with short duration. The preliminary finding suggests that these excrescent vowels aid listeners' perception of the preceding voiced stop as a form of the strongest audible release, although it is also possible that they just appear as a result of gestural mistiming without purpose. Certainly, this paper is limited in scope; the cases reported here are limited to the seven examples from the read speech only. To reach a fuller understanding of the excrescent vowels in American English, future studies will investigate more examples in spontaneous speech.

References

Boersma, Paul and David Weenink. 2022. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/.

Browman, Catherine P. and Louis Goldstein. 1992. "Targetless" schwa: An articulatory analysis. In Gerard J. Docherty and D. Robert Ladd (eds.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 26-56. Cambridge: Cambridge University Press.

Crystal, Thomas H. and Arthur S. House. 1988. The duration of American-English vowels: an overview. *Journal of Phonetics*, 16, 263-284.

Dilley, Laura C. and Mark A. Pitt. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11), 1664-1670.

Garofolo, John. S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. 1993. DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia, PA.

Hall, Nancy. 2006. Cross-linguistic patterns of vowel intrusion. *Phonology*, 23(3), 387-429.

Hall, Nancy. 2011. Vowel epenthesis. The Blackwell companion to phonology, 1-21.

Huilgol, Shreya, Jinwoo Baik, and Stefanie Shattuck-Hufnagel. 2019. A framework for labeling speech with acoustic cues to linguistic distinctive features. *The Journal of the Acoustical Society of America* 146(2), EL184–EL190.

Levin, Juliette. 1987. Between epenthetic and excrescent vowels. *Proceedings of the West Coast Conference on Formal Linguistics* 6, 187-201.

Pennington, Martha C. 2013. *Phonology in English Language Teaching: An International Approach*. London and New York: Routledge.

Steriade, Donca. 1990. Gestures and autosegments: comments on Browman and Goldstein's paper. In John Kingston and Mary Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 382-397. Cambridge: Cambridge University Press.

Stevens, Kenneth. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The*

Journal of the Acoustical Society of America 111(4), 1872–1891.
Wright, Joseph. 1905. *The English Dialect Grammar*. Oxford: Frowde.
Yun, Suyeon, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. 2020. A landmark-cue-based approach to analyzing the acoustic realizations of American English intervocalic flaps. *The Journal of the Acoustical Society of America* 147, EL471-477.