**Frequency Effects of the Variation on Mandarin Zero Onsets: deletion vs. augmentation processes**

**Shuo Zhang**

The current study addresses the occurrence of frequency effects associated with a reductive phonological process (e.g., deletion) vs. an augmentation process (e.g., epenthesis). According to Coetzee & Kawahara (2013), the deletion rates in an observed variable process (such as English t/d deletion) vary across words with different rates of usage frequency, such that the words in the higher frequency interval (e.g., 'send') has a higher deletion rate comparing to those in a lower frequency interval (e.g., 'lard'). Efforts have been made to incorporate such frequency biases into the noisy harmonic grammar in order to derive a more accurate model of representation of such variable processes (ibid).

The processes focused in the C&K (2013) paper are both examples of simplification or reductive processes—i.e., the form that has undergone the process is in some sense articulatorily simpler or more reduced than the input. In the current study, I follow their discussion on whether such frequency effects can affect the application of augmentation processes, and explore the following research questions: (1)are augmentation effects affected by frequency biases in a similar way comparing to reduction processes? (2)if so, how can we incorporate this factor into our modeling of variation? (3)if not, how are they different? Why?

The variation in the Beijing Mandarin Chinese (BMC) zero onsets exhibits an interesting array of variable phenomena involving both reductive and augmentation processes. In the traditional analysis of Standard Mandarin (which is established based on BMC), a full Mandarin syllable can begin with a consonant, a glide, or a Consonant-Glide (CG) combination. When a full syllable does not begin with any of these three cases, there is still an articulatory effort in the onset, which has been called the ‘zero onset’, indicated by [ø](Chao 1968, Li 1966). This zero onset has four realizations (Chao 1968:20): [ɣ][

ʔ][ɦ][ŋ]. The distribution of these is not entirely free variation or predictable. The important function of the zero onset is to prevent 'linking' of an onset-less nuclear vowel (e.g., /a/ in /an/) to the previous consonantal (nasal) coda or high vowel coda (Chao 1968). Duanmu (2007) gives the following example to illustrate the variants and the prevention of the linking of the onset-less vowel /a/ to the coda nasal of the previous word:

(1) [mian øau] → (a) [mʲan ʔau]

'cotton coat' (b) [mʲan ɣau]

(c) [mʲaŋ ɣau]

(d) [mʲaŋ ŋau]

*(e) [mʲanau]

All of these variants (a)-(d) are augmentation processes. In addition to these variants, the current BMC data suggests that there is one more variant possible, which is a reductive process:

(2) [mian øau] → (a) [mʲãau]

in (2)a, we observe that the consonant coda is deleted from the first word, resulting in a V-#-V sequence. Here, however, we have two considerations:

(i). Contrary to the claim that an onset slot is required in Standard Mandarin, the V-#-V sequence is legal in BMC, i.e., onset-less syllables is legal when they follow a vowel coda. This is to be verified through acoustic analysis.

(ii). the application of (2)a is restrictive. For instance, when the onset-less syllable begins with a high vowel, a glide is inserted and deletion will not be observed (e.g., /i/→/ji/, /u/→/wu/).

Our initial approximation to this variation phenomenon in BMC can be summarized as:

*(3) When a syllable following a consonantal coda is onset-less, two strategies are observed in order to prevent 'linking' between the coda consonant and the nuclear vowel of the onset-less syllable: (a) insert a realization of the zero onset; (b) delete the coda consonant.*

Having established this characterization, I define the new research tasks of the current paper:

- Design a C#V target production task and collect data from BMC speakers;
- If variants from (3) above is indeed observed from the data, what is the observed frequency of the reduced vs. the augmented variants?
- What factors account for this distribution? (And what is its implication, in relation to C&K's model and discussion?)
- Is there a usage frequency effect associated with the variation of reduction vs. augmentation processes?
- How can we incorporate the frequency effects into a grammar that models this variation?

**1. /C-#-V/ Target Production Experiment Design**

In order to observe the frequency effects associated with the variants, a relatively large amount of production data in natural BMC speech is required. Unfortunately, no known corpus of BMC has annotations for the target of interest in this paper. Moreover, as will be discussed below, the frequency of the words involving target segments C-#-V is in general low, as is the number of total possible words/phrases that involve such a sequence. Therefore, obtaining abundant data for analyzing frequency from a corpus can be a difficult task. I turn to the alternative method, to collect production data from BMC speakers.

### 1.1 Format of Speech Production Task

Several considerations are taken into the design of the format of the production experiment. First, speakers will need to engage in a rather informal speech style that they usually employ in everyday casual speech. There are reasons to believe, from my own observation as a native speaker, that a formal style will induce less variation (especially of the reductive variant) in BMC speech (and in standard Mandarin in general). We do not want the subjects to read formally, word by word, a word list, which will likely lead to a more standardized style/mode of production. Second, we want to have control over the amount of structural variation in the produced sentences that contain the target words. In other words, since we're not sure at this stage what is the influence of prosody (stress, duration, metrical organization, etc., correlated with the syntactic position of the words in the sentence) on the choice of the variants (and I believe they do), we want to make the sentences as uniform as possible and limit the amount of other factors in order to derive reliable results. In order to achieve these, at the same time obtaining tokens for the needed target words, I used a simple production formula that can induce a spontaneous speech style and also maintain a simple and consistent sentence structure. Here is an example of the test sentence from this formula, where the blank space will be replaced by a target word from a list of words given to the speaker:

(4) wo xihuan/ buxihuan_____.
I like dislike
"I like/don't like_____".

The subjects are instructed to speak in a natural speech style similar to their everyday life speech.

### 1.2 Design of Word List

As noted above, in order to observe the frequency effects on the variants, a decent word list with a good amount of tokens of the target phonological structure sequence is crucial. However, this has been proved to be not an easy task. First of all, there are only two consonants that are allowed to be in the coda position of a possible syllable: /n/ and /ŋ/

(this is expected since Mandarin only allows segments in the coda position that are more sonorant—mostly vowels). The possible types of syllable structures associated with these to codas are expected to be limited. However, a search through an exhaustive list of possible syllable types in Mandarin (Duanmu 2007, appendix) reveal that contrary to our expectation, the syllable types with these two kinds of coda account for almost 45% of all the possible syllable types in Mandarin. This gives us a lot more of possibilities in constructing a word list with a consonant coda.

The next step is to look at what vowels in onset-less words/syllables are available to follow a consonant coda of the previous syllable, and to allow for the two variants that we are interested. Duanmu (2007) listed the following Mandarin vowel inventory:

Mandarin Vowels:

- High: [i y u],
- Mid: [ə o e ɤ],
- Low: [a]

Each of which carries several variants according to their phonological positions. The high vowels are excluded from the consideration for the current task because a glide is always inserted when they follow a consonant coda (Duanmu 2007). Among the remaining mid and low vowels, mid vowels only occur in the syllable-initial position very infrequently (there are less the three syllable types I can list), and for low vowels, there are relatively more syllable types with a low vowel in the initial position (<=5, a, ai, ao, an, ang, where a is a meaningless segment as in 'ah', and does not appear in a word-internal position, assuming a two-syllable or three-syllable word).

Overall, neither C# or #V gives us a lot of options for target words. Luckily, with the combination of a few syllable types and the four tones, we finally obtain a word list of 93 targets. My initial observation and corpus query confirm that none of these words is very frequent (according to C&K, only two or three words account for about half of all the tokens in their corpus, which count as 'frequent' word. I assume this observation holds across languages, in Mandarin as well), but there is indeed a difference across frequency

bins (mostly within the bottom 25% percentile in a ranked frequency list). It is to be decided late whether we can actually observe a frequency effect based on this word list (it could be that these words at the bottom 25% occupy the same frequency bin and has no observable effects on the deletion/epenthesis rates).

9 subjects who are native speakers of BMC (college students) participated in the production experiment.

**2. Data Annotation, Word Frequency and Data Processing**

The production task yielded a total of 93*9=837 tokens of the target sequences. Among these sequences, most are C#V sequences. There are a few V#V sequences to be analyzed in order to confirm the validity of a V#V sequence in BMC (as opposed to the Standard Mandarin, in which V#V is not licit according to literature). A total of 22 tokens are not used for data analysis (including the V#V sequences) because beside the V#V sequences, there are many C#V sequences that cross syntactic boundaries (P#NP, i.e., it is not a 'word' per se), and they are almost always categorically realized with a pause or a epenthesized glottal stop. The final number of tokens for analysis is therefore 837-22*9=639 tokens.

All tokens are hand labeled as epenthesis or deletion. Only the glottal stop variant is observed in the case of epenthesis for the current study. Among the deletion tokens of a V1C1#V2C2 sequences, a perceptual difference can be made between those that realize the post-deletion V1-V2 sequence as a diphthong (shorter duration) and those that realize this sequence as two vowels (longer duration). Both types are labeled as deletion token and no distinction has been made in the annotation of these two types of realizations. Also, caution has been taken not to mistake the longer V1-V2 sequence for an epenthesis token (since the existence of some variants of the epenthesized segments are not very salient to distinguish perceptually).

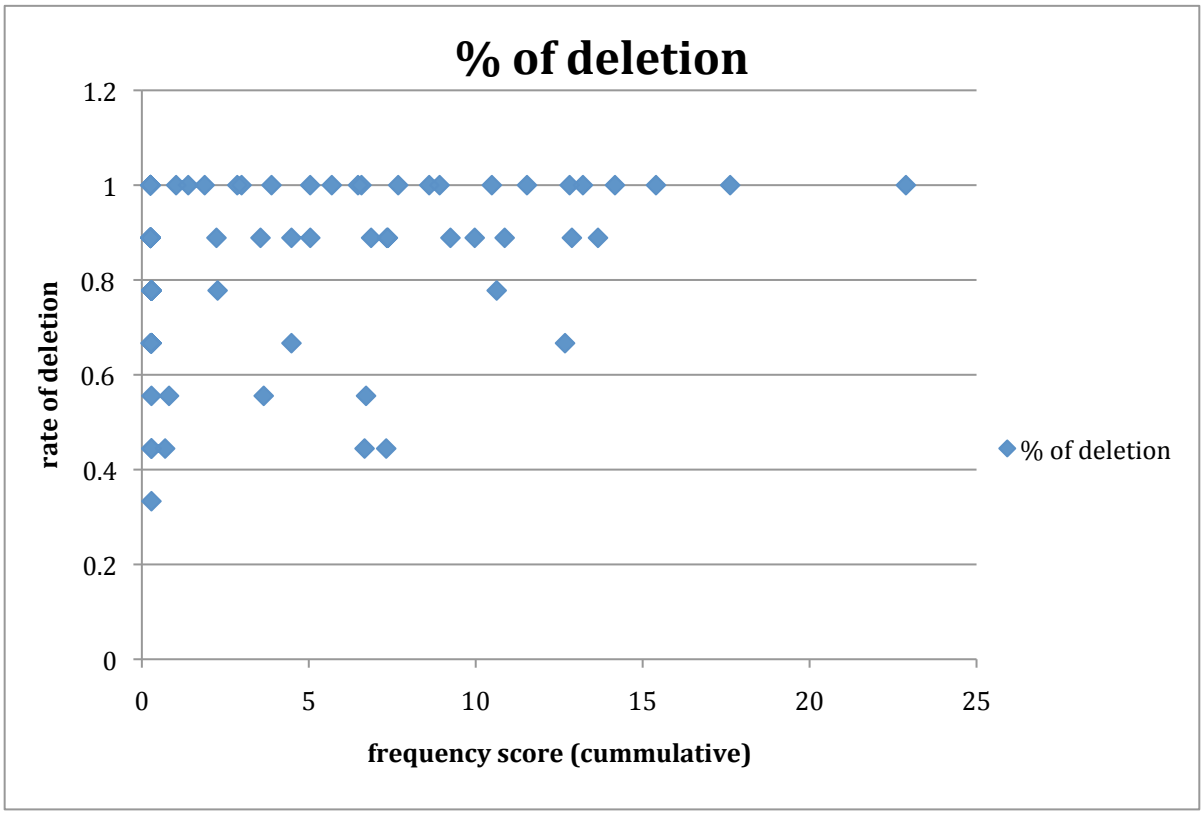
The important task in the data processing is to establish word frequencies. For this task, I used the *Word List with Accumulated Word Frequency in Sinica Corpus 3.0* (5 million tokens, 146,876 types, Taiwan based). The corpus query interface can be accessed online at (http://elearning.ling.sinica.edu.tw/CWordfreq.html). A query of a particular word frequency would return the following results (Nc in the parenthesis is a syntax tag and is irrelavent):

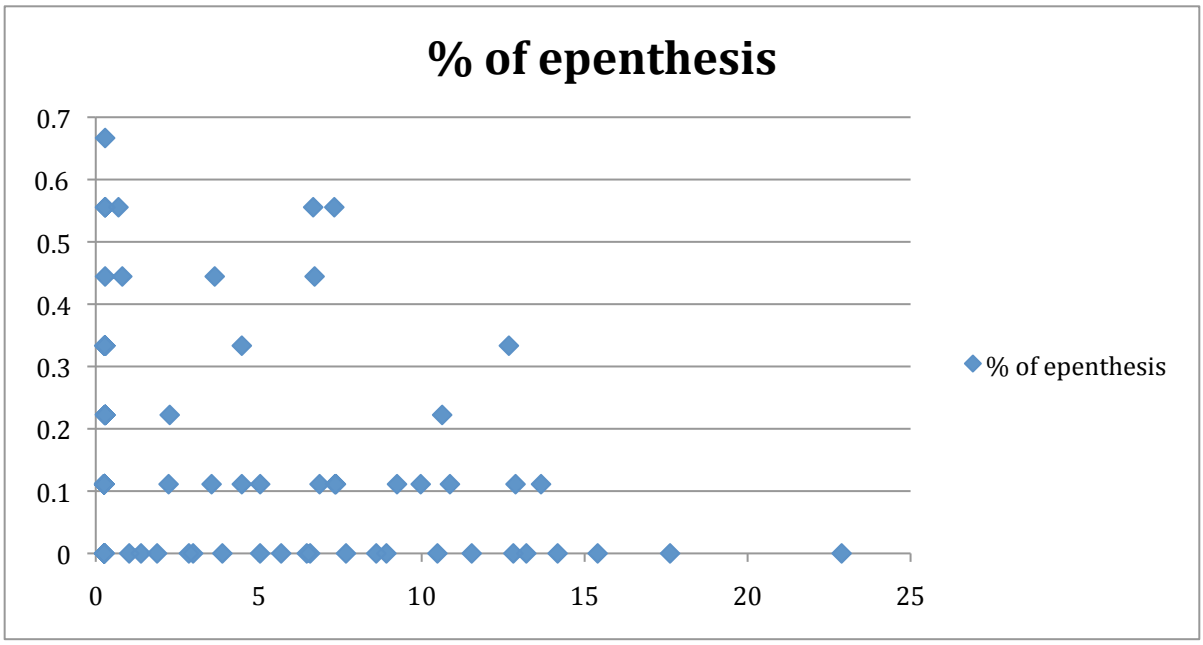
**現代漢語個別詞的頻率 Word frequency: 長安**

| No | Rank | Word | Frequency | Percent | Cumulation |
|---|---|---|---|---|---|
| 10970 | 10827 | 長安(Nc) | 39 | 0.001 | 85.829 |

There are several measures I took to establish a word-frequency score for the data analysis. First, many of the target words with the C#V sequence are street names or city names of China (the morpheme -/an/ is often used in place names), which would yield a very low frequency in Taiwanese Mandarin if the name is local in Beijing or mainland China and is rarely heard of in Taiwan. To resolve this misrepresentation, I used the frequencies of the main street and city names in Taiwan as a substitute. Second, as discussed above, many of the target word are in the low frequency bin and quite a few will return a '0', which means they did not appear in the current corpus. Here I follow the tradition of Natural Language Processing and Computational Linguistics to set the frequency of these words as equal to the second lowest, non-zero frequency I found within all targets. Finally, I leave the more strict discussion of the establishment of word frequency in a particular words and its possible disparity from the actual word frequency for a particular speaker or speech community out of the scope of discussion of this paper. To compute the frequency score in a manner that higher frequency words will have a higher frequency score, I used the formula F=100-f, where f is the cumulative frequency ranking of a word (if f=82 it is ranked on the 82% percentile in the corpus, i.e., about 82% of the words are more frequent than this word in this corpus).

### 3. Results and Discussion



The above two charts shows the rates of deletion (y) across frequency bins (x), and the corresponding results for rate of epenthesis across different frequency bins. In this section I discuss several aspects of the results that are observed from the data.

First of all, the production data from the 9 subjects who are native speakers of BMC confirms that /V#V/ is a licit sequence in BMC, and the deletion variant is observed in a /C#V/ sequence, alongside the epenthesis variant. Thus we have empirically confirmed our first characterization of this variable phenomena described in (3) above. Figure 1 and 2 show the spectral contrast between an epenthesis variant (E) and a deletion variant (D) of the word [*Nan An*](city name). The continuous and stable formants of the deletion variant are highly visible, showing the two vowels connecting with each other.

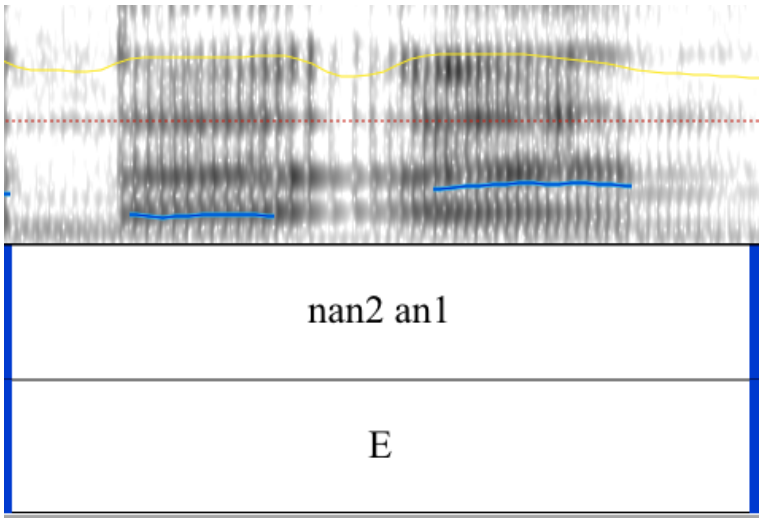


Figure1 Epenthesis variant with glottal stop

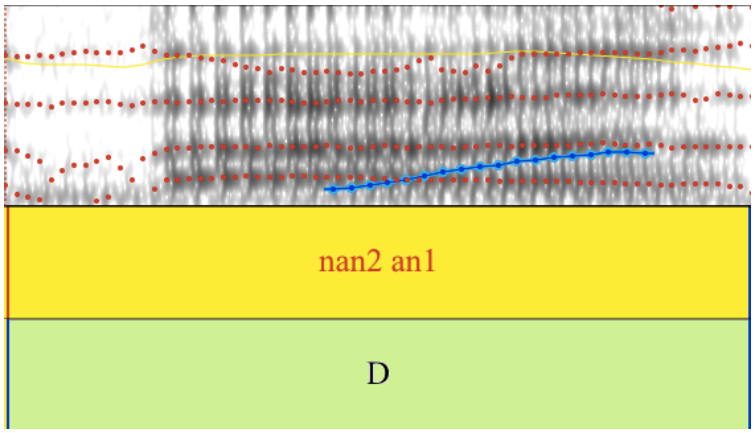


Figure2 Deletion variant

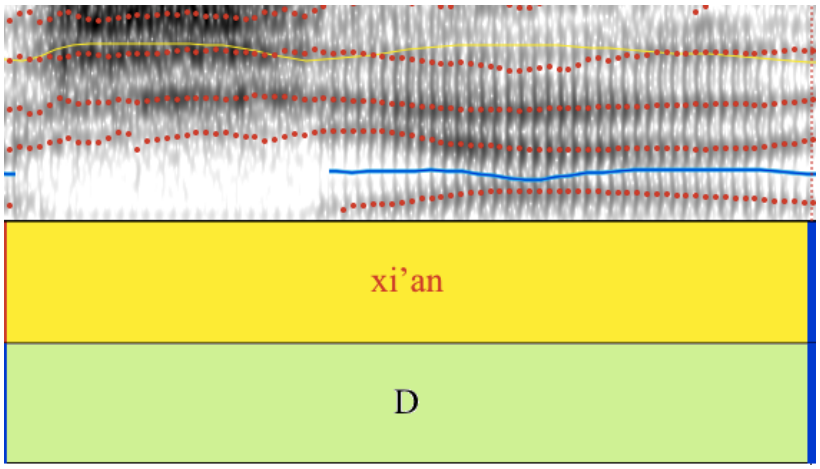



Figure 3 Deletion variant of the word 'Xi'an'

Our data directly confirms the validity of /V#V/ sequence in BMC with tokens of a word with such a sequence, as shown in the case of the spectrogram of "*Xi'an*" (the city name), which is perceptually indistinguishable from the word '*xian*' (to open) when speaking in a faster rate (in which /i/ and /a/ are concatenated into a diphthong).

Having established the validity of (3), we proceed to analyze the relationship between the word usage frequency and the rate of deletion vs. epenthesis. Several observations can be made from charts showing the distribution of deletion rates and epenthesis rates across the different frequency bins. First, strictly speaking, the data is not linearly distributed since we see a categorical deletion ($F_{del}$=1) across all frequency bins. However, we do see a general linear tendency between the deletion rates and the word frequency as in higher frequency words have higher deletion rates. Also, comparing out data with C&K's data on English t/d deletion, we see that in their data, the relationship is not strictly linear either in that there are words with higher frequencies that have lower (or the same) deletion rates comparing to the lower frequency words. Overall, our data turned out to be matching the predictions about a relationship between frequency and deletion rates. In both our data and C&K's data, because there are only very few words (around or less than 3) that can be considered as frequent, the validation for a high frequency words with higher deletion rates only come from these very few words, whereas the lower frequency

words occupy a large space that show a noisy distribution with a general (not strict) linearity. One may question the validity of such results.

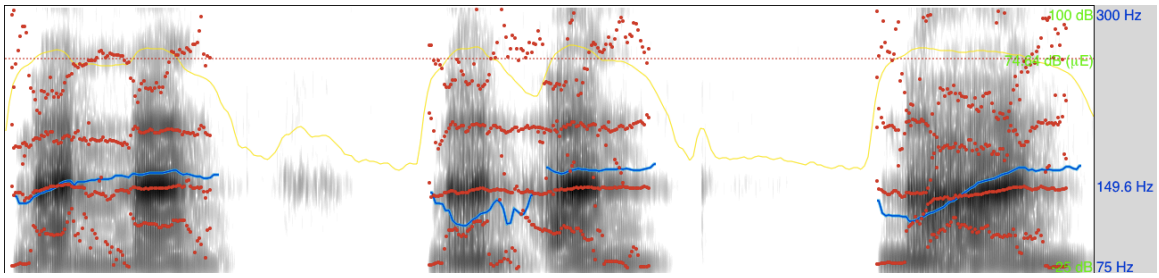
Second, comparing our data with C&K's data, the case of our data showed a tendency for the lower frequency words to have more variation than higher frequency words. This is clearly demonstrated by the fact that categorical deletion is observed almost across every frequency bin. This observation may be explained by the exemplar theory: lower frequency words are heard less, therefore there are less exemplars stored for these items in our brain, which in turn produces more variants across speakers. In C&K's data, however, this tendency is not clear. However, there is another factor that may obscure the observation on the amount of variation, in both our data and C&K'd data: the fact that there are much more low frequency words than there are the high frequency words. In other words, low frequency words may have more variation because there are more tokens of them; high frequency words cannot have much variation because there are only few of them. This analysis is not incompatible with the exemplar theory. In fact, the fact that the very few high frequency words turned out to be of high deletion rates can be a confirmation of the relationship between frequency and deletion rates.

The third point to be made regarding the experiment results is that of the relationship between epenthesis and deletion rates, which is one of the first motivating questions of this paper. C&K argued that analogous to the reduction processes they have examined, an augmentation variable process also depends on the relative weights of markedness and faithfulness constraints. They further postulated that their model would predict that epenthesis (in a epenthesis-or-faithful situation) will be observed more often in more frequent words than in less frequent words. This prediction will not apply to the current data because the current phenomenon is not an epenthesis-or-faithful type of situation. Rather it is an epenthesis-or-deletion situation, where the faithfulness constraints always sink to the bottom. In the next section I will discuss the above results from an OT perspective and sketch a possible model in Noisy Harmonic Grammar.

## 4. A Noisy Harmonic Grammar Approach

The similarity in the behavior of deletion rates vs. epenthesis rates across different frequency bins in my data and the behavior of deletion rates vs. faithful candidates is striking. Is there a reason why that is the case? In other words, is it possible that epenthesis of a glottal stop is closer to a faithful candidate than deletion? If so, in order to account for our data, we need a model that can assign different scores (or weights) to the different constraints that constitutes a violation: that is to say, to epenthesize a glottal stop (violating DEP) is more faithful (or not as bad) than to delete a coda consonant (violating MAX). This would explain the similarity of behavior between the epenthesis candidates in my data and the faithful candidate in the case of t/d deletion.

This idea is conceivable in the current data because the glottal stop is not a contrastively salient segment as most other types of consonant. One evidence for this is that glottal stop is often allowed to be optionally inserted in the beginning of syllables where the onset slot is empty (i.e., vowel initial) cross-linguistically (/an/ vs. /ʔan/) without changing the identity of the word. According to the licensing by cue theory (Steriade 1999,2001), a segment is protected by special faithfulness constraints in contexts where its perceptual cues are robustly licensed. The consonant release burst can cue both place and manner information (Lahiri et al. 1984; Stevens and Blumstein 1978; Stevens and Keyser 1989), and formant transitions into a following vowel also carry information about place and manner (See C&K section 3.2.1 for a list of reference). In the current data, if we consider (1)the faithful candidate (even though it never is allowed to win), (2)the epenthesis candidate and (3)the deletion candidate, since the second syllable initiates with a vowel, both (1) and (2) carry the two kinds of cues (burst and formant transition) in this context, but (3) does not. This is a possible explanation of why the epenthesis candidate in our data behaves much like the faithful candidate in the deletion-or-faithful situation as showed by C&K's data. The spectrogram below which shows (1), (2), and (3) from left to right further illustrates the similarity between the faithful candidate and the epenthesis candidate.




(1)/nanan/ (2)/nanʔan/ (3)nãan

**Figure 4 (1)faithful candidate; (2)epenthesis candidate; (3)deletion candidate**

The above discussion has an implication that even though both candidates (2) and (3) violate the faithfulness constraints, there is a distinction between the degree of violation in that (2) is closer to faithful candidate than (3). To capture this distinction, we need a model that can assign different violations different scores, analogous to the Harmonic Noisy Grammar, which assigns different noise scores to the faithful vs. deletion candidate in the t/d deletion study in order to capture the variation with the same constraint ranking(C&K section 2.1). Here I will sketch such an OT analysis.

First I will show a preliminary analysis of this variable phenomenon based on the standard OT. We need to keep in mind that in this process, faithfulness is always violated and faithful candidate never wins. In addition to the standard faithfulness constraints, MAX and DEP, we propose a markedness constraint that captures the fact that a syllable initial vowel can never be linked to the previous consonant coda. In other words, this constraint forbids the resyllabification (i.e.,the coda of the first syllable must stay as a coda and cannot become the onset of the next syllable). We'll call this constraint *RESyl. Below is a standard analysis, using two crucial rankings of MAX and DEP to produce the variation:

- Deletion candidate wins

| nan+an | *RESyl | DEP | MAX |
|---|---|---|---|
| nanan | *! | | |
| ⇒nãan | | | * |
| nanøan | | *! | |

- Epenthesis candidate wins

| nan+an | *RESyl | MAX | DEP |
|---|---|---|---|
| nanan | *! | | |
| nãan | | *! | |
| ⇒nanøan | | | * |

Using a mock Noisy Harmonic Grammar (as in the values are made up for the sake of demonstration), we will follow this formula:

(5) $$H(cand) = \sum_{i=1}^{n} (w_i + nz_i) C_i(cand)$$

Where $w_i$ is the weight of constraint $C_i$, $nz_i$ the noise associated with constraint $C_i$ at this evaluation occasion, and $C_i(cand)$ is the number of times that *cand* violates $C_i$, expressed as a negative integer.

and we will need to assign different noise scores for MAX and DEP in each situation:

- Deletion candidate wins

| nan+an | w nz<br>5 -0.7 | w nz<br>1.5 0.1 | w nz<br>1 -0.1 | H |
|---|---|---|---|---|
| | *RESyl(4.3) | DEP(1.6) | MAX(0.9) | |
| nanan | -1 | | | -4.3 |
| ⇒nãan | | | -1 | -0.9 |
| nanøan | | -1 | | -1.6 |

- Epenthesis candidate wins

| nan+an | w nz<br>5 -0.7 | w nz<br>1.5 -0.3 | w nz<br>1 0.5 | H |
|---|---|---|---|---|
| | *RESyl(4.3) | DEP(1.3) | MAX(1.5) | |
| nanan | -1 | | | -4.3 |
| nãan | | | -1 | -1.5 |
| ⇒nanøan | | -1 | | -1.3 |

This analysis effectively captures the current data with a consistent ranking of constraints. Along a similar line of C&K, we can incorporate a scaling factor that represent the frequency effect for the deletion vs. epenthesis rates. This is beyond the scope of the current paper and will be highly analogous to the modeling carried out by C&K, for which reason I will leave it out for future inquiry.

**Reference**

Chao, Yuan-Ren. (1968) A Grammar of Spoken Chinese. (Berkeley and LA: UC Press).

Duanmu, San. (2007) The Phonology of Standard Chinese. Oxford University Press.

Li, Fangkui. (1966) 'The Zero Initial and the Zero Syllabic'. Language, 42:300-2.

Coetzee & Kawahara. (2013). Frequency biases in phonological variation. Nat Lang Linguist Theory.