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

Shuo Zhang

November 1, 2013

Abstract

This paper examines frequency effects associated with a reductive phonological process (deletion) vs. an augmentation process (epenthesis) in Beijing Mandarin. According to Coetzee & Kawahara (2013), the English t/d deletion rates vary across words with different usage frequencies, such that higher frequency words (e.g., 'send') have a higher deletion rate compared to those with a lower frequency (e.g., 'wend'). I investigate the case of two-way variation (augmentation vs. reduction) variation in Beijing Mandarin Chinese (BMC) zero onsets. In BMC, coda consonants are prohibited from re-syllabifying as the onset of the following onsetless syllable. Instead, one of the two processes occur: a "zero onset" segment such as [ʔ] acn be inserted, ('zero onset' segment, [ø](Chao 1968, Li 1966), or the consonant coda of the first syllable may be deleted. A production task eliciting /C#V/ sequences in a word list was designed to determine whether there is any frequency effects on the distribution of the two processes. 9 subjects who are native speakers of BMC completed the semi-spontaneous speech-style production task, yielding 837 target tokens for statistical analysis. Word frequency bins are established by querying the online interface of Word List with Accumulated Word Frequency in Sinica Corpus 3.0 (5 million tokens, 146,876 types). Results show a positive correlation between frequency and deletion rate (r=.29,high frequency words show high rate of deletion), whereas a negative correlation is found between frequency and epenthesis rate. Moreover, lower frequency words showed a greater amount of variation than higher frequency words. Considering the similarity of the results in the current study (deletion vs. epenthesis) to those between deletion vs. faithful in C&K(2013), I argue that the low contrastive salience of the epenthesized segment (such as [ʔ]) makes the epenthesis rate behaving similarly to the faithful candidates in C&K(2013). In addition, I consider the results in the light of the Exemplar Theory (Keith 1997) and the Licensing-By-Cue hypothesis (Steriade 1999,2001). Finally, I apply C&K(2013)'s computational learning model on the BMC data, incorporating the frequency effects using a Noisy Harmonic Grammar(Smolensky and Legendre 2006) with weighted scaling factor.

Recent years have seen a rising interest in generative phonology to take into account evidence from lingusitic variation phenomena as a way to approach phonological theory from an empirical and data-driven perspective. In the Optimality Theory research, for instance, a model's ability to account for observed variable phenomena is often a crucial criterion for the evaluation of the grammar(Anttila 2002). A recent improvement in this paradigm involves the consideration of non-grammatical factors in addition to linguistic factors, such as speech-genre, age, sex, education background, among others. In this paper I am concerned with the usage frequency of a word as a non-grammatical factor, which also plays an important role in the research of

usage-based account of language in cognitive linguistics, as demonstrated by the exemplar theory (Johnson 1997).

This paper examines 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., 'wend'). 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 (t/d deletion in English and geminate devoicing in borrowings in Japanese) 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. The authors raised a question of whether the same model would account for the frequency effects observed in a augmentation phonological process, which would provide additional evidence to the incorporation of frequency weight scaling factor in their model of OT grammar. In the current study, I follow their discussion and explore the following research questions: (1)are augmentation processes affected by frequency biases in a similar way as the reduction processes? (2)if so, how can we incorporate this factor into our modeling of the variable output distribution pattern?

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 Chinese (MC, 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 is described to have four possible realizations (Chao 1968:20): [ɣ][ʔ][ɦ][ŋ] in free variation (somewhat depending on dialects and place of articulation for velar variants). 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 glide 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 give an outline of the rest of the sections in this paper. In the first half of the paper I investigate the effect of usage frequency on the observed distributions of the deletion vs. epenthesis candidates by designing and carrying out a speech production experiment. This also includes acoustic analysis of BMC data to verify our statement in (3) above (i.e., the validity of deletion candidate and V#V sequence in BMC, since to my knowledge, this variation in the context of BMC has not be formally described in literature regarding its deviance from standard MC). The results of the experiment is discussed in the light of the frequency effect on reduction vs. augmentation processes and licensing-by-cue hypothesis. In the second part I focus on building a Noisy Harmonic Grammar model (NHG,closely related to stochastic OT by Boersma 1997, see Coetzee 2009;Coetzee and Pater 2011) and incorporate the frequency weight scaling factor to improve the model by running a Praat simulation of the NHG.

1. C#V Target Production Experiment Design [1]

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

[1] In this paper I adopt the view that words in BMC are monosyllabic. Therefore, my use of word boundary sign # is equivalent to syllable boundary in this case.

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 have such an influence), 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___".

In this format, the speaker is given a choice of saying verb "xihuan" (like) or "bu-xihuan"(dislike) in order to maximally induce a informal speech style (as opposed to a reading style) as they will have to focus on the semantic judgment of the sentence instead of their phonetic realization. In addition, 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 task has no easy solution at first glance. 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 two 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 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 in. Duanmu (2007) listed the following Mandarin vowel inventory:

Mandarin Vowels:
a.High: [i y u],
b.Mid: [ə o e ɤ],
c.Low: [a]

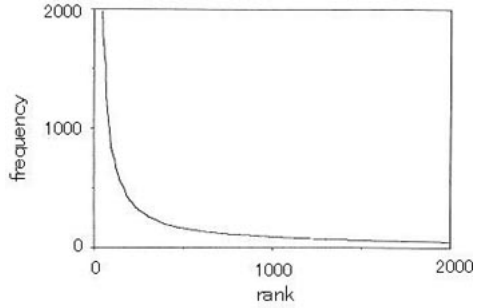
Each of which carries several variants according to their phonological positions. The high vowels are excluded from the consideration for the current task because a glide is always inserted when they follow a consonant coda (Duanmu 2007). Among the remaining mid and low vowels, mid vowels only occur in the syllable-initial position very infrequently (there are less than 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' words. I assume this observation holds across languages, in Mandarin as well, as seen in the Zipf's Law, see Figure 1), 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.

Figure 1: Zipf's Law: Word frequency as a function of the rank of the frequency rank of words in a corpus



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 sequences 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), [2] and word frequency is computed according to bisyllabic or trisyllabic word as a unit[3].

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 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 (meanwhile with a scale that is easy to see on a plot), I used the log10(count) of the words, where count is the count of the number of times the word has appeared in the corpus.

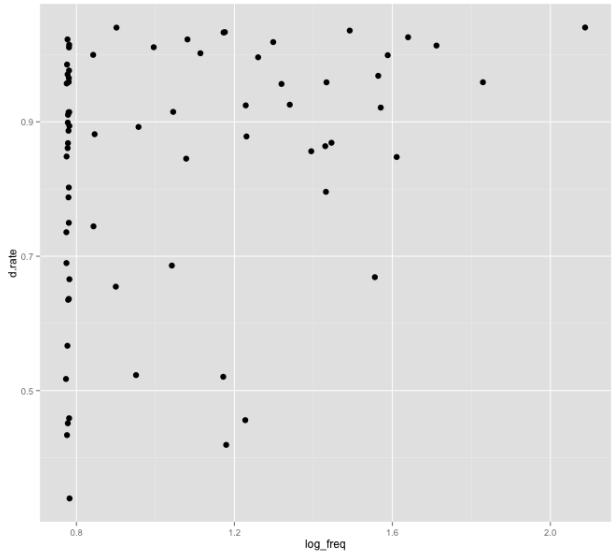
3. Results and Discussion

Figure 2 shows the results of rates of deletion (y) as a function of log frequency bins (x) (with jitter added to show overlapping data points). As deletion and epenthesis are the only two licit candidates for this output, the corresponding results for rate of epenthesis across different frequency bins is a mirror image of Figure 2 and is not shown here. In this section I discuss several aspects of the results that are observed from the data.

[2]The corpus query interface can be accessed online at (http://elearning.ling.sinica.edu.tw/CWordfreq.html).

[3]Alnternatively we can also use the strategy to compute word frequency based on monosyllabic unit, but that also raises the question of relationship between the monosyllabic morphme and the bisyllabic morpheme.

Figure 2: Deletion rate by Word frequency



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 3 and 4 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.

Figure 3: Epenthesis variant with glottal stop

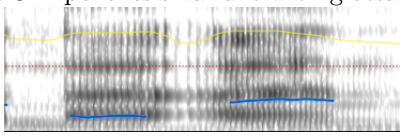
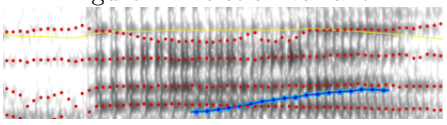
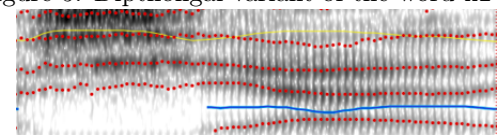


Figure 4: Deletion variant



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, see Figure 5).

Figure 5: Dipthongal variant of the word `xi'an`



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 the results in Figure 2. First, strictly speaking, the data is not linearly distributed since we see a categorical deletion in some items (Fdel=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 tend to have higher deletion rates. Also, comparing our 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 positive correlation (r=.29) 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.[4]

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 in some items 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.

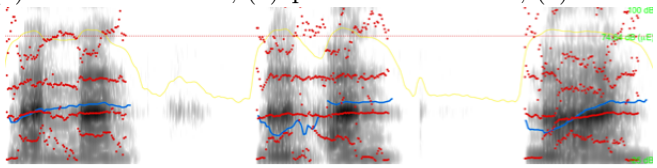
[4]This may also show that factors other than frequency are relevant for conditioning deletion. Frequency may interact with other factors - while high freq items may strongly prefer deletion, this preference may diminish at lower frequencies, with other factors (such as phonological context) being more important for determining repair. This point will be further discusses in our modeling in later sections.

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 segments in the BMC inventory[5]. 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 relevant of cues (burst and/or 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.

Figure 6: (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)[6]. To capture this distinction in output variation, we need a model that can assign different violations different scores, as in the Noisy Harmonic Grammar, which assigns different noise scores to the faithful vs. deletion candidate in the t/d deletion

[5]e.g., there is no glottal stop ~ zero contrast in BMC. There is, though, an /a/ ~ zero contrast (having an /a/ vs not having an /a/ would change the word)

[6]Strictly speaking though, the analogy only applies to the way frequency distribution behaves, not the constraint ranking itself. This is due to the fact that a wining faithful candidate violates no faithful constraints whereas an epenthesis candidate, no matter how un-salient acoustically, always violates a faithfulness constraint DEP.

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) across syllable boundaries. We'll call this constraint *RESyl. Below is a standard analysis, using two crucial rankings of MAX and DEP to produce the variation:

a.Deletion candidate wins

nan+an	*RESyl	DEP	MAX
nanan	*!		
☞nãan			*
nanøan		*!	

b.Epenthesis candidate wins

nan+an	*RESyl	MAX	DEP
nanan	*!		
nãan		*!	
☞nanøan			*

In order to better capture this variable output pattern with a consistent ranking, I propose a mock Noisy Harmonic Grammar (NHG, 'mock' here as in the values are made up for the sake of demonstration) ranking, as proposed by Smolensky and Legendre (2006). In NHG, the ranking of constraints is reflected by assigning a different numeric value to each constraint, whereas variation in output is generated by assigning a noise score unique to each constraint. We then compute the sum of violation score of each constraint (H-score) following this formula:

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

where w_i is the weight of constraint Ci , and Ci (cand) is the number of times that candidate cand violates Ci , expressed as a negative integer.

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

a.Deletion candidate wins

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

b.Epenthesis candidate wins

nan+an	w	nz	w	nz	w	nz	H
	5	-0.7	1.5	-0.3	1	-0.5	
	*RESyl(4.3)		DEP(1.2)		MAX(1.5)		
nanan	-1						-4.3
nãan					-1		-1.5
☞nanøan			-1				-1.2

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 (sf) that represent the frequency effect for the deletion vs. epenthesis rates. In this case, the incorporation of a weight scaling factor in the computation of faithfulness constraints takes into account the effect of usage frequency on deletion rates (vs. epenthesis rates)[7]. The NHG with a weight `scaling factor` is based on the following formula:

$$H(cand) = \sum_{i=1}^{n}(w_i + nz_i)M_i(cand) + \sum_{j=1}^{m}(w_j + nz_j + sf)F_j(cand)$$

Where Mi is the i-th markedness constraint, wi the weight associated with Mi , nzi the noise associated with Mi at this evaluation occasion, and Mi (cand) the number of times that cand violates Mi (expressed as a negative integer); and where Fj is the j -th faithfulness constraint, wj the weight associated with Fj , nzj the noise associated with Fj at this evaluation occasion,and Fj (cand) the number of times that cand violates Fi (expressed as a negative integer); and where sf is the scaling factor associated with the specific word being evaluated.

In the next section I carry out a computational modeling and simulation of this Noisy Harmonic Grammar(implemented in Praat), where the actual constraint ranking and disharmony scores will be learned by running a simulation in Praat. I will also discuss the effect of frequency scaling factor that improves our models of deletion vs. epenthesis rates in Mandarin zero onsets across frequency bins.

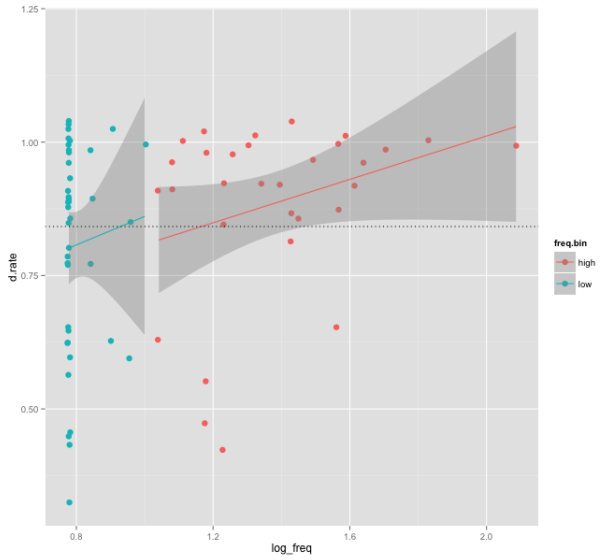
5. Machine Learning Simulation and Results

The goal of incorporating weight scaling factors into the NHG grammar is to make a model that can better predict the output distributions of deletion vs. epenthesis rates across different frequency bins so that they approach the observed patterns more closely. This builds on the problem that in a NHG model without frequency scaling, the predicted distribution has a mismatch in different frequency bins with observed patterns. This mismatch is shown in Figure 7. In this plot, we divide the frequency into two bins, and added a linear model line to represent the increase in deletion rate as a function of the log frequency of the word. Notice that the deletion value as predicted by the baseline NHG model is shown with a dotted horizontal line (around 84%), and this prediction has quite a bit of deviation from the actual observation across

[7]According to C&K 2013, the expected deletion rates in a variable process like t/d deletion correspond to the overall observed deletion rates across the entire corpus, where the prediction deviates considerably from the actual observed rates when deletion rates are viewed as a function of the word frequency. The NHG with weight scaling improves this model and gives more accurate predictions on deletion rates across different frequency bins. For more details of the model see C&K 2013.

frequency bins[8].

Figure 7: Deletion rate by Word frequency (dotted line showing expected deletionr rate without sf)



The basic assumption of weight scaling factor modeling is that there is a universal pattern that we observe cross linguistically with the frequency effects on languages (for C&K 2013, one example is that more frequent words are more likely to undergo reduction processes). The current study assumes the same with regard to reduction vs. epenthesis process (where epenthesis is a non-salient segment as in zero-onsets of BMC). A second crucial assumption is that every word is associated with a distribution function, whose shape is determied by the frequency of the word, as modeled by the `beta distribution` (Gupta and Nadarajah 2004). A well studied distribution that is similar to the normal distribution, beta distribution is a family of distributions whose exact shape is determined by its parameters α and β, with a finite range for input variable x on the interval [0,1]. [9]. In the current study I follow C&K (2013) in adopting a generalized form of Beta distribution, with three parameters α , β, and ρ, whose probability density distribution is given by the following formula:

$$f(x;\alpha,\beta,\rho) = \rho\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)dx}$$

where ρ specifies the range of the function as spanning from $-\rho$ to ρ. α and β are shape parameters that determine the skewness of the distribution. When α=β,the distribution is symmetric around zero. When α>β, it is left-skewed, and when α<β, it is right-skewed. The computation of scaling factor is given by this formula, where the constants α and β take on different values (to be determined later) and with different values of ρ. In the modeling process, we experiment with different values of ρ and in the end decide on the best value based on the

[8]Jitter is added to this plot, due to the big amount of overlap in deletion rate vlues in the current experiment.

[9]C&K 2013 suggested that an important reason that they picked Beta distribution is that comparing to normal distribution, which has a range of (-inf, +inf), beta distribution has a finite range that places an absolute limit on the influence that non-grammatical factors such as frequency can have via weight scaling.

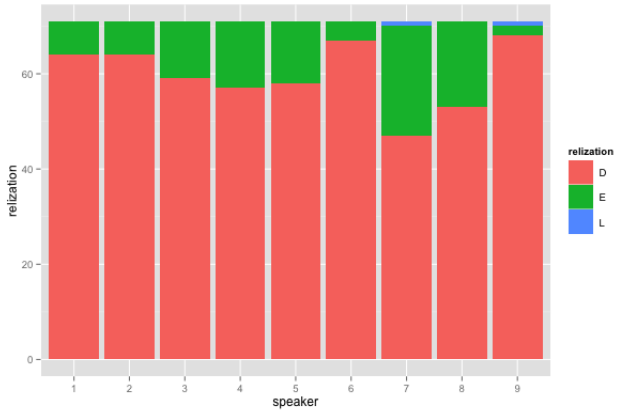
root mean squre errors of their predicted results comparing to observed distribution (of deletion rates across different frequency bins).

5.1 Noisy Harmonic Grammar Modeling and Simulatiom

5.1.1 Initial NHG (Linear OT) Grammar Learning Simulation in Praat

The first step of the modeling is to obtain our base grammar model and constraint scores in Praat by running a learning simulation of Praat's implementation of Noisy Harmonic Grammar (a.k.a Decision Strategy: Linear OT in Praat) that we could improve upon with weight scaling later on. To do this, a Praat OTGrammar object file is created[10], along with a PairDistribution object file that gives the observed distribution of deletion vs. epenthesis candidates. In the current speech-production experiment discussed above, even though faifthful candidates are theoretically prohibited, we do observe that two instances of linked variants (where the coda and the initial vowel of the next syllable are directly linked, violating the *RESyl constraint) surfaced in the pool of 639 tokens.(This distribution of individual speakers is shown in Figure 8, where D=Deletion, E=Epenthesis, L=Link/faifth candidate). In the simulation we take this distribution as the input PairDistribution pattern (Table 2, observed pattern).

Figure 8: Individual speaker variation on deletion, epenthesis, and linking



The OTGrammar file contains the three constraints *RESyl, MAX and DEP with initial score of 100 for each constraint. Decision strategy is set to `Linear OT` while all other settings are kept Praat's default in the learning simulation. Once the grammar is learned, we view the initial constraint ranking scores, shown in Table 1:

Table 1: Initial constraint ranking scores

Constraints	Score
*RESyl	107.152
DEP	101.2
MAX	97.49

[10]To see the grammar, PairDistribution, and praat script source code files used in this experiment, please visit zangsir.weebly.com/otresults.html

With the initially learned grammar, we use `To Output Distributions` to draw the expected distribution pattern, shown in Table 2 in comparison with the observed distribution:

Table 2: NHG constraint simulation of expected patterns

	observed	expected
deletion	84.05%	84.06%
epenthesis	15.65%	15.935%
faithful	0.3%	0.005%

The expected distribution in Table 2 represents the baseline prediction of the initial grammar model on overall deletion rates (with no frequency effect taken into account yet). Notice that this grammar reflects our earlier discussion that violations of DEP is ranked as worse than violations of MAX, generating the desired output distribution patterns comparable to our observation. Our next task is to compute the weight scaling factor for different frequency bins on different values of ρ.

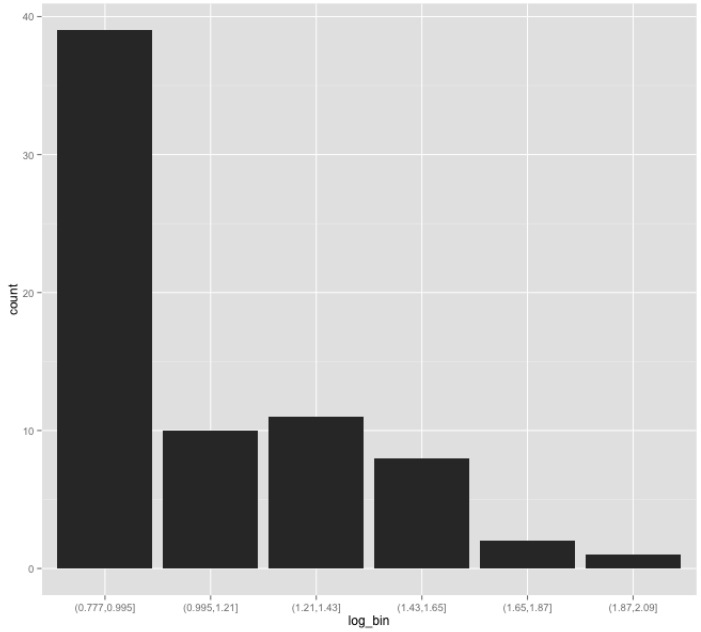
5.2 Frequency Bins and Estimation of Scaling Factors with Different Values of ρ

As discussed above, the word list involved in the C#V zero-onset production task consists of relatively low frequency words. In comparison with a more widely observed phenomenon such as the t/d deletion, the variation of zero-onset involves only a very constrained set of phonological environments that can be found in a limited number of instances in BMC. The resulting difference is that in the current data, we do not get a wide range of frequency values. As shown in the frequency distribution in Figure 9(with LaPlace smoothing to set the zero values to non-zero), most of the words in the list have a log frequency below 1 (count:less than 10 times in the corpus), whereas the highest frequency words are in the 2 bin. This posits a problem with data binning in terms of both the numbers of bins, and the binning strategy. Any attempt to bin the data into reasonable number of bins according to convention (No. of bin=log2(N)+1 or No. of bin=sqrt(N), where N is the total number of data points) would result in a extreme imbalance in number of points per bin. This is due to the large number of data points with very low frequency (appearing only once in the corpus). This will inevitably result in having only 1 or 2 data points in the higher frequency bins, which is an undesired situation when we try to incorporate the frequency weight scaling factor to account for deletion rates. For the same reason, any equal width (each bin has the same interval from lower to upper limit) or equal frequency (each bin has the same number of data points) will not work. Therefore I decided to bin the data into two bins only, which has a cut around log frequency value 1(deriving the two bins as (0,1] and (1,2]. Having decided the frequency bins to account for, I computed the deletion frequency for C#V words in the two bins(Table 3).

Table 3: Observed deletion rates arcoss frequency bins

bins (log frequency)	observed deletion rates
1	81.27%
2	91.36%

Figure 9: Histogram of Log10 frequency values (Sinica Corpus) in the zero-onset word list



Scaling factors across different bins are to be computed. In doing this, we follow C&K(2013)'s method to first establish the reference frequency in order to obtain the value of α in the beta distribution pdf (α=reference frequency, β=frequency of the word bin to compute the pdf), by using the mean log frequency of the two words right above and below the word with 50% cumulative frequency in the corpus. I computed this value as α=log10(1213)=3.1. Table 4 shows the scaling factors derived with different values of ρs (1,2,3,4,5,6,7,8,9), for the two frequency bins[11].

Table 4: scaling factors for words across frequency bins at different values of ρ

frequency bin	baseline	ρ=1	2	3	4	5	6	7	8	9
1.0	0	0.91	1.82	2.73	3.64	4.55	5.45	6.36	7.27	8.18
2.0	0	0.35	0.71	1.06	1.42	1.77	2.13	2.48	2.84	3.19

5.3 Sf-weighed Constraint Ranking: Experimenting with Different Values of ρ

Now that we have obtained different scaling factor values across different ρs, we will need to first modify our initially learned constraint ranking values (Table 1) by adding scaling factors to the values of the faithfulness constriants, for different bins. The resulting constraint scores are shown in Table 5.

[11]The computation of the scaling factor can be accessed from the spreadsheet at http://www.quantitativeskills.com/sisa/rojo/distribs.htm, with the mode being the scale factor.

Table 5: constraint score adjustment according to different sfs

Constraints	baseline	ρ=1	2	3	4	freq.bin
*RESyl	107.152	107.152	107.152	107.152	107.152	1.0
DEP	101.2	102.11	103.02	103.93	104.84	1.0
MAX	97.49	98.4	99.31	100.22	101.13	1.0
*RESyl	107.152	107.152	107.152	107.152	107.152	2.0
DEP	101.2	101.55	101.91	102.26	102.62	2.0
MAX	97.49	97.84	98.2	98.55	98.91	2.0

Constraints	baseline	5	6	7	8	9	freq.bin
*RESyl	107.152	107.152	107.152	107.152	107.152	107.152	1.0
DEP	101.2	105.75	106.65	107.56	108.47	109.38	1.0
MAX	97.49	102.04	102.94	103.85	104.76	105.67	1.0
*RESyl	107.152	107.152	107.152	107.152	107.152	107.152	2.0
DEP	101.2	102.97	97.14	103.68	104.04	104.39	2.0
MAX	97.49	99.26	99.62	99.97	100.33	100.68	2.0

Once constraint values are decided, we need to adjust the constraint values of the initial OTGrammar to derive a set of new grammars for each value of ρ and across frequency bins. For this task, I implemented a praat script to adjust the constraint values and to generate output distribution predictions for each grammar. The results of the expected distributions are shown in Table 6.

Table 6: Predicted deletion rates (%) in BMC C#V sequence at different values of ρ

bins	observed	baseline(expected)	ρ=1	2	3	4
1	81.27%	84.06%	90.46%	90.49%	90.2%	89.57%
2	91.36%	84.06%	90.344%	90.44%	90.53%	90.38%

bins	observed	baseline(expected)	5	6	7	8	9
1	81.27%	84.06%	88.51%	86.2%	81.76%	75.65%	66.62%
2	91.36%	84.06%	90.33%	90.39%	90.41%	90.26%	90.01%

In order to decide on a value of ρ that gives the model the best fit, we compute the root mean squre (RMS) error[12] for each value of ρ. The result shows that the best fit between the predicted and observed values is obtained when ρ=7. (Table 7)

Table 7: Mean square errors and percentage of improvement relative to the baseline, unscaled grammar at different values of ρ

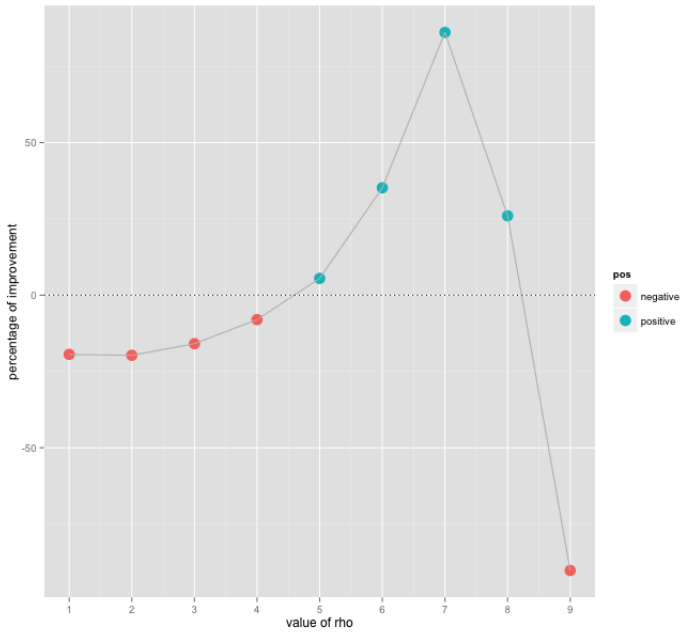
	baseline	ρ=1	2	3	4	5	6	**7**	8	9
mean square root error	7.74	9.24	9.26	8.97	8.36	7.31	5.01	**1.07**	5.73	14.72
improvement	-	-19.4%	-19.7%	-15.9%	-8%	5.5%	35.2%	**86.1%**	26.0%	-90.2%

[12]Mean square root error is calculated according to the formula E=$\sum$(Pi − Oi)^2, where Pi is the value predicted for observation i, and Oi the observed value for observation i.

5.4 Discussion

Table 7 shows that comparing to the baseline model, our new model with frequency scaling factor gives a 86.1% improvement of the predicted rate of deletion across the two different frequency bins at ρ=7. Comparing to the t/d deletion data of C&K(2013), interestingly, the current data performs worse than the baseline model when ρ is small (<5) or quite big (>8)[13]. Figure 10 shows the improvement curve of the model with a peak performance at 7. The differences in the behavior of ρ may have to do with the different patterns of frequency bins and constraint ranking scores in the current data. A most striking difference between our data and C&K's data is that, in their initial grammar, many candidates violate multiple constraints while others only violate one constraint, which results in a drastic difference in their ranking score. In our data, however, the ranking scores of all constraints are in a simlar range.This difference has consequences in the behavior of the beta distribution with different values of ρ. Finally, all of the frequency bins are well below the value of the reference frequency in the current data (F(ref)=log10(1213)=3.1), which may have a effect on the behavior of the model.

Figure 10: Improvement (%) of the model at different values of ρ



As discussed above, one thing worth discussing in the current implementation of the model to zero-onsets data is that of the choice of frequency bins. In the final model we built in this paper, even though we see a great improvement in the two frequency bins in terms of predicted vs. observed deletion rates, we nonetheless observe that it is still not a good enough representation of the actual data distribution with deletion rates. If we consider Figure 7, we see that there are a majority of points clustered at the very low end of the frequency bins (many appear only 1 time in the raw frequency of the corpus), and the wide range of deletion rate values in these low-frequency items is not well represented by our model. However, in the current study, as I already mentioned, the conundrum is that there are such imbalance in data points across frequency bins

[13]To be accurate, the C&K paper did not show results for ρ<3.

so that any binning strategy that gives abundant account of the lower end would yield very few data points in the higher frequency bins, which is not a desirable situation with data analysis. On the other hand, there is also the doubt as to whether these clustered low frequency points would show any differentiation in deletion rates that represents the effect of usage frequency. In other words, it is possible that frequency effects can only be observed beyond a threshold of scale difference. If the frequency difference that our model can account for is too fine-grained, the model runs the risk of overfitting. Future steps of improving the model might take into account this paradox.

References

[1] Anttila,Arto.(2002)Variation and phonological theory. In Handbook of language variation and change, eds. Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 206-243. Oxford:Blackwell.

[2] Boersma, Paul. (1997) How we learn variation, optionality, and probability. University of Amsterdam Institute of Phonetic Sciences Proceedings 21:43-58.

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

[4] Coetzee, Andries W. 2009a. An integrated grammatical/non-grammatical model of phonological varia- tion. In Current issues in linguistic interfaces: Volume 2, eds. Young-Se Kang, Jong-Yurl Yoon, Hyunkyung Yo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Kyoung-Ae Kim, and Hye-Kyung Kang, 267–294. Seoul: Hankookmunhwasa.

[5] Coetzee & Kawahara. (2013). Frequency biases in phonological variation. Natural Language and Linguisic Theory, 31: 47-89.

[6] Coetzee, Andries W., and Joe Pater. 2011. The place of variation in phonological theory. In Handbook of phonological theory: 2nd Edition, eds. John Goldsmith, Jason Riggle, and Alan Yu, 401–434. Cambridge: Blackwell.

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

[8] Gupta, Arjun K., and Saralees Nadarajah, eds. (2004). Handbook of the beta distribution and its applications. New York: Marcel Dekker.

[9] Johnson, Keith. (1997). Speech perception without speaker normalization: an exemplar model . In K. Johnson & J. Mullennix (Eds.), Talker variability in speech processing (pp. 145–165). Academic Press.

[10] Lahiri, Aditi, Letitia Gewirth, and Sheila E. Blumstein. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. The Journal of the Acoustical Society of America 76: 391–404.

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

[12] Smolensky, Paul, and Géraldine Legendre, eds. (2006). The harmonic mind: From neural computation to Optimality-Theoretic grammar, Volume 1: Cognitive architecture, Volume 2: Linguistic and philosophical implications. Cambridge: MIT Press.

[13] Steriade, Donca. (1999). Phonetics in phonology: The case of laryngeal neutralization. In UCLA working papers in linguistics 2 (Papers in phonology 3), ed. Matthew K. Gordon, 25–146. Los Angeles:Department of Linguistics, UCLA.

[14] Steriade, Donca. (2001). Directional asymmetries in place assimilation. In The role of speech perception in phonology, eds. Elizabeth Hume and Keith Johnson, 219–250. San Diego: Academic Press.

[15] Stevens, Kenneth N., and Sheila E. Blumstein. (1978). Invariant cues for place of articulation in stop conso- nants. The Journal of the Acoustical Society of America 64: 1358–1368.

[16] Stevens, Kenneth N., and Samuel J. Keyser. (1989). Primary features and their enhancement in consonants. Language 65: 81–106.