Probability Distributions of Sounds and Phonotactics in Taiwan Mandarin Syllables

I-Ping Wan[1], Chiung-Wen Chang[2], Chainwu Lee[3], Pu Yu[2*]
[1] Graduate Institute of Linguistics/Research Center for Mind, Brian, and Learning/Program in Teaching Chinese as a Second Language, Phonetics and Psycholinguistics Laboratory, National Chengchi University, Taipei, Taiwan
[2] Graduate Institute of Linguistics, Phonetics and Psycholinguistics Laboratory, National Chengchi University, Taipei, Taiwan
[3]Phonetics and Psycholinguistics Laboratory, National Chengchi University, Taipei, Taiwan
[1] ipwan@g.nccu.edu.tw
[2] 111555006@g.nccu.edu.tw
[3] chainwu_lee@yahoo.com
[2*] acadyupu@gmail.com

Abstract

This study examines the influence of phonotactic probabilities, phonological structures and articulatory complexity on speech production in Mandarin. By analyzing a natural spoken corpus comprising 202 hours of daily conversation in Taiwan Mandarin, which includes 2,384,567 lexical items and yields 6,272,394 tokens involving 3,852,987 consonant tokens and 2,419,407 vowel tokens, the dataset is precisely categorized into 12 syllable structure types. The study employs frequency-based probabilistic phonotactics, with probability distributions calculated using Zipf's Law and Yule's distribution, where Yule's distribution provides a better prediction for the segment distribution. Phonotactic probabilities are further determined by the bigram or biphone frequencies of phonological segments and sequences within Mandarin word types. The results reveal a departure from previous research that found a strong correlation between speech production, phonological structure and articulatory complexity, such as markedness in phones or syllable structures. Instead, Taiwan Mandarin speakers demonstrated sensitivity to frequency variations, with phonotactic probabilities independently influencing speech production, suggesting that these probabilities are encoded within speech production processes. This research contributes to the understanding of how phonotactic constraints, independent of articulatory complexity, shape speech production in Mandarin.

1 Introduction

It is generally believed that speakers can process certain sound sequences faster than others. The possible sound sequences in languages are not all equiprobable as some are more frequent than others. The increasing variety of approaches to probability in phonology indicates a growing consensus that phonological analysis needs to incorporate probability and frequency into the theoretical framework (Alderete and Finley, 2023). Therefore, phonological complexity and probabilistic constraints are essential concepts in the study of natural languages. Their strong correlation significantly influences various aspects of linguistic theory and practice. Articulatory complexity refers to the intricate features of a language's sound system, including the number and types of phonemes, syllable structures, and phonotactic rules.

A number of researchers suggested that certain sound sequences have attributed similar behavioral effects that are easier to articulate (i.e., less phonological complexity), but others attributed the patterning to varying degrees of probabilistic constraints (e.g., Jusczyk et al., 1994). Such constraints can be referred to as phonotactic probabilities where phonological phones and sound sequences are legally arranged in lexical items. For example, in English, the initial sequence [str] is allowable whereas the sequence [stn] does not form a legal arrangement. Or, in Mandarin, the initial sequence [kwa] is permissible while the sequence [kja] or [kwn] is not. In addition, the single phone unit in the above phone sequences does not distribute evenly. The glide [w] or [j] occurs more frequently than the consonant [k] in Mandarin due to the fact that glides have a wider distribution (i.e., syllable-initially, syllable-medially, and syllable-finally) than the true consonant [k] (syllable-initially exclusively) (Wan, 2022).

In experiments by Goldrick and Larson (2008), English speakers were sensitive to variations in frequency, demonstrating that phonotactic probabilities are encoded by speech production processes. These novel phonotactic constraints were found to be correlated with the phonotactic probability of specific phonological structures. However, other research has shown a highly correlated association between speech production and phonological structure and articulatory complexity such as markedness in phones or syllable structure (e.g., Jakobson, 1941/1968; Romani and Calabrese, 1998). Evidence from these studies presents a limited number of structures that have yielded mixed and uncertain findings.

Further studies have found that phonotactic probabilities exhibit a strong correlation with neighborhood density, which refers to the number of lexical items that share phonological similarity with a target (e.g., Goldrick and Rapp, 2007; Vitevitch et al., 2004). These effects manifest at separate and independent levels within the spoken production system. In this study, we aim to compute frequency-based probabilistic phonotactics in Mandarin syllables by categorizing a spoken dataset into 12 syllable structure types via Biphone/Phone or Bigram/Gram frequencies (i.e., segment-to-segment co-occurrence probability of sounds within the lexical items; Vitevitch and Luce, 2004), with tone omitted from the calculation. In addition, the effects of phonotactic probabilities and likelihood will be measured across the different syllable structure types.

2 Methodology

The spoken data used in the study that has been collected over decades were drawn from Wan et al. (2024) involving 202 hours of daily conversation in Taiwan Mandarin involving 2,384,567 lexical items. The topics of the recorded spoken content that were recorded in a naturalistic setting varied from lecture notes, class discussions, interviews, presentations, conversations of daily lives, etc., among multiple speakers in Taiwan.

Sound files collected after 2020 were transcribed into the International Phonetic Alphabet (IPA) via Chinese characters using a Speech-to-Text (STT) system. This system was developed using the pyTranscriber application (https://github.com/raryelcostasouza/pyTranscriber) in the Phonetics and Psycholinguistics Laboratory. Transcribing a 60-minute audio file into Chinese characters took approximately 80 seconds. However, the accuracy of the transcription varied significantly, depending on factors such as voice quality, background noise, speaker gender, age, and speech speed. The accuracy rate ranged between 70% and 90%, depending on the combination of these factors. The output of the STT system was then manually checked for accuracy. Subsequently, the entire transcript was automatically segmented by the CKIP parser (Ma and Chen, 2003) and POS tagged by the CKIP tagger from the Chinese Knowledge and Information Processing group (CKIP, 1998). The parsed and tagged transcription was also manually reviewed according to the word segmentation and POS tagging criteria of the Academia Sinica Corpus (CKIP, 1998), which are commonly applied in corpora such as the Linguistic Data Consortium (Ma and Huang, 2006) and the Peking University corpus (Huang et al, 2008).

It is important to note that the spoken data samples collected in this study were analyzed based on the frequency of occurrence across various topics recorded in naturalistic settings. Word counts are up to date and are not derived from movie subtitles. The following (1) and (2) shows

$$F_r = \frac{a}{r^b} \qquad (1)$$

the formula for calculating the frequency distribution and probability in Mandarin.

Formula (1) represents the mathematical expression of Zipf's law (Zipf, 1949). It describes the frequency distribution of words or other linguistic units, where the frequency of the most common unit (such as a word or phoneme) is inversely proportional to its rank in the entire corpus. In other words, the highest-ranking word or phoneme has the greatest frequency, the second-ranking unit has approximately half the frequency of the first, and this pattern continues accordingly. In the function, *r* represents the rank of an item, and *Fr* is its frequency. *a* is a constant, typically representing the frequency of the highest-ranked item; b is a constant that describes the inverse relationship between frequency and rank.

$$F_r = \frac{a}{r^b} C^r \qquad (2)$$

Equation (2), Yule equation (Yule, 1924), is similar to Zipf's Law. However, the Yule equation incorporates an additional exponent, *Cr*, which accounts for the dominance of a few highly frequent distributions. Specifically, Yule's distribution is a discrete probability distribution used to model the frequency of particular distributions, reflecting the underlying processes, whereas Zipf's Law focuses on rank-order distributions. Both Zipf's Law and the Yule equation have demonstrated a relatively high degree of fit in past research concerning sound distribution (e.g. Kłosowski, 2017; Tambovtsev and Martindale, 2007). Therefore, this study employs these two formulas to examine the phonetic distribution within the dataset.

Using a corpus-based and data-driven analysis to investigate the probability of sound frequency represents a recent trend in speech communication, language learning and psycholinguistic experiments (Wan et. al., 2024; Hsieh and Wan, to appear; Chien and Wan, 2023; Wan, 2021). Therefore, the questions to be investigated involve the following:

- What is the distribution pattern of speech tokens in Mandarin? Will consonants, vowels or glides be distributed evenly?

- Are the behavioral effects of certain sound sequences due to lower or higher phonological complexity, or do they result from varying degrees of probabilistic constraints?

- How do phonotactic probabilities influence the legality of sound sequences in Mandarin? How do they impact and are encoded by speech production processes? What is the relationship between phonotactic probabilities and articulatory complexity? How do phonotactic probabilities correlate with phonological structures and articulatory complexity, such as markedness in phones or syllable structures?

- How can frequency-based probabilistic phonotactics in Mandarin be computed and analyzed? How will these effects be measured in the study?

3 Results and Discussions

Token counts and probability of sound frequency using a log 10 frequency distribution were extracted from 202 hours of daily conversation involving consonants (N= 3,852,987 tokens) and vowels (N=2,419,407 tokens) in Taiwan Mandarin, as shown in Table 1 and Table 2.

Consonants are distinguished by three primary parameters involving place of articulation, manner of articulation and voicing (voiced vs. voiceless). One of the key distinctive features in Mandarin is the use of aspiration to differentiate six minimal pairs: [p/p^h, t/t^h, k/k^h, tʂ/tʂ^h, ts/ts^h, tɕ/tɕ^h]. In each pair, the first consonant is unaspirated, while the second is aspirated. Aspiration in Mandarin is a significant phonological feature, where the presence or absence of a burst of breath below following the consonant can change the meaning of a word entirely. In addition, three series of affricates and fricatives, including voiced, voiceless unaspirated, and voiceless aspirated features, [ʐ, tʂ, tʂ^h, ʂ], [ts, ts^h, s] and [tɕ, tɕ^h, ɕ], occur in consonant inventory. The inclusion of voiced fricatives such as [ʐ] is relatively rare in Mandarin, with most of the fricatives and affricates being voiceless. The log frequency analysis shows that the distribution pattern of the single phone units in Taiwan Mandarin is uneven. For example, the glide [w] or [j] occurs more frequently than the consonant [k] in Mandarin due to their wider distribution (i.e., syllable-initially, syllable-medially, and syllable-finally) compared to the consonant [k], which occurs exclusively in syllable-initially position. This results in a highly structured and distinctive phonological system that does not correspond to traditional markedness in phones.

A major distinction among Mandarin vowels involves differences in tongue height, anterior-posterior tongue position, and lip rounding. Consistent with previous findings, all the single vowel units are not distributed evenly. The following bar chart illustrates the rank order of frequency for each single phone unit.

	Bilabial	**Labio-dental**	**Dental**	**Retroflex**	**Palatal**	**Velar**
Plosive (Unaspirated)	p 97 191 (4.99)		t 252 633 (5.40)			k 137 394 (5.14)
Plosive (Aspirated)	p^{h} 16 720 (4.22)		t^{h} 90 879 (4.96)			k^{h} 49 864 (4.70)
Fricative		f 36 249 (4.56)	s 36 913 (4.57)	ʂ / ʐ 210 908 / 49 284 (5.32) / (4.69)	ɕ 104 968 (5.02)	x 136 644 (5.14)
Affricate (Unaspirated)			ts 75 612 (4.88)	tʂ 124 869 (5.10)	tɕ 152 966 (5.18)	
Affricate (Aspirated)			ts^{h} 19 220 (4.28)	tʂ^{h} 39 001 (4.59)	tɕ^{h} 58 030 (4.76)	
Nasal	m 108 959 (5.04)		n 495 299 (5.69)			ŋ 247 892 (5.39)
Liquid			l 97 524 (4.99)			
Glide	(w) (ɥ)				j / ɥ 586 407 / 37 283 (5.77) / (4.57)	w 590 278 (5.77)

Table 1: Mandarin consonant phones.

	Front		**Central**	**Back**	
	Unround	**Round**	**Unround**	**Unround**	**Round**
Close (High)	i 311 464 (5.49)	y 34 587 (4.54)	ɨ 188 426 (5.28)	u 101 988 (5.01)	
Close-mid (Mid)	e 111 487 (5.05)		ə 175 696 (5.24) ɚ 8846 (3.95)	ɤ 302 544 (5.48)	o 221 937 (5.35)
Open-mid (Lower Mid)	ɛ 164 202 (5.22)			ɔ 164 040 (5.21)	
Open (Low)			a 634 190 (5.80)		

Table 2: Mandarin vowel phones.

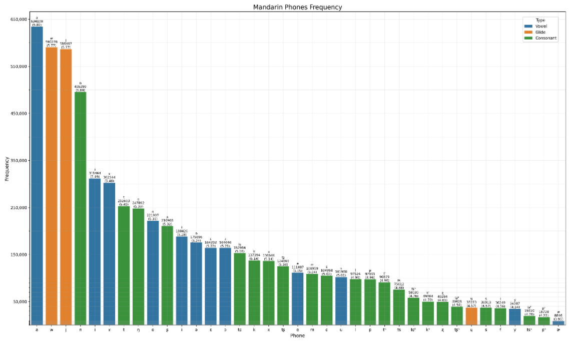

Figure 1: Mandarin phone frequency.

In this figure, the bars contain three colors: blue representing vowels, orange representing glides, and green representing consonants. It is clearly seen that the vowel [a] occurs most frequently in daily conversation in Taiwan Mandarin, followed by the glides [w] and [j], with the nasal [n] being the next most common sound. The least common vowel is the retroflex vowel [ɚ], and the least common consonant is [pʰ], followed by [tsʰ]. This distribution partially reflects the syllable structure of Mandarin, where CGVX can occur; X can be either the nasal [n] or the glides [j, w]. The glides can occur word-initially, word-medially after true consonants, and word-finally, while the nasal [n] can occur both word-initially and word-finally. Since Taiwan Mandarin does not use Erhua syllables, the retroflex vowel [ɚ] is rarely used and is commonly replaced by the vowel [ɤ]. The following shows the distribution according to two statistical quantifier measurements.

Figure 2 illustrates the phone frequency distribution of Mandarin on a log-log scale. The figure presents spoken data points alongside fitted curves using two models that include Zipf and Yule.

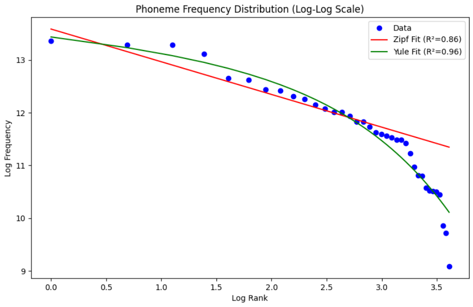

Figure 2: Mandarin phone frequency distribution (log-log scale) with Zipf and Yule fit curves.

The blue dots represent the empirical phone frequency data, while the red and green lines correspond to the Zipf and Yule fits, respectively. Compared to these two models, the Zipf fit, with a correlation coefficient (R) of 0.86, initially follows the data but diverges as the rank increases, suggesting that this model may not fully capture the distribution of less frequent phones. In contrast, the Yule fit, with an R value of 0.96, aligns closely with the data across the entire range, providing a more accurate representation of the phone frequency distribution. The higher R value of the Yule fit signifies a stronger correlation and better explanation for the observed data. Therefore, at this stage, the Yule model appears to be more suitable for representing this distribution.

Mandarin is analyzed as having a range of possible phonetic (i.e., surface) syllables: V, CV, GV, VG, VN, CVG, CVN, CGV, GVG, GVN, CGVG, and CGVN. The maximal syllable is CGVX, with C a [+consonantal] segment, G a glide, V the nucleus vowel, and X either a nasal of a glide (i.e., Wan 1999). The samples of types and token frequencies of a syllable structure, CGVN, in Mandarin are shown in Table 3.

IPA	Freq.	IPA	Freq.	IPA	Freq.
ɕjaŋ	18031	swan	1727	ʂwən	337
ɕjɛn	14461	tʂwan	1498	lwan	324
tɕjaŋ	11453	tʂʰwan	1496	tɕʰɥən	307
mjɛn	10427	tɕʰjaŋ	1337	tɕɥɛn	287
pjɛn	9907	kwaŋ	1226	kʰwan	281
tɕʰjɛn	8286	tʂwaŋ	1171	ʐwan	228
tɕjɛn	8224	kʰwaŋ	1045	twən	228
tjɛn	8019	ɕɥən	972	swən	216
njɛn	7820	xwaŋ	724	tswən	214
tʰjɛn	6324	tʂwən	688	njaŋ	118
ljaŋ	5760	xwən	682	tɕʰjoŋ	115
kwan	5689	ɕjoŋ	641	kwən	55
tɕʰɥɛn	3705	tʂʰwaŋ	619	nwan	54
xwan	2990	tsʰwən	576	tʰwən	49
ljɛn	2766	tɕɥən	466	ʐwən	39
ɕɥɛn	2452	tʂʰwən	453	ʂwan	27
twan	2096	tʰwan	378	tswan	20
lwən	1874	ʂwaŋ	361	tsʰwan	3
pʰjɛn	1806	kʰwən	348	tɕjoŋ	1

Table 3: Samples of CGVN in IPA and token frequencies.

$$PhonProb_{[\text{ɕ}ja\text{ŋ}]} = \frac{1}{n}\sum_{i=1}^{n}\frac{log(freq(S_i))}{log(freq(P_i))} = \frac{1}{3}\sum_{i=1}^{3}\frac{log(freq(S_i))}{log(freq(P_i))} =$$

$$[\frac{Sum\ of\ log\ frequencies\ of\ words\ with\ [\text{ɕ}j]\ inintial\ biphone\ position}{Sum\ of\ log\ frequencies\ of\ words\ with\ any\ biphone\ in\ intial\ biphone\ position} + \frac{Sum\ of\ log\ frequencies\ of\ words\ with\ [ja]\ in\ second\ biphone\ position}{Sum\ of\ log\ frequencies\ of\ words\ with\ any\ biphone\ in\ second\ biphone\ position} + \frac{Sum\ of\ log\ frequencies\ of\ words\ with\ [a\text{ŋ}]\ in\ third\ biphone\ position}{Sum\ of\ log\ frequencies\ of\ words\ with\ any\ biphone\ in\ third\ biphone\ position}]/3 =$$

$$[\frac{log(freq(\text{ɕ}ja\text{ŋ}))+log(freq(\text{ɕ}j\text{ɛ}n))+\cdots}{log(freq(\text{ɕ}ja\text{ŋ}))+log(freq(\text{ɕ}j\text{ɛ}n))+log(freq(t\text{ɕ}ja\text{ŋ}))+log(freq(mj\text{ɛ}n))+log(freq(pj\text{ɛ}n))+\cdots} + \frac{log(freq(\text{ɕ}ja\text{ŋ}))+log(freq(t\text{ɕ}ja\text{ŋ}))+log(freq(lja\text{ŋ}))+log(freq(t\text{ɕ}^{\text{h}}ja\text{ŋ}))+\cdots}{log(freq(\text{ɕ}ja\text{ŋ}))+log(freq(\text{ɕ}j\text{ɛ}n))+log(freq(t\text{ɕ}ja\text{ŋ}))+log(freq(mj\text{ɛ}n))+log(freq(pj\text{ɛ}n))+\cdots} + \frac{og(freq(\text{ɕ}ja\text{ŋ}))+log(freq(t\text{ɕ}ja\text{ŋ}))+log(freq(lja\text{ŋ})+log(freq(t\text{ɕ}^{\text{h}}ja\text{ŋ}))+log(freq(kwa\text{ŋ}))+log(freq(t\text{ʂ}wa\text{ŋ}))+log(freq(k^{\text{h}}wa\text{ŋ}))+\cdots}{log(freq(\text{ɕ}ja\text{ŋ}))+log(freq(\text{ɕ}j\text{ɛ}n))+log(freq(t\text{ɕ}ja\text{ŋ}))+log(freq(mj\text{ɛ}n))+log(freq(pj\text{ɛ}n))+\cdots.}]/$$

$$3 = [\frac{log(18031)+log(14461)+\cdots}{log(18031)+log(14461)+log(11453)+log(10427)+log(9907)+\cdots} + \frac{log(18031)+log(11453)+log(5760)+log(1337)+\cdots}{log(18031)+log(14461)+log(11453)+log(10427)+log(9907)+\cdots} + \frac{log(18031)+log(11453)+log(5760)+log(1337)+log(1226)+log(1171)+log(1045)+\cdots}{log(18031)+log(14461)+log(11453)+log(10427)+log(9907)+\cdots}]/3 = [\frac{26.18886307683207}{1225.5136668570954} + \frac{59.79234887458323}{911.7019871418528} + \frac{34.658475044915086}{263.65257960492846}]/3 = 0.07280267170780302 \tag{3}$$

Table 3 includes only tokens of CGVN syllables, although Mandarin features additional tokens with various syllable structures in the spoken dataset for CGVX syllables, showing all possible sound sequences and their token frequencies in Mandarin (note that tone is excluded from this study). Formulas (3) and (4) demonstrate the calculation of bigram/biphone phonotactic probability in Mandarin, using the CGVN syllable [ɕjaŋ] as an example, which yields a probability value of 0.073.

$$PhonPrab_{[\text{ɕ}ja\text{ŋ}]} = 0.07280267170780302 \cong 0.073 \tag{4}$$

Formula (3) and (4) demonstrates the calculation of the phonotactic probability for [ɕjaŋ]. In [ɕjaŋ], there are three biphones: the initial biphone [ɕj], the second biphone [ja], and the third biphone [aŋ]. The formula calculates the average positional probability of these three biphones. For the first biphone position, it sums the log10 frequencies of all words beginning with [ɕj] (e.g., [ɕjaŋ], [ɕjɛn], and others) and divides this by the sum of the log frequencies for words containing the first biphone sequence. For the second biphone position, it sums the log10 frequencies of all words containing [ja] in the second position (e.g., [ɕjaŋ], [tɕjaŋ], [ljaŋ], [tɕʰjaŋ], and others) and divides this by the sum of the log frequencies for words containing the second biphone sequence. For the third biphone position, it sums the log10 frequencies of all words containing [aŋ] in the position (e.g., [ɕjaŋ], [tɕjaŋ], [ljaŋ], [tɕʰjaŋ], [kwaŋ], [tʂwaŋ], [kʰwaŋ], and others) and divides this by the sum of the log frequencies for words containing the third biphone sequence. Finally, the average of these ratios is calculated, resulting in a phonotactic probability of 0.07280267170780302, which can be approximated to 0.073.

In this model, the phonotactic probability is calculated for a given syllable using the token frequencies and a dataset of word types that involve different syllable structures. Initially, the syllable is segmented into a series of bigrams, which represent pairs of adjacent units. Subsequently, for each position within the syllable, the model computes two sums involving one for the logarithm of the same bigram occurring at the position and another for the logarithm of the frequency of all bigrams at that position. The phonotactic probability of the syllable is determined by summing the ratio of these two sums for each bigram in the syllable and dividing by the total number of bigrams in the syllable. This ratio reflects the relative frequency of each bigram in its specific position, as shown below. When a syllable contains only a single vowel, its phonotactic probability is calculated as the ratio of the vowel's logarithmic frequency to the total logarithmic frequency of all single sound syllables.

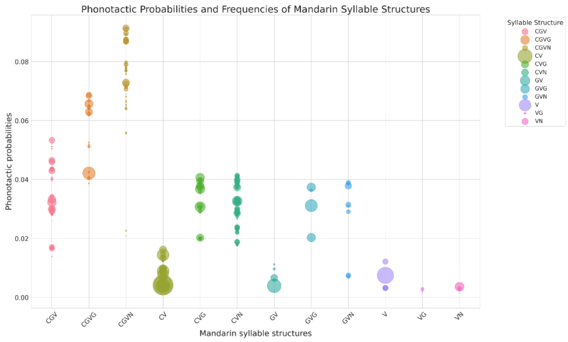

Figure 3: Distribution of phonotactic probabilities across Mandarin syllable structures**.**

In this figure, the x-axis displays all possible Mandarin syllable structures, including combinations of consonants (C), vowels (V), glides (G), and nasals (N) in legal sound sequences. The y-axis illustrates the phonotactic probabilities, reflecting the likelihood of each syllable structure occurring in Mandarin. Each colored dot corresponds to a specific syllable structure, with the size of the dot indicating its relative frequency or prevalence within the language. Larger dots signify more common syllable structures, while smaller dots represent less frequent ones. Among the structures with four legal sound units, CGVN exhibits the highest phonotactic probability, with a cluster of large dots in the 0.09-0.10 range, followed by CGVG. This is evident based on the formula, where the inclusion of four sound sequences can generate a higher probability (i.e., N+1). In contrast, the CV structure, despite having lower phonotactic probabilities, indicates the most frequent syllable structure in Taiwan Mandarin. The VG structure displays the lowest probabilities and small dot sizes, highlighting its relative rarity in Taiwan Mandarin.

In conclusion, the distribution pattern of speech tokens in Mandarin reveals an uneven distribution of segments, reflecting the legitimate structures within Mandarin syllables. The study suggests that the performance or behavior effects of certain sound sequences are primarily influenced by probabilistic constraints rather than articulatory complexity, such as markedness. Taiwan Mandarin speakers demonstrate sensitivity to frequency variations, with phonotactic probabilities playing a crucial role in shaping speech production, independent of articulatory complexity. These probabilities influence the legality of sound sequences by determining the likelihood of specific phonological segments and sequences within word types. The study further indicates that phonotactic probabilities are encoded within speech production processes and operate independently from traditional measures of phonological/articulatory complexity. While previous research emphasized a strong correlation between speech production and articulatory complexity, this study finds that phonotactic probabilities have a distinct and independent impact. The analysis of frequency-based probabilistic phonotactics in Mandarin, computed using Zipf's Law and Yule's distribution, highlights the importance of these probabilistic constraints in influencing speech production, as evidenced by the examination of a natural spoken corpus of daily conversations.

In this study, we examine the phonotactic probability distribution calculated in a given Mandarin syllable using the token frequencies and a dataset of word types involving different syllable structures. Type and token frequencies in the current spoken data confirm the studies found in English where the possible sound sequences are not all equiprobable as some are more frequent than others. More importantly, certain sound sequences are related to probabilistic constraints and do not fall in the articulatory complexity since the CV-type structure is supposed to be the easiest pattern at a more flexible range, whereas its phonotactic probability is the lowest. The study suggests that phonotactic constraints in Mandarin disassociate articulatory complexity and phonotactic probabilities influence speech production regardless of the markedness complexity. The spoken samples via data computation confirm an emerging agreement within the field that phonological theories need to consider phonotactic probabilities.

4 Limitations

A limitation of the current study is that the Levenshtein edit distance needs to be measured to further calculate neighborhood density. Neighborhood density refers to the number of words that sound similar to a target word. Words with a sparse neighborhood are generally recognized more quickly and accurately, while those with a dense neighborhood may be recognized more slowly and less accurately. Future research should investigate neighborhood density in Mandarin, focusing on how sound-similar words are stored in the mental lexicon.

Acknowledgments

We sincerely appreciate the valuable and constructive comments from the three anonymous reviewers and the editor, which have significantly improved this manuscript. We are also grateful to the audience who attended the conference. All remaining errors in the analysis and interpretation are solely our own. This research was supported in part by the National Science and Technology Council in Taiwan (MOST 111-2410-H-004-091) to the first author.

References

Chu-Ren Huang, Lung-Hao Lee, Wei-guang Qu, Jia-Fei Hong, and Shiwen Yu. 2008. Quality Assurance of Automatic Annotation of Very Large Corpora: a Study based on heterogeneous Tagging System. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2725-2729. Marrakech, Morocco. European Language Resources Association (ELRA). https://aclanthology.org/L08-1106/.

CKIP. 1998. *Academia Sinica Balanced Corpus (Version 3) [CD-ROM]*. Taipei: Chinese Knowledge and Information Processing Group, Academia Sinica.

Cristina Romani and Andrea Calabrese. 1998. Syllabic constraints in the phonological errors of an aphasic patient. *Brain and Language*, 64(1):83-121. https://doi.org/10.1006/brln.1998.1958.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Addison-Wesley.

George Udny Yule. 1924. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. *Philosophical Transactions of the Royal Society of London Biological Sciences*, 213: 21-87. https://doi.org/10.1098/rstb.1925.0002.

I-Ping Wan. 1999. *Mandarin phonology: Evidence from speech errors (Order No. 9943387)* [State University of New York at Buffalo]. Available from ProQuest Dissertations & Theses A&I; ProQuest Dissertations & Theses Global. (304551099)

I-Ping Wan. 2021. Interlanguage tone patterns in Thai pre-school children: A preliminary corpus analysis. *Taiwan Journal of Chinese as a Second Language*, 22(1):1-33. https://doi.org/10.29748/TJCSL.202106_(22).0001.

I-Ping Wan. 2022, April 15. Error analysis in Mandarin corpus phonology, [Invited talk in the Institute of Linguistics at Academia Sinica, Taipei, Taiwan]. Academia Sinica Institute of Linguistics Phonetics Laboratory.

I-Ping Wan, Marc Allassonnière-Tang, and Pu Yu. 2024. Early Segmental Production in Thai Preschool Children Learning Mandarin. *International Journal of Asian Language Processing*. 34(2): 2450005-1-2450005-22 https:///dx.doi.org/10.1142/S271755452450005X.

John Alderete and Sara Finley. 2023. Probabilistic phonology: A review of theoretical perspectives, applications, and problems. *Language and Linguistics*, 24(4):565-610. https://doi.org/10.1075/lali.00141.ald.

Matthew Goldrick and Brenda Rapp. 2007. Lexical and post-lexical phonological representations in spoken production. *Cognition*, 102(2):219-260. https://doi.org/10.1016/j.cognition.2005.12.010.

Matthew Goldrick and Meredith Larson. 2008. Phonotactic probability influences speech production. *Cognition*, 107(3):1155-1164. https://doi.org/10.1016/j.cognition.2007.11.009.

Michael S. Vitevitch and Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3):481-487. https://doi.org/10.3758/BF03195594.

Michael S. Vitevitch, Jonna Armbrüster, and Shinying Chu. 2004. Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):514-529. https://doi.org/10.1037/0278-7393.30.2.514.

Peter W. Jusczyk, Paul A. Luce, and Jan Charles-Luce. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5):630-645. https://doi.org/10.1006/jmla.1994.1030.

Piotr Kłosowski. 2017. Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling. *EURASIP Journal on Audio, Speech, and Music Processing*, *2017*:1-16. https://doi.org/10.1186/s13636-017-0102-8.

Roman Jakobson. 1941/1968. *Child Language Aphasia and Phonological Universals*. (A. R. Keiler, Trans.) The Hague: Mouton.

Wei-Yun Ma and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-5)*, pages 2182-2185, Genoa, Italy. European Language Resources Association (ELRA). https://aclanthology.org/L06-1163/.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Sapporo, Japan. Association for Computational Linguistics. https://aclanthology.org/W03-1726.

Yu-Fu Chien and I-Ping Wan. 2023. Production of Mandarin tones by Thai preschool children. Paper presented at the 14th Annual Pronunciation in Second Language Learning and Teaching, Purdue University, USA.

Yun-Shan Hsieh and I-Ping Wan. (To appear) A study on continued word association responses in Mandarin. Chinese Lexical Semantics: Lecture Notes in Computer Science. Springer, Singapore.

Yuri Tambovtsev and Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, 4(2):1-11.

A Supplementary Material

A link to supplementary materials is provided as follows: https://osf.io/qczh2/.