Phonotactic probabilities in Mandarin syllables

Anonymous ACL submission

Abstract

In this study, a spoken data sample is computed via frequency-based probabilistic phonotactics in Mandarin syllables by categorizing the spoken dataset into 12 syllable structure types. Phonotactic probabilities are measured by the bigram or biphone frequencies with which phonological segments and phone sequences occur in word types in Mandarin. Spoken data drawn from 2,384,567 lexical items show that correlations between speech production measures and phonological/articulatory complexity are not found. Instead, phonotactic probabilities influence speech production processes in Mandarin speakers independent of phonological complexity.

Keywords: Mandarin spoken corpus, phonotactic probability, bigram/biphone frequency, speech production

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. Some 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. Such constraints can be referred to as phonotactic probabilities where phonological phones and sound sequences are legally arranged in lexical items (Jusczyk et al., 1994). 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. Additionally, 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 Goldrick and Larson's (2008) experiments, English speakers were sensitive to variations in frequency, demonstrating that phonotactic probabilities are encoded by speech production processes. Such 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/articulatory complexity such as markedness in phones or syllable structure (e.g., Jakobson, 1941/1968; Romani and Calabrese, 1998). Evidence from these studies presented a rather small number of structures that yielded mixed and uncertain findings. The increasing variety of approaches to probability in phonology indicates a growing agreement that phonological analysis needs to incorporate probability and frequency into the theoretical framework (Alderete and Finley, 2023). Therefore, determining whether phonological/articulatory complexity or probabilistic constraints having a strong correlation in natural languages is not straightforward.

A number of studies further found that phonotactic probabilities exhibit a strong correlation with neighborhood density, which refers to the quantity 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 are going to compute frequency-based probabilistic phonotactics in Mandarin syllables by categorizing the 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) and tone is omitted in the calculation. In addition, the effects of phonotactic probabilities, the likelihood of occurrence of a phone sequence, will be measured along with the syllable structure types.

2 Methodology

Spoken data in the study that has been collected over decades were drawn to be discussed from the first author's lab, Phonetics and Psycholinguistics Laboratory (N= 2,384,567 syllables, 202 hours). The topics of the spoken content that were recorded in a naturalistic setting varied from lecture notes, class discussions, interviews, presentations, conversations of daily lives, etc., between multiple speakers in Taiwan. The sound files collected after 2020 were sent to a Speech-to-Text (STT) system for transcription into the International Phonetic Alphabet (IPA) via Chinese characters; the STT Package was developed from the application pyTranscriber (`https://github.com/raryelcostasouza/pyTranscriber`) in the Phonetics and Psycholinguistics lab. The transcription in Chinese characters of a 60-minute audio file took 80 seconds. The accuracy of the output can vary a lot, depending on factors such as voice quality, noise clarity, gender, age, and/or speech speed of speakers. The accuracy rate varied between 70% and 90% depending on the combination of these factors. The output of the STT system was then manually checked. Afterward, 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 checked manually according to the criteria of word segmentation and POS tagging 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 noted that spoken data samples collected in this study are calculated on the frequency of occurrence information in various topics which were recorded in a naturalistic setting. Word counts are up to date and are not from movie subtitles. 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).

Based on the measurement in Vitevitch and Luce's study (2004) where Bigram/Gram or Biphone/Phone frequencies are computed by dividing the sum of the log frequencies of all of the words with element A at position N (or N & N+1 in the case of Bigrams/Biphones) by the sum of the log frequencies of all words with any unit in position N (or N & N+1). Samples of types and token frequencies of a syllable structure, CGVN, in Mandarin are shown in Table 1.

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 1: Samples of CGVN in IPA and token frequencies

Table 1 lists the spoken dataset in CGVN with all possible sound sequences of token frequencies in Mandarin (Note that tone is omitted in the study).

The following (1-2) shows the formula for the bigram/biphone phonotactic probability in Mandarin with an example of the syllable type CGVN [ɕjaŋ] (value at 0.0821).

$$PhonProb_{tokens} = \left[\frac{\text{Sum of log frequencies of words with } [\text{ɕj}] \text{ in initial biphone position}}{\text{Sum of log frequencies of words with any biphone in initial biphone position}} + \frac{\text{Sum of log frequencies of words with } [\text{ja}] \text{ in initial biphone position}}{\text{Sum of log freqecies of words with any biphone in initial biphone position}} + \frac{\text{Sum of log frequencies of words with } [\text{aŋ}] \text{ in initial biphone position}}{\text{Sum of log freqecies of words with any biphone in initial biphone position}}\right] / [\text{Number of biphone positions}] \quad (1)$$

$$PhonProb_{tokens} = \left[\frac{\log(18031)+\log(14461)+\log(641)}{\log(18301)+\log(14461)+\log(11453)+\log(10427)+\cdots+\log(1)} + \frac{\log(18031)+\log(5760)+\log(118)}{\log(18301)+\log(14461)+\log(11453)+\log(10427)+\cdots+\log(1)} + \frac{\log(18031)+\log(5760)+\log(1226)+\log(724)+\log(361)+\log(118)}{\log(18301)+\log(14461)+\log(11453)+\log(10427)+\cdots+\log(1)}\right] / 3 = 0.0821 \quad (2)$$

In this model, the phonotactic probability is calculated in a given syllable using the token frequencies and a dataset of word types involving different syllable types. Initially, the given syllable is segmented into a series of bigrams, representing 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 one for the logarithm of the frequency of all bigrams at the 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 indicates the relative frequency of each bigram in its position.

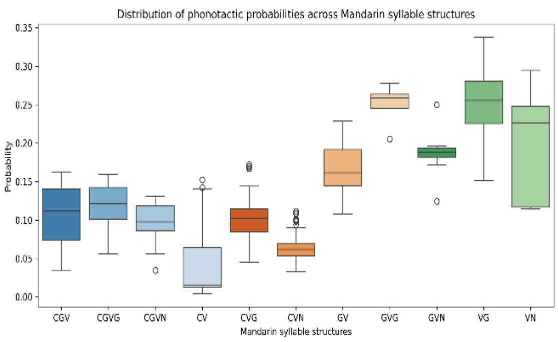

Figure 1: Distribution of phonotactic probabilities across Mandarin syllable structures.

In Figure 1, the picture visualizes the distribution of phonotactic probabilities across different syllable structure types, each depicted by a box plot. Each box shows the interquartile range (IQR) of probabilities, with the median marked by a line across the box. The 'whiskers' extend to the furthest points within 1.5 times the IQR, highlighting the range of typical data, while circles denote outliers, representing syllable structures with probabilities outside the typical range. The VN structure has the widest range of phonotactic probabilities, suggesting a high variability in how often it occurs in Mandarin, and it has a higher median probability, indicating the phonotactic constraints in the vowel-nasal sequences in Mandarin are less restricted. The VG and GVG structures have the highest median probability, suggesting that the vowel and the glide sequences can allow a wider range of varieties in Mandarin. On the other hand, CV has the lowest median probabilities, indicating that there is a restricted constraint on the consonant-vowel sequences in Mandarin. The presence of outliers in several structures implies that there are some specific sequences within those structures that are significantly less or more restricted than others.

3 Results and limitations

In this section, 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.

It is interesting to note that among the entire syllable structure types, CV used to be categorized as the easiest type for children to articulate cross-linguistically in acquisition studies. The study suggests that phonotactic constraints in Mandarin disassociate articulatory complexity and phonotactic probabilities influence speech production regardless of the markedness complexity. Our spoken samples via data computation confirm an emerging agreement within the field that phonological theories need to consider phonotactic probabilities. The limitation for the current study is that Levenshtein edit distance will need to be measured in order to further calculate the neighborhood density. The future step will need to investigate the neighborhood density in Mandarin, where the sound-similar words are stored in the mental lexicon.

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).

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

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. 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. https://drive.google.com/file/d/11giA-12_AlAAH5kkyRPOwK0jra4AZmev/view

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.

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

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. 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 SIGHAN*, pages 168-171. https://aclanthology.org/W03-1726/

A Supplementary Material

Supplementary material on the spoken data and coding will be released to the public once the paper is accepted for publication.