Appendix

Phonotactic probability

Vitevitch and Luce (2004) define phonotactic probability positionally, giving options for both uniphone- and diphone-based measures. The lexical frequency information comes from Kučera and Francis (1967), while drawing the phonetic transcriptions and words from an online version of the Merriam-Webster Pocket Dictionary. For the uniphone positional measure, and for each particular phone and position pairing possible in the lexicon, the sum of the logged frequency counts of all the words containing that pairing was divided by the sum of the logged frequency counts of all words with that position available for a phone. Formally, this is given in Equation 1:

$$p(s,i) = \frac{\sum\limits_{\sigma \in S_i} \log_{10} f_\sigma}{\sum\limits_{\omega \in W_i} \log_{10} f_\omega}\,, \tag{1}$$

where $p(s,i)$ is the probability of segment s at position i, S_i is the set of all words containing s at position i and W_i is the set of all words containing any segment at position i (that is, all words that have at least i segments). For a given item, then, its phonotactic probability is taken as the sum of this calculation for each of its segments, accounting for their positions. The function for calculating the phonotactic probability of a word, $pp(w)$ is given formally in Equation 2:

$$pp(w) = \sum_{(s,i) \in w} p(s,i)\,, \tag{2}$$

where (s,i) is a pair containing a segment s of word w and the position i where s occurs in w. The diphone version was calculated analogously, only they used diphones instead of single phones.

Vitevitch and Luce (2004) claimed that the use of the log function helps the measure better represent human sensitivity to log frequency effects. However, some characteristics of this definition of phonotactic probability are undesirable. Principally, taking the log transform of count data before performing the division operation, as in Equation 1, makes it more difficult to interpret the output in a well-defined manner. This is perhaps easier to see when expressed in an equivalent manner, as in Equation 3, which is clearly not recognisable as a traditional probability value. It may well be that the result is a good predictor of participant behaviour, but it can't reasonably be conceived of as a probability value to represent phonotactic probability.

$$p(s,i) = \log_{10}\left(\left(\prod_{\sigma \in S_i} f_\sigma\right)^{\frac{1}{\prod_{\omega \in W_i} f_\omega}}\right) \tag{3}$$

There are other methods that can be used to account for the concern that logged values better reflect human perception. One example is to calculate the probability based on count data first and then log that probability value. This is what is known as "log probability", and it can be easily mapped back to a standard probability value between 0 and 1. It is also more transparent in terms of what it represents about the count data.

An additional concern is that words that have a frequency count of 1 will not come to affect the probability values for any sequence because $\log(1) = 0$. Vitevitch and Luce (2004) also do not state how they account for items with a frequency of 0, for which the log function is undefined. Finally, while 0 is the lower bound for their method of calculating phonotactic probability for a word, there is no theoretical upper bound. As such, a word or pseudoword could be assigned a phonotactic probability value greater than 1, violating the definition of classical probability. Consider counting the beginning of a word as a phone as a simple example demonstrating this property. For example, the word *cat* would be represented as /#kæt/ with the "#" representing the beginning or onset of a word. All words would have such a beginning symbol, so the numerator and denominator in the fraction defining phonotactic probability are equal when considering /#/ in the first position, thus its calculated value of occurring at the first position is 1. Having even one phone following the /#/ with a non-zero value will yield an overall value greater than 1. Whether in practice such values

are often observed remains to be seen. But, it is nevertheless difficult to argue that these values can be interpreted as a proper probability if it is even possible for values greater than 1 to be obtained.

Bailey and Hahn (2001) used a similar approach to Vitevitch and Luce (2004) without the log functions to calculate transitional probabilities for diphones and triphones, though the source of the frequency counts for this metric does not seem to be mentioned in their paper. For the composite word scores, they took the geometric mean of the conditional probability scores of the segments that make up the word to calculate the score for the word. By calculating the geometric mean instead of the arithmetic mean, a true probability value is initially calculated, but it becomes less clear what the value is once it is raised to the power of $\frac{1}{n}$ to finish calculating the geometric mean. Janse and Newman (2013) used a similar method involving a mean, though CELEX (Baayen et al., 1995) was used for frequency counts. This manner of calculating phonotactic probability will converge toward a value for the word or item in question, though it is unclear what this value would be or represent. Adding subsequent segments would not necessarily drive the probability of the sequence down, which does not match the intuition that a word or pseudoword consisting of, say, 500 segments is improbable.

For the purposes of this study, we operationalised phonotactic probability as the probability that a particular sequence of diphones would co-occur, based on the relative frequency counts of each diphone in the language. We made this decision based on the results from Bailey and Hahn (2001), in that diphones seemed the least complex unit to achieve the greatest predictive power. They also stated that diphone treatments of phonotactic probability are the most common. As well, Pierrehumbert (2003) claimed that triphones are difficult to learn in comparison to diphones, so diphones seem the best choice for predictive power and closeness to speaker knowledge. The idea of using a co-occurrence probability, which is calculated with a product like Coleman and Pierrehumbert (1997) do, is not new. Yet, because previous and popular methods of calculating and defining phonotactic probability have not done this, we believe it worth being explicit about this choice.

We used the same augmented CMU Pronouncing Dictionary version 0.6 (Weide, 2005) used by Tucker et al. (2019), as well as COCA in our calculations. We began by finding the overall frequency-based probability of occurrence for each diphone found in the CMU Pronouncing dictionary. The frequency of each diphone was calculated by taking each occurrence of it in the CMU Pronouncing

Dictionary, multiplying it by the frequency of the word in COCA, and adding the resulting product to the tally of occurrences of the diphone in question. This results in a token frequency instead of a type frequency. Richtsmeier (2011) suggested using type frequencies instead, but they correlated at a value of approximately 0.99, so we don't believe there would be much of a difference. Word onset and word offset were considered phones, such that *cat*, for example, would be processed as onset+k, kæ, æt, t+offset. Words that did not occur in both datasets were dropped. The probability of occurrence of each diphone was calculated as the diphone's frequency divided by the total count of all diphones observed. This process is given formally in Equation 4

$$p(s) = \frac{\sum_{\sigma \in S} f_\sigma}{F}\,, \tag{4}$$

where S is the set of all words containing the diphone s, f_σ is the frequency of a word σ containing diphone s, and F is the number of diphones in a word times the word's frequency, summed over all words occurring in both the CMU Pronouncing Dictionary and COCA. Effectively, F is the total number of diphones observed in the subset of COCA words that have pronunciations in the CMU Pronouncing Dictionary.

We then took the phonotactic probability of a pseudoword to be the product of the probabilities of occurrence of each diphone in the pseudoword, which is what Vitevitch and Luce (2004) and Bailey and Hahn (2001) refer to as the co-occurrence probability. Formally, our function for calculating the phonotactic probability of a word $pp(w)$ is given in Equation 5:

$$pp(w) = \prod_{s \in w} p(s)\,. \tag{5}$$

Note that defining phonotactic probability as a product instead of a mean or summation of pseudo-probabilities has a few important properties. The first among them is that it concentrates the information revealed about the phonotactic probability of a sequence at the beginning. While multiplication is often commutative and associative, there is a natural given order in which to carry out the operations here, that being the order in which the diphones occur in the pseudoword. And, the rate at which the probability converges toward 0 will slow down as later and later terms are encountered. Analogously, the first few segments in a word or pseudoword are likely where the most discriminative information would be contained. This is due to the fact that the number

of good possible matches for what's being heard decreases quickly at the start of the sequence and slowly at the end of the sequence, with the largest decreases happening upon hearing the first few phones. Second, the probability converges asymptotically toward 0 for sufficiently long sequences of segments, matching the linguistic intuition that a sequence of, for example, 500 phones is an improbable occurrence for a word in a language. The implementation of this method of calculating phonotactic probability used in this study is available in the `Phonetics.jl` package (Kelley, 2020) for the `Julia` programming language (Bezanson et al., 2017).

References

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database (cd-rom).

Bailey, T. M. and Hahn, U. (2001). Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language*, 44(4):568–591.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98.

Coleman, J. and Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Janse, E. and Newman, R. S. (2013). Identifying nonwords: Effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech*, 56(4):421–441.

Kelley, M. C. (2020). Phonetics.jl.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In Bod, R., Hay, J., and Jannedy, S., editors, *Probabilistic Linguistics*, pages 177–228. MIT Press.

Richtsmeier, P. T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, 2(1):157–183. Publisher: De Gruyter Mouton Section: Laboratory Phonology.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., and Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, 51(3):1187–1204.

Vitevitch, M. S. and Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

Weide, R. (2005). The Carnegie Mellon pronouncing dictionary [cmudict. 0.6]. *Pittsburgh, PA: Carnegie Mellon University.*