Stem-controlled Vowel Harmony Can Emerge By Averaging Over Acoustic Exemplars

We introduce *Mnemorphon*, an exemplar-based model of morphophonological production, operating over word-sized tokens of human speech. We demonstrate how such a model can provide a novel explanatory framework for a well known and much studied set of morphophonological alternations, in addition to providing evidence for the feasibility of whole-word approaches to phenomena that are typically analyzed at the morphemic, segmental, or subsegmental levels. Finally, we address connections between explicit exemplar models and neural network *aka* connectionist accounts of human linguistic capacities and behaviour (cf. various commentaries in Ambridge 2020).

The core claims of exemplar-based approaches to human language are (i) that linguistic knowledge is encoded via memorization of perceptual experiences, and (ii) that generalized patterns of linguistic categorization and production emerge from analogical, similarity-driven mechanisms (Johnson 2007, Kaplan 2017). Computational implementations of exemplar models that operate on human speech require that the computation of similarity and analogy be robust to temporal variation, given that tokens of words show nontrivial variability in duration (Baker & Bradlow 2009, Bürki 2018). Previous accounts have tended to obviate this requirement by abstracting away from temporal variation, either by operating on point data (Pierrehumbert 2001) or else on sequential data that are discretized and restricted to fixed length (Mailhot 2010). In contrast, Mnemorphon's inputs and outputs are whole-word spectrograms, high-dimensional representations of actual human speech incorporating rich acoustic information, along with the durational variation characteristic of human production.

The phenomena under consideration are the patterns of labial and backness harmony in Turkish inflection, specifically plural and case marking. Vowel harmony has historically been described and explained using discrete and/or symbolic methods, whether rule- or constraint-based (e.g. Clements & Sezer 1982; Nevins 2010; Walker 2012). In contrast, Mnemorphon's representations have no explicit segmental or morphemic representation, demonstrating the feasibility of whole-word exemplar models of productive morphophonological knowledge (*contra* many claims in the literature, cf. Albright & Hayes 2003 and Finley 2020 *inter alia*).

Mnemorphon learns by storing "perceived" lexical items in a look-up table. These stored exemplars are (*form*, `meaning`) pairs: *forms* are spectrograms of word tokens, and `meanings` are tripartite symbolic tags (structured as `LEMMA-CASE-PLURAL`), meant to model salient aspects of Turkish inflectional morphology. To produce an output, a set of relevantly-alike stored items are collected (by grouping over `meanings`), then an output form (spectrogram) that is a summary representation of the aggregate, is output. In order to accommodate temporal variation in the output computation, Mnemorphon borrows from the literature on time series analysis and incorporates *DTW barycenter averaging*, (Petitjean *et al* 2011; DBA) an algorithm for computing a well-defined mean of a set of time series of potentially variable length. DBA in turn leverages *dynamic time warping* (Vintsyuk 1968; DTW), an alignment algorithm that provides an optimal response to the challenge of defining a

distance function between pairs of sequences that is robust to temporal variability (Kirchner *et al* 2010 use DTW in an exemplar model to model static lexical generalizations, rather than alternations).

We evaluate Mnemorphon both quantitatively and qualitatively. In modeling vowel harmony, success can be defined as outputting the "correct" vowel(s) in a given context, e.g. outputting a front vowel in a novel stem-affix combination if the stem contains front vowels. Following standard practice in computational linguistics, we hold out a test set of forms that Mnemorphon is tasked with generating. For the quantitative evaluation, we train a convolutional neural network to output a most likely vowel category when given a short sequence of spectrogram frames. Although this is a challenging problem, and state of the art accuracy is only around 85% (CITE), we show that Mnemorphon reliably outputs correct forms, effectively passing a sort of acoustic wug-test. Error analysis suggests that erroneous outputs can typically be attributed to a lack of representative data. Qualitatively, we make use of recent advances in speech synthesis and use a neural vocoder to convert Mnemorphon's output spectrograms into raw audio. Impressionistically, generated samples resemble held-out output acoustic forms.

We share code for running Mnemorphon, along with samples of generated speech.

Clements, G. & E. Sezer (1982) Vowel and Consonant Disharmony in Turkish. In H. van der Hulst and N. Smith (eds.) *The structure of Phonological Representations (Part II)*. Dordrecht, Foris. **Salor, O. *et al* (2006)** Middle east technical university turkish microphone speech v1.0 *ldc2006s33*. **Kaplan, A. (2017)** Exemplar-based models in linguistics. In *Oxford Bibliographies Online*. OUP. **Kirchner R. *et al* (2010)** Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, 38. **Petitjean F. *et al* (2011)** A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3). **Pierrehumbert, J. (2001)** Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language*, 45:1–11. **Vintsyuk, T. (1968)** Speech discrimination by dynamic programming. *Cybernetics*, 4:52–57. **Walker, R. (2012)** Vowel Harmony in Optimality Theory. *Language and Linguistics Compass*, 6. **Nevins, A. (2010)** *Locality in vowel harmony*. Cambridge, MA. MIT Press. **Johnson, K. (2007)** Decisions and Mechanisms in Exemplar-based Phonology. In Solé, M.J., Beddor, P. & Ohala, M. (eds) *Experimental Approaches to Phonology: In Honor of John Ohala.* OUP. **Ambridge, B. (2020)** Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40 (5-6). **Albright, A. & B. Hayes (2003)** Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90 (2). **Finley, S. (2020)** The need for abstraction in phonology: A commentary on Ambridge (2020). *First Language*, 40 (5–6). **Al-Badawy, E.A. *et al* (2022)** Vocbench: A Neural Vocoder Benchmark for Speech Synthesis. *ICASSP 2022*, Singapore. **Mailhot, F. (2010)** Instance-Based Acquisition of Vowel Harmony. In *Proceedings of the 11th Meeting of the ACL SIG on Computational Morphology and Phonology*, Uppsala, Sweden. ACL. **Baker R.E. & A. Bradlow (2009)** Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*; 52. **Bürki, A. (2018)** Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomics Bulletin & Review*, 25.