

The end of phonology, or: What kind of representation is relevant for the language user?

Background

Motivation

- Many prominent models of auditory word recognition and comprehension assume two levels of representation and processing [1-4].
- At the prelexical level the audio signal is parsed into discrete abstract units of various grain sizes, sequences of which are then used for lexical access.
- Recent work within the Discriminative Lexicon framework [5] has argued that lexical processing can be modelled using auditory representations derived directly from the speech signal rather than discrete symbolic representations [6,7].
- Their model relied, however, on comparatively heavily engineered acoustic representations.

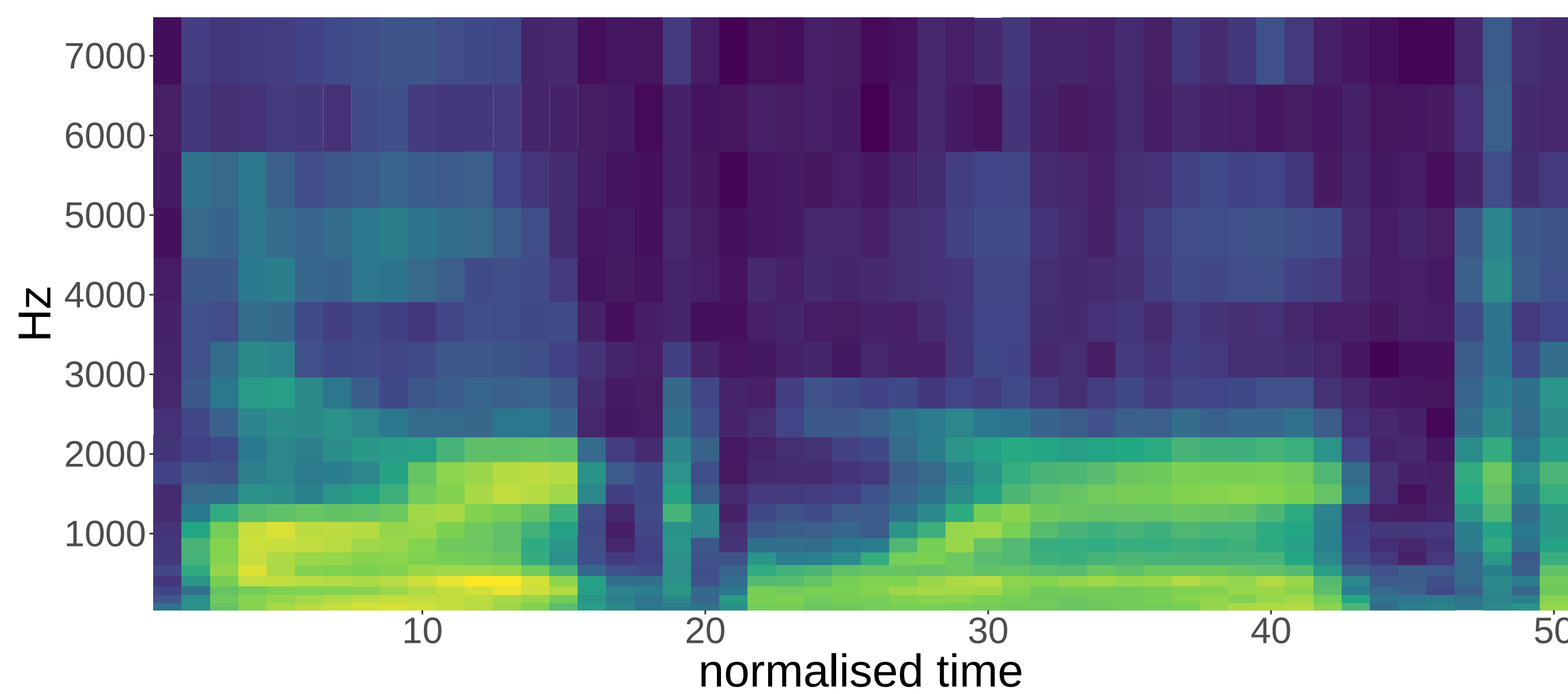


Figure 1. Time-normalised Mel-spectral representation of a token of *art work*, showing spectral energy across 21 Mel bands over 50 normalised time frames.

Aim

- We propose an alternative, and rather simple, approach, introducing Time-normalised Mel-Spectral representations (TMS).
- TMS representations preserve continuous spectrotemporal structure while requiring only minimal preprocessing and abstraction of the acoustic signal.

The Mel spectrum: spectral energy across Mel bands over time

- Frequency in Hz does not correspond linearly to perceived pitch.
- The Mel scale was designed so that equal distances on the scale correspond approximately to equal perceived pitch differences.
- Unlike a conventional spectrogram with linearly spaced frequency bins, the Mel filter bank provides finer resolution at lower frequencies and increasingly broader bands at higher frequencies.
- Each token is represented by spectral energy across 21 Mel bands and time-normalised to 50 frames, see Figure 1.
- Each column of the TMS representation in Figure 1 is then turned into a vector that is then concatenated with the vectors for each column.
- That is, each cell in Figure 2 represents one cell in Figure 1, resulting in a vector of length 1051.



Figure 2. Vectorised TMS representation of the same token of *art work*, obtained by flattening the 21 × 50 Mel-spectral matrix into a fixed-length feature vector.

Phonology vs. the signal

Test setup

- Data:** 971 noun-noun compound tokens (150 compound types) from the Boston University Radio Speech Corpus [8], at least 3 tokens per type from at least 2 different speakers
- Forms:** TMS vectors (token-level and centroids), phonological trigrams
- Semantics:** context-dependent embeddings derived from BERT [9]
- Modelling:** comprehension using linear discriminative learning [5], see Figure 3
- Accuracy testing regimes**
 - Leave-1-out cross-validation
 - Leave-150-out cross-validation (= 10-fold cross-validation)
 - The nearest neighbour of a predicted vector must be its observed vector

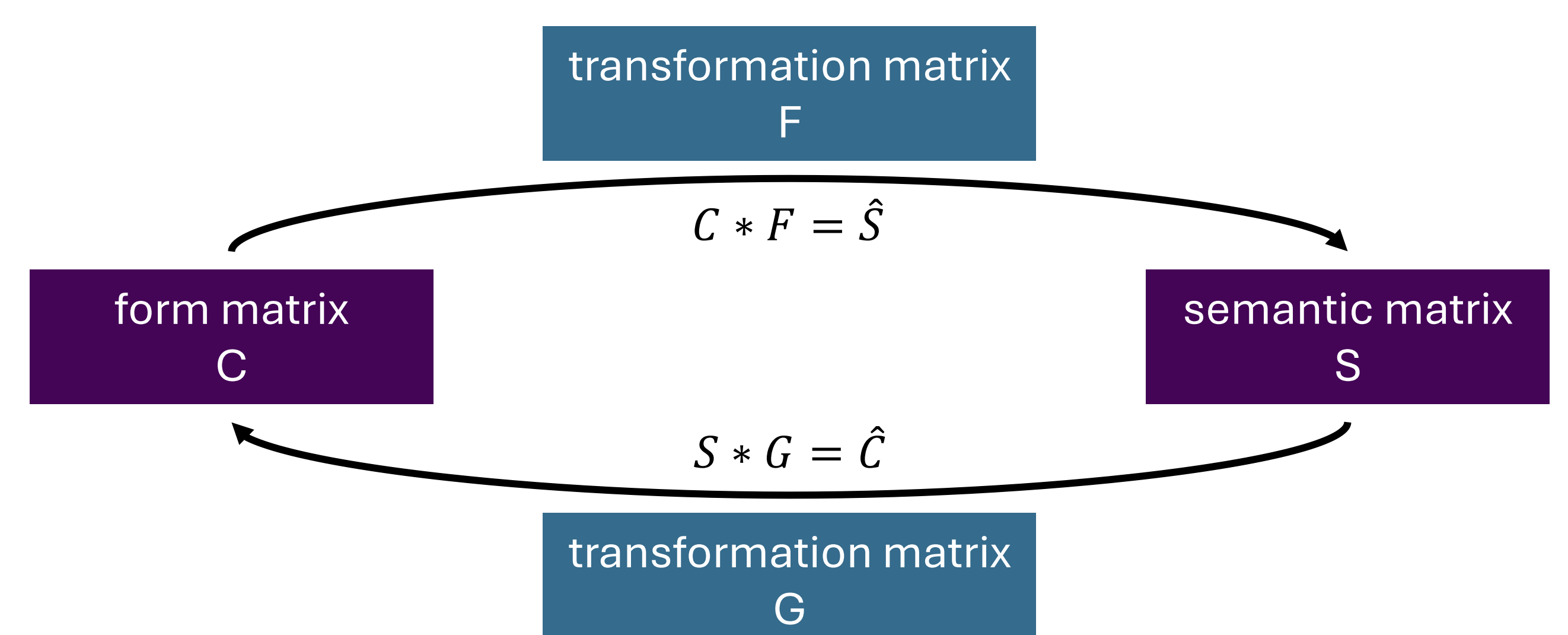


Figure 3. Illustration of mapping between C and S via F (i.e. comprehension), and S and C via G (i.e. production).

Accuracy

	Leave-1-out cross-validation	Leave-150-out cross-validation	Strict nearest neighbour
Trigrams	0.976	0.187	0.173
TMS	0.984	0.187	0.173
Baseline	0.018	0.007	0.010

- TMS vectors perform **as well as** phonological trigrams in predicting semantic embeddings in all three testing regimes.

Discussion

- Continuous spectrotemporal representations can be used successfully to access meaning in the lexicon without involving abstract segmental representations involving various researcher degrees of freedom.
- Richly structured subphonemic detail constitutes a linguistically informative representational level for lexical processing and learning.
- Outlook:** We are currently benchmarking TMS vectors against two earlier approaches to form representations derived more directly from the speech signal.

References



- [1] Lahiri, A., & Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38(3), 245–294. [2] McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. [3] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234. [4] Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395. [5] Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, 2019, 4895891. [6] Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., & Baayen, R. H. (2023). LDL-AURIS: a computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience*, 38(4), 509–536. [7] Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, 12(4), e0174623. [8] Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1996). *Boston University Radio Speech Corpus*. Linguistic Data Consortium. [9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, 4171–4186.