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

Dominic Schmitz & Ingo Plag (Heinrich-Heine-Universität Düsseldorf)
{dominic.schmitz, ingo.plag}@uni-duesseldorf.de

Extant theories of auditory word recognition and comprehension assume two levels of representation and processing. 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 (Shafaei-Bajestan et al. 2023) has argued that lexical processing can be modelled using auditory representations derived directly from the speech signal rather than discrete symbolic representations. Their model (LDL-AURIS) relied, however, on comparatively heavily engineered acoustic representations.

In this paper 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. Figure 1 illustrates a time-normalised Mel spectrum, which is then flattend into a single vector ('TMS speech vector', or 'speech vector' for short) that encapsulates the spectral properties of the compound token in question.

In our study we test the performance of TMS speech vectors against the performance of phoneme-based representations in simple two-layer networks, in which formal representations are mapped onto context-dependent representations of meaning, i.e. contextual embeddings. As a test bed we use 971 noun-noun compound tokens (representing 150 compound types) from the Boston University Radio Speech Corpus (Ostendorf et al. 1996), and linear discriminative learning networks (Baayen et al. 2019) for modeling comprehension. For formal representations we used TMS speech vectors that were directly derived from the speech signal on the one hand, and phonological trigrams on the other. For the semantics, we used context-dependent semantic embeddings derived from BERT (Devlin et al. 2018). For the modelling of comprehension, semantic vectors were predicted from the formal representations. Accuracy was evaluated using different regimes for measuring successes and failures in finding the correct semantics (e.g. held-out compound tokens and nearest-neighbour classification). To keep the variability within types comparable between triphone-based and TMS-based representations, type-level TMS centroids were used.

The phoneme-based models and the TMS-based models perform equally well in the different accuracy-measuring regimes. When mapping triphones and TMS-centroids onto token-level semantics and employing a very strict accuracy regime (i.e. finding the nearest neighbour token of the predicted vector), both models reach a comprehension accuracy of 0.173 (baseline: 0.01). With type-level semantics, the triphone-based model

reaches an accuracy of 0.976 and the TMS-based model reaches an accuracy of 0.984. In regimes with held-out tokens (cross-validation and leave-one-out), both triphone- and TMS-based models reach accuracies of 1. This is expected for triphone models, but perhaps unexpected for the speech vectors.

The results for the TMS speech vectors suggest a close relationship between acoustic realisation and semantics in compound comprehension. The TMS-based models demonstrate that continuous spectrotemporal representations extracted from the acoustic signal can be used succesfully to access meaning in the lexicon without involving abstract segmental representations, which, on top of general theoretical problems, introduce various researcher degrees of freedom. The present study demonstrates that richly structured subphonemic detail in the speech signal itself constitutes a linguistically informative representational level for lexical processing and learning.

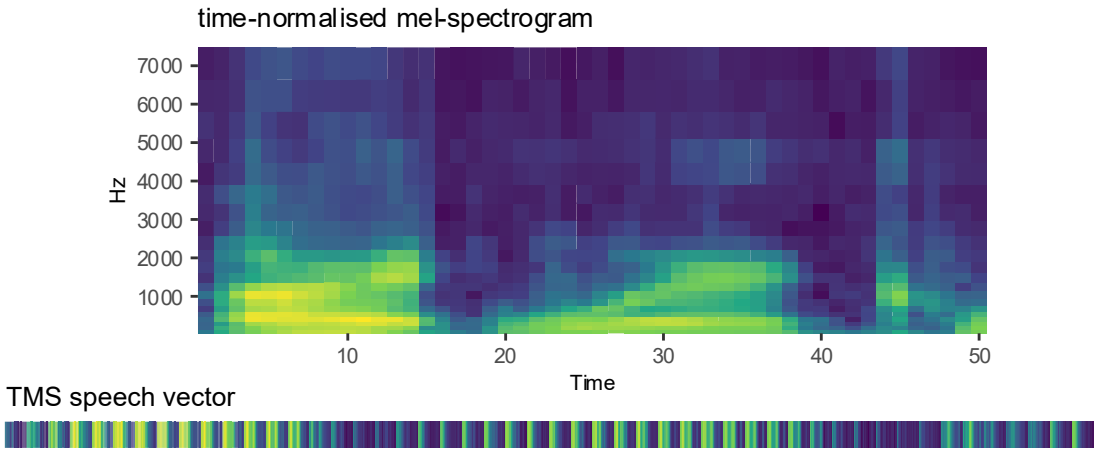

Figure 1: Time-normalised Mel Spectrum of a token of *art work* and the corresponding TMS speech vector.

References

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 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.

Devlin, Jacob, Ming Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

Ostendorf, Mari, Patti Price & Stefanie Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.

Shafaei-Bajestan, Elnaz, Masoumeh Moradipour-Tari, Peter Uhrig & R. Harald Baayen. 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.