The realization of tones in spontaneous spoken Taiwan Mandarin: a corpus-based survey and theory-driven computational modeling

Yuxin Lu[1], Yu-Ying Chuang[2] & R.Harald Baayen[3]
[13]University of Tübingen
[2]National Taiwan Normal University

A growing body of literature has demonstrated that fine phonetic details in how words are actually spoken can reflect subtle differences in meaning. These studies point to a complex and as yet understudied interaction between semantics and phonetic realization, particularly in the tonal realization in Mandarin.

The current study investigated the tonal realization of disyllabic words with 20 bi-tonal combinations using a corpus of spoken Taiwan Mandarin. Two hypotheses guided our research. First, the meanings of words co-determine the phonetic details of how the tones of these words are produced. Second, the pitch contours of word tokens can be predicted from their token-specific meaning vectors with above-chance accuracy using computational modeling.

To test Hypothesis 1, we made use of Generalized Additive Mixed Models (GAMMs) to model pitch contours as a function of predictors, including normalized time, tone pattern, gender, neighboring tones, speech rate, word position, bigram probability, speaker, word, and last but not least, words' meaning. The results show that word and words' meaning emerged as the most crucial predictors of f0 contours, with a greater effect than tone pattern. The strong of effect of words' meaning further suggests that the word-specific tonal realization may be semantic in nature.

Furthermore, to test Hypothesis 2, we used Discriminative Lexicon Model (DLM) to map the context-specific meaning onto pitch contours. DLM seeks to model the relation between form and meaning, both represented by nuermic vectors in the simplest possible set-up. The semantics vector was represented by 768-dimensional Contextualized Embeddings obtained from the GPT-2 large language model. The mean prediction accuracy was 12.3% for the training dataset and 7.7% for the testing dataset. The DLM-predicted tone pattern contours showed remarkable similarity with the GAMM-isolated contours of tone patterns. Thus, the pitch contours of word tokens can be predicted to a considerable extent from these contextualized embeddings, which approximate token-specific meanings in contexts of use.

The results obtained in the present study have several theoretical implications. First, words' pitch contours can be predicted with above chane accuracy from their meanings in context. This finding indicates that the mapping from context-sensitive meaning to pitch contours is machine-learnable. The finding that just a simple linear mapping is all that is needed suggests that human speakers should also be able to learn this simple mapping between meaning and form. Second, our findings challenge the axioms of the arbitrariness of the sign and the dual articulation of language. If the relation between form and meaning would be truly and fundamentally arbitrary, this would imply learning words and their meanings is extremely difficult, and would not allow any generalization. However, our simple linear mapping falsifies the axiom that the relation between form and meaning (here, pitch and meaning) is completely arbitrary. In summary, meaning in context and phonetic realization are far more entangled than standard linguistic theory predicts.