I let the music speak: a model of music perception that predicts speech segmentation

Geraint Wiggins
Goldsmiths, University of London

Abstract: To study the relationship between language and music, we apply a successful model of music perception to segmentation of phoneme streams into syllables.

Our model is a complex mixed-order multiple-feature n-gram model, with advanced back-off and smoothing capabilities. It has a long-term component, learned by unsupervised mere exposure, and a short-term component, exposed to the current stimulus; entropic weighting biases predictions between components. It was invented to simulate implicit learning of melodic pitch expectation, but it also predicts melodic segmentation, subjective expectation strength, associated neurophysiological activity, and aspects of expert musicologists' judgements. It is unusual as a Markov model in being multidimensional: it is capable of modeling sequences of objects with multiple features, using those features independently or together, and combining resulting mulitple predictions in a principled way.

Here, we model phoneme/stress sequences from the TIMIT speech resource metadata, using 2,342 phoneme sequences from US English, containing 21,427 syllables and 82,611 separate phoneme occurrences. We predict syllable boundaries by rise-picking in the resulting sequence of information-content values. The model predicts given segmentation with kappa=0.48, precision is .71, recall is .63, F1=.67, correct, using phoneme and stress only, over this surprisingly small learning corpus.

The results suhhest that our model may be a cross-modal model of preceptual sequence learning.