

Making Speech Synthesis Sound More Human: Introducing Variation

Eric Keller, LAIP, IMM Lettres, University of Lausanne, Switzerland
eric.keller@unil.ch

Abstract

Current speech synthesis systems show increasing technical perfection, but are still recognized as non-human when producing extended passages. This perception is a major hindrance in the general acceptance of speech synthesis technology in areas such as film and television. We propose a review of issues that contribute to the "perceived humanness" of speech. Central to this review will be the issue of regularity and variation, particularly with respect to an utterance's temporal structure. We have new evidence to show that human speakers show considerable interspeaker agreement with respect to the placement of strong vowel onsets, but that interspeaker agreement is much less for weak vowel onsets. Strength or weakness is determined automatically from the acoustic signal, and is thus likely to correlate with both motoric and perceptual saliency within the utterance. This suggests that current statistical or neural network models for temporal structuring of speech may well be fundamentally flawed. Instead of a rigid, totally predictable structure, temporal prediction systems should probably provide a set of main "temporal anchor points" within the utterance, and introduce "motivated variation" for the remaining aspects of temporal structure. We will pass in review the different types of psycholinguistic and pragmatic events that can motivate such variation, and we will present a modified linear prediction system that can handle these new requirements.