Abstract

In this paper we present a new cascading conditional random field based phrase break model for text-to-speech systems, trained on the speaker specific acoustic data that the text-to-speech voices are trained on. The training phase does not require any manually labeled phrase break tags, as these are derived directly from the speaker specific recordings used for building the synthetic voices. We present objective evaluations on various corpora, and show that the proposed model compares well with state-of-the-art data-driven phrase break models, with the added benefit of being in a unified framework.