CCIT REPORT #754 January 2010

Statistical Text-To-Speech Synthesis based on Segment-wise Representation with a Norm Constraint

Stas Tiomkin[†‡], David Malah[†], and Slava Shechtman[‡]

[†] Department of Electrical Engineering, Technion-I.I.T, Israel Institute of Technology, Haifa 32000, Israel.
[‡] Speech Technologies group, Haifa Research Lab, IBM

Email: {stast@tx, malah@ee}.technion.ac.il, slava@il.ibm.com

Abstract

In statistical HMM-based TTS systems (STTS), speech feature dynamics is modelled by first- and second-order feature frame differences, which, typically, do not satisfactorily represent frame to frame feature dynamics present in natural speech. The reduced dynamics results in over-smoothing of speech features, often sounding as muffled and buzzy synthesized speech. In this work we propose a method to enhance a baseline STTS system by introducing a segment-wise model representation with a norm constraint. The segment-wise representation provides additional degrees of freedom in speech feature determination. We exploit these degrees of freedom for increasing the speech feature vector norm to match a norm constraint. As a result, statistically generated speech features are not over-smoothed, resulting in more natural sounding speech, as judged by listening tests. The proposed method consumes less real-time memory during synthesis, and the applied iterative algorithm has faster convergence than a Global-Variance (GV) approach, reported earlier, with comparable quality.

Index Terms

Text to speech (TTS) synthesis, statistical TTS, speech feature dynamics, segment-wise model representation.