CCIT Report #777 November 2010

A Hybrid Text-to-Speech System that Combines Concatenative and Statistical Synthesis Units

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

Email: stasti@gmail.com, malah@ee.technion.ac.il, {slava, zvi}@il.ibm.com

Abstract

Concatenative synthesis and statistical synthesis are the two main approaches to text-to-speech (TTS) synthesis. Concatenative TTS (CTTS) stores natural speech features segments, selected from a recorded speech database. Consequently, CTTS systems enable speech synthesis with natural quality. However, as the footprint of the stored data is reduced, desired segments are not always available in the stored data, and audible discontinuities may result. On the other hand, statistical TTS (STTS) systems, in spite of having a smaller footprint than CTTS, synthesize speech that is free of such discontinuities. Yet, in general, STTS produces lower quality speech than CTTS, in terms of naturalness, as it is often sounding muffled. The muffling effect is due to over-smoothing of model-generated speech features.

In order to gain from the advantages of each of the two approaches, we propose in this work to combine CTTS and STTS into a hybrid TTS (HTTS) system. Each utterance representation in HTTS is constructed from natural segments and model generated segments in an interweaved fashion via a hybrid dynamic path algorithm. Reported listening tests demonstrate the validity of the proposed approach.

Index Terms

TTS synthesis, concatenative TTS, statistical TTS, hybrid TTS, dynamic path.

† Department of Electrical Engineering, Technion-I.I.T, Israel Institute of Technology, Haifa 32000, Israel.

‡ Speech Technologies group, IBM Research Laboratory in Haifa University Campus, Mount Carmel Haifa, Israel 31905.