

Abstract

A developer wanting to create a speech synthesizer in a new voice for an under-resourced language faces hard problems. These include difficult decisions in defining a phoneme set and a laborious process of accumulating a pronunciation lexicon. Previously this has been handled through involvement of a language technologies expert. By definition, experts are in short supply.

The goal of this thesis is to lower barriers facing a non-technical user in building “TTS from Zero.” Our approach focuses on simplifying the lexicon building task by having the user listen to and select from a list of pronunciation alternatives. The candidate pronunciations are predicted by grapheme-to-phoneme (G2P) rules that are learned incrementally as the user works through the vocabulary. Studies demonstrate success for Iraqi, Hindi, German, and Bulgarian, among others. We compare various word selection strategies that the active learner uses to acquire maximally predictive rules.

Incremental G2P learning enables iterative voice building. Beginning with 20 minutes of recordings, a bootstrapped synthesizer provides pronunciation examples for lexical review, which is fed into the next round of training with more recordings to create a larger, better voice... and so on. Voice quality is measured through transcription on heldout sentences, AB listening tests, and through mel cepstral distortion (MCD). We have discovered a log-linear law relating corpus size to mel cepstral distortion, and measured the gain attributed to a better lexicon. Data also supports a log-linear relation between MCD and AB listening tests, thereby grounding the most commonly used objective measure to the easiest form of subjective evaluation.

Finally, we introduce a novel approach to inferring a lexicon directly from acoustic samples recorded by the user. Our algorithm combines evidence provided by an all-phone decoder with synthesizer output to discover accurate as-spoken surface pronunciations.