Ksenia Shalonova

OutsideEcho, UK
ksenia@outsideecho.com

TTS Evaluation

Keywords: *TTS, quality evaluation, intelligibility, naturalness, letter-to-sound module*

Contents

1. Introduction

This document provides practical information for evaluating the quality of the TTS system on different levels.

2. Letter-to-sound module

2.1. Morphological decomposition

In testing the performance of the Morphological Decomposition module, the general principle can be the more data are checked manually – the better.

One possible solution can be based on dividing the word frequency list into 3 parts (frequent, in the middle, not frequent). Word frequency list can be obtained from the available text corpora. From this list it is possible to obtain the reference set (manually checked) containing 300 most frequent wordforms, 200 in the middle, 100 not frequent. The necessity of testing not the most frequent words as these words may have different morphological characteristics, be totally different in length etc.

Wrong decomposition of a wordform can be caused by the following reasons:

1. Missing root in the root dictionary
2. Missing suffix in the suffix dictionary
3. Missing rule for morpheme concatenation
4. Missing general rule

Morphological learning in future can be considered as one of semi-automated tools for Morphological decomposition evaluation.

For the languages with free stress and for the languages with grammatical tones it is worth carrying out a separate test for proper stress/tone assignment. The criteria can be the same – to obtain the reference set with 300 most frequent wordforms, 200 in the middle and 100 not frequent.

2.2. Phrasing

It is possible to subdivide the annotated speech corpora (minimum – 400 sentences) into 2 sets – one is for obtaining the phrase rules and another is for testing. On the limits of this project it will be not feasible to carry out more testing.

2.3. G2P

First of all, testing of isolated words can be carried out. Such testing can be performed on the same principle described in 2.1. (obtaining reference sets – manually checked transcriptions for the wordforms with different frequency score).

Testing connected texts can be performed in 2 ways:

1. Random set of sentences
2. Some difficult cases (e.g., when several rules are applied at one place) based on the linguist feedback.

It is important to notice that the preliminary knowledge about allophone (or phone) realisations (taken from the scientific books, articles etc.) can be verified/rejected during segmentation/annotation procedure.

3. General tests of Synthesized Speech

The purpose of the general tests is evaluating intelligibility and naturalness of the synthesized speech. By intelligibility we mean the proper phoneme realisations, whereas the naturalness comprises pitch, duration and allophone realisations.

For testing **intelligibility** the following approaches can be used [1]:

- Minimal pairs intelligibility test
- Nonsense sentences (Semantically Unpredictable Sentences test)
- Diagnostic intelligibility tests (limited)

For testing **intelligibility and naturalness** the overall quality test can be performed with the quality scale: good, bad, mediocre on the following data sets:

- isolated words (100 most frequent, 50 of the least frequent and 50 of intermediate frequency).
- compound words
- short affirmative sentences (appr. 50 random sentences)
- long affirmative sentences with phrase breaks Long affirmative sentences (appr. 50 random sentences).
- prosodically rich sentences
- interrogative sentences
- sentences containing text normalisation items
- emails
- …

Possible errors to be marked by subjects for each stimulus

For isolated words:

- Incorrect pronunciation
- Incorrect word stress
- Incorrect duration (ask to mark a particular sound with incorrect duration)
- Bad sound quality (ask to specify: metallic, bad juncture at sound boundary etc.)
- Can't understand the word
- …

For sentences:

- Incorrect pronunciation (ask to mark the word with incorrect pronunciation)
- Incorrect word stress (ask to mark the word with incorrect word stress)
- Incorrect duration (ask to mark the word/sound with incorrect duration)
- Bad sound quality (metallic, bad juncture at sound boundary etc.)
- Bad rhythm/intonation
- Lack of pause/phrase boundary
- Can't understand the sentence
- …

The major problems can be fixed in the following ways

- test the output of the Text Processing module (see 2)
- improve tuning weights in Unit selection algorithm
- determine what DB extension is required. (eg, rare diphone coverage, triphone coverage for frequent letter combinations, final fall coverage, final rize coverage, etc.).

References

[1] Multilingual Text-to-Speech Synthesis. The Bell Labs Approach. Ed. Richard Sproat. Kluwer Academic Publishers, 1998. pp. 229-244.