Large Scale Russian Hybrid Unit Selection TTS

Ilya Oparin, Vitaly Kiselev, Andre Talanov,
Speech Technology Center,
and dept. of Phonetics, St.Petersburg State University,
St.Petersburg, Russia
ilya@speechpro.com

Abstract

This paper outlines a project on the development of a new hybrid unit-selection and concatenative Russian TTS system. Project is held within Federal Research and Development Program in Priority Directions of Development of Scientific and Technological Complex of Russia in 2007-2012. A new generation Russian TTS that makes use of syntactic and semantic analysis and can be implemented in various types of electronic devices is the major aim of the project.

1 Introduction

Most of the TTS systems that were developed for Russian are diphone or allophone concatenative systems. It is well known that it is not possible to attain natural sounding of the synthesized speech using these approaches. The unit selection (US) approach currently provides the best synthesis quality. There are very few Russian US TTS systems and those are developed by international companies. Probably due to this fact these systems are generally characterized by relatively weak text analysis (that is heavily language-dependent) and are characterised by numerous mistakes in stress and break assignment, inability to handle numerals, abbreviations and special signs correctly and, finally, monotonous speech. Russian is an inflectional language with variative stress position and is full of homography. There is much room for improving the quality of the synthesized Russian speech that has been left comparatively intact.

The project is held within Federal Research and Development Program in Priority Directions of Development of Scientific and Technological Complex of Russia in 2007-2012 and financed jointly by Russian Federal Agency for Science and Innovations (http://www.fasi.gov.ru) and Speech Technology Center (http://www.speechpro.com). Project started in April 2007 and is due to finish in October 2009. Major aim of the project is to develop a new Russian TTS system that provides highly natural synthesized speech. Natural speech means both good acoustic quality (provided by the US technique), expressivity and “linguistically” correct reading. Additional push to the to the latter is given by the implementation of special speech-oriented kinds of syntactic and semantic analysis.

Another feature of the system is its scalability and state-of-the-art technological implementation. The scalability means the system should be prone to use in PCs, server platforms and mobile devices. This is attained by implementing the so-called hybrid synthesis that combines features of concatenative allophone and US synthesis. We use large US speech databases (10 hours of speech for each speaker (totally 8 speakers), segmented on different linguistic levels) that include neutral and emotional speech.

The system is designed for wide spectrum of possible applications that are concerned with the implementation of TTS system for optimization of different industrial tasks. Major features of the system may be summarized as thus:

- New solutions for Russian speech synthesis that would result in higher quality of synthesized speech when compared to existing Russian TTS systems;
- Implementation of syntactic and semantic analysis for better naturalness of speech;
- Scalability of the system that allows for its implementation for both personal computers and mobile devices such as pocket PCs and smart phones;
- Support of international standards (e.g. MRCP) that would allow for easy integration of the TTS system in other software modules;
- Large speech databases that are rich in intonation and include expressive speech;

- Wide range of different synthesized voices.

The system is designed and implemented by Speech Technology Center in collaboration with the department of Phonetics, St.Petersburg State University.

2 Outline of the System

A hybrid unit selection - concatenative TTS system is being developed. The system is scalable, so that it would attain best possible quality for different computational environments. Full version of the system includes both modules of US and concatenative synthesis for best performance. Concatenative module is bound to complement the US one in case there are allophones that are not found in the US database (or there is no good match). That is actually very unlikely in the full version, so it can be regarded as the US system. Lite version for mobile devices relies on the restricted US database and is to larger extent hybrid since it inserts allophones from the the concatenative TTS database in the US-generated signal. The whole system is designed and configured so that it supports major mobile operating systems, requires little memory space, imposes low computational load on the CPU of a mobile device and attains high speech quality for the conditions given.

Large speech databases are used for US synthesis. Each database contains about 10 hours of speech for each speaker. Two hours are segmented on different levels manually, the rest of the segmentation is performed automatically in the force aligment mode of a Russian speech recognition system developed at Speech Technology Center. The database contains reading of different texts by the speakers. Some of texts are aimed at obtaining intonation-rich and expressive speech.

2.1 Text Analysis

The following stages of text analysis is performed at the stage of text analysis to minimize linguistic errors in the synthesized speech:

- Handling of numerals. Correct grammatical characteristics should be inferred from text and assigned to digits in order to “read” them properly, i.e. substitute digits with wordforms in correct cases.
- Abbreviations. Abbreviations in Russian may be read in different ways: letter by letter or as a word. In the latter case it is also necessary to detect the correct stress position.
- Special signs. Correct reading of special signs ($, %, etc.) should bear certain grammatic characteristics that depend on the context.
- Shortenings. The same shortening can correspond to different referents. For example, in Russian the words *year*, *mountain*, *city*, and *gram* are conventionally shortened to the same letter. Thus linguistic analysis should be implemented in order to solve this ambiguity.
- Transliteration. Latin words should be transliterated to cyrillics and read properly.
- Homonymy and homography disambiguation. The same wordforms may have different grammatical characteristics and be pronounced in different ways (normally because of different stress position). Even in case homonyms are pronounced in the same way the homonymy should be disambiguated since it may affect reading of other words. Full homonymy disambiguation is possible with syntactic and semantic analysis.
- Break assignment. Russian is characterised by relatively long sentences (as compared to English). Many intra-sentence breaks should be placed, to large extent not bound to punctuation. In this case linguistic analysis is needed to place prosodic boundaries. Such analysis is hindered by the relatively free word order in Russian.
- Intonation types. In order to avoid monotonous speech and attain natural sound, intonation is modelled with a rich system of intonation types. Currently we distinguish 31 different intonation types.
- Intonation center. Usually the phrasal stress falls on the last stressed syllable. However in some cases it may shift and this shift should be predicted from text.

Automatic syntactic and semantic analysis is needed to find better solutions to a number of tasks mentioned above. Pure syntactic analysis has strong restrictions in implementation due to the inaccuracy of modern parsers, large number of concurring parse-trees for one sentence and high computational demands. At the same time context analysis was shown to be a good option to implement in TTS systems. Thus a proper balance between precision and computational costs should be found for the implementation of text analysis.