WikiSpeech – enabling open source text-to-speech for Wikipedia

John Andersson[1], *Sebastian Berlin*[1], *André Costa*[1], *Harald Berthelsen*[2], *Hanna Lindgren*[2], *Nikolaj Lindberg*[2], *Jonas Beskow*[3], *Jens Edlund*[3], *and Joakim Gustafson*[3]

[1]Wikimedia Sverige, [2] STTS, [3] KTH

{John.andersson, andre.costa, sebastian.berlin}@wikimedia.se,
{harald, hanna, nikolaj}@stts.se,
{beskow, edlund, jocke}@speech.kth.se

Abstract

We present WikiSpeech, an ambitious joint project aiming to (1) make open source text-to-speech available through Wikimedia Foundation's server architecture; (2) utilize the large and active Wikipedia user base to achieve continuously improving text-to-speech; (3) improve existing and develop new crowdsourcing methods for text-to-speech; and (4) develop new and adapt current evaluation methods so that they are well suited for the particular use case of reading Wikipedia articles out loud while at the same time capable of harnessing the huge user base made available by Wikipedia. At its inauguration, the project is backed by The Swedish Post and Telecom Authority and headed by Wikimedia Sverige, STTS and KTH, but in the long run, the project aims at broad multinational involvement. The vision of the project is freely available text-to-speech for all Wikipedia languages (currently 293). In this paper, we present the project itself and its first steps: requirements, initial architecture, and initial steps to include crowdsourcing and evaluation.

Index Terms: speech synthesis, accessibility, crowdsourcing, multilingual

1. Introduction

We introduce WikiSpeech[1], a long-term effort to create the platform and tools needed for delivering freely available Wikipedia optimized text-to-speech through Wikimedia Foundation's server architecture. At its inauguration, the project is backed by The Swedish Post and Telecom Authority (PTS) and headed by Wikimedia Sverige, Södermalms Talteknologiservice (STTS) and KTH Speech, Music and Hearing (KTH). In the long-term, the development and maintenance of speech synthesis for Wikipedia will hopefully involve parties ranging from other organizations within the Wikimedia movement (to attract volunteers and partnerships in different countries), through disability organizations and companies, to academia and the development aid sector (to enhance the service in smaller languages when there is an urgent need to reach the public with important information). In other words, the project is highly open for collaborations. The list of interested parties is growing rapidly as this paper is being written, and more formalized forms of collaboration will be put in place as the project develops.

The project will benefit both researchers and companies. For researchers, it delivers unique possibilities. Access to large amounts of data is one of the main conditions for modern speech technology and Wikispeech will generate data on all levels, from recorded narrations, through user data, to user-generated judgements, transcriptions and labels. The project not only supports, but may be at the forefront of new methods in user-centered iterative research and development, crowdsourcing and online evaluation. For companies, the material generated by volunteers to improve the text-to-speech will be free to reuse and integrate in products/services.

Particularly interesting for the accessibility aspect of speech technology is the fact that the project deals with reading of longer connected texts on a wide range of topics. Existing text-to-speech solutions are typically not designed for that kind of reading. The project also contributes to research outside text-to-speech technology. For example, the feedback from users that is generated could be regarded as a form of perception tests that can provide insights into how real listeners perceive speech of various kinds, and from the continuous updates and additions of words and pronunciations by users we can learn things that were not previously known on how language develops and how languages relate to each other. In the development version built in the project, Swedish, English and Arabic (a right-to-left language) will be included.

2. Background

2.1. Wikipedia

For more than a decade and a half now, the online encyclopaedia Wikipedia consistently ranks as one of the top-ten most visited websites in the world, with approx. 500 million visitors and 20 billion page views monthly. Wikipedia is currently available in 293 languages (making it the second language-version rich of any website, surpassed by jw.org only). The impact of providing information in people's own languages is well known, and is the driving idea behind a number of targeted Wikimedia projects aiming to deliver free high quality information across all languages about issues that are often complex (e.g. in medicine[2]).

[1] https://www.mediawiki.org/wiki/Wikispeech

[2] https://meta.wikimedia.org/wiki/Wiki_Project_Med and http://blog.wikimedia.org/2016/03/29/wikipedias-essential-vaccines/

2.2. The MediaWiki Server Environment

Wikipedia is built on the MediaWiki software, which is in turn used by many thousands of other websites. The MediaWiki software is written in PHP. WikiSpeech will be created as an *extension*, and extensions can use other programming languages. In the Wikimedia environment, there are currently some 800 MediaWiki installations, all of which can potentially utilize WikiSpeech.

Any extension that is to be enabled in production on the Wikimedia servers must fulfil strict requirements: for example, all parts must be freely licensed, and all parts must support multilingual use (MediaWiki exists in 371 languages and Wikipedia on 293). Extensions must support being activated and configured separately on each wiki, and must be translatable on the Translatewiki.net platform. This means that all code must be written with this in mind and that some additional effort to install on this platform is required so that internationalization is possible.

2.3. Text-to-speech and accessibility

In many parts of the world, the main target audience for text-to-speech in terms of accessibility is differently-abled user groups, for example those with visual impairment or dyslexia. Those with a medical diagnosis affecting reading comprehension (e.g. dyslexia; visual or cognitive impairments) often have access to technological aids. This access is, however, largely a matter of sheer coincidence. In order to get the technological aids, one must be lucky enough to live in a high income country; to speak a language for which a working text-to-speech solution is available; and to get the required diagnosis. In less fortunate parts of the world, where access to Wikipedia can make a real difference in terms of equality and democracy, other audiences become equally important and considerably more plentiful. A substantial portion of the world's population is, to this day, illiterate.

In effect, people with poor reading comprehension (from unaccustomed readers to the illiterate) often have limited or no access to the commercial tools that could potentially improve their understanding. This is especially true if they do not wish to share their personal user data with one/all of the IT behemoths.

In total, approximately 25% of people find it easier to learn from spoken text than by reading, for example Dunn and Dunn [1] found that only 20-30 percent of the school-age children were auditory learners. 25% of the readers of Wikipedia means that another 115-125 million people could benefit from the project in the long run, not counting the thousands of organizations that also use MediaWiki[1]. Taken together, the number of people that would benefit from built in text-to-speech on Wikipedia and other MediaWiki installations is very large, by any standards.

2.4. Wikipedia texts

Wikipedia, and wikis in general, contain large quantities of specialised text, which for example frequently includes names and words from different languages (e.g. the article in Swedish about Paris has plenty of French place names in it). This places hard requirements on text-to-speech, such as the need for an unusually extensive pronunciation lexicon is required. The average sentence length is relatively high, with a large proportion of numbers, names and technical terms, and the texts often describe complex relations and processes. Readers generally read or peruse paragraphs or entire articles, rather than singular utterances. The speaking style, then, should be narrative. The most helpful interactive feature is likely pausing and resuming, which if handled properly should help readers digest more complex texts.

2.5. Multilingual support

As mentioned, Wikipedia currently exists in 293 different languages, all of which should, in the long-term, be supported by the text-to-speech system. This number is likely to grow. Commercial solutions only exist for a minority of all languages, and the incitement for developing commercial solutions for many smaller languages are weak, so for many of the languages, WikiSpeech represents one of few plausible pathways to a functional text-to-speech system. For this to become reality, the platform must be quite scalable, and flexibility is crucial.

2.6. Crowdsourcing

Wikipedia development in general, and of text-to-speech in particular, offers a unique opportunity for crowdsourcing. In addition to the sheer quantity of Wikipedia visitors, Wikipedia has a vivid community built around collaborative, shared work. At any moment in time, tens of thousands of volunteers with widely varying language expertise are involved in Wikimedia projects. This, and the fact that Wikipedia is extraordinarily well-established and well-known, provides an unparalleled opportunity for crowdsourced solutions, and we hope to engage a variety of people (e.g. people active in disability organizations, linguists, computer scientists) in the efforts. Text-to-speech users may themselves assist the development and help note and correct annoying errors, just like many readers appreciate being able to correct for example typos when reading texts.

The possibilities range from the trivial to the complex; from user generated pronunciations lexica to recordings of specialised texts and more. We hope that this will provide a basis for refined and high-quality text-to-speech even for obscure subjects in languages which previously had no working text-to-speech solutions.

3. Related work

The last decades there have been a number of efforts in developing open source speech synthesis platforms. There have also been efforts in providing speech synthesis for free for non-commercial applications, such as the MBROLA[2] project that was initiated by TCTS Lab of the Faculté Polytechnique de Mons. The aim of the project is to provide diphone speech synthesis in as many languages as possible. There is no open source code, but binaries for several platforms, and the team built voices for free if they get recorded diphone databases [2].

One of the first large-scale open source toolkits is the unit selection synthesizer Festival[3], which was a developed in the 90s by Edinburgh University [3]. The framework has been developed continuously and today, it is maintained at CMU as part of the Festvox[4] system [4]. Festvox provides tools and scripts for building new speech synthesis voices. Flite[5] is a small footprint synthesis engine that can use voices built with Festvox [5].

[1] https://mediawiki.org/wiki/MediaWiki_Usage_Report_2015
[2] http://tcts.fpms.ac.be/synthesis/
[3] http://www.cstr.ed.ac.uk/projects/festival/
[4] http://www.festvox.org
[5] http://www.festvox.org/flite/

FreeTTS[1] was developed by Sun Microsystems in order to compare the efficiency of their Java™ programming language with C. In order to get an optimal solution they based the architecture on Festival and the synthesis engine on Flite [6]. It can use voices built with Festvox and MBROLA.

Several other systems exist, notably Espeak[2], which runs on Linux and Windows and can be used as a frontend to MBROLA voices [7] and provides support for adding new languages, and GNUspeech[3], an open source real-time articulatory synthesizer that allows for the setting up of rules in new languages [8].

One of the most widely used open source speech synthesis frameworks is the HMM-based HTS[4] system from NITECH [9]. It was first released in 2002 and is continuously updated with new and improved methods for data-driven speech synthesis. It does not have its own text-processing frontend, but can be used with Festival and Festvox.

MaryTTS[5] is an open-source, multilingual Text-to-Speech Synthesis platform written in Java, that was developed by DFKI and Saarland University [10]. It can be used to build Festvox unit selection voices or HTS HMM synthesis voices. MaryTTS has a workflow for building synthesis voices in new languages in which the first step is to download a dump of Wikipedia in the target language [11].

Speect[6] is an open source multilingual synthesis framework developed by the HLT group at the Meraka Institute in Pretoria [12]. The system is built with an object oriented design with a plugin architecture, which allows for separating the synthesis engine from the linguistic and acoustic dependencies. The system uses the HTK toolkit [13] to force align the speech corpus used for training and HTS as the backend synthesizer.

One of the latest open source initiatives is Idlak[7], which is a project that aims to build an end-to-end parametric synthesis system [14] within Kaldi, a free, open-source toolkit for speech recognition research [15].

There are also a number of initiatives to publish open source speech corpora that can be used to build synthetic voices. The CMU_Arctic[8] speech synthesis databases are a set of phonetically balanced corpora designed for unit selection synthesis [16], and the Blizzard Challenge[9] provides data for speech synthesis [17].

IPS WikiSpeech[10] provides a platform for web-based creation of speech databases for the development of spoken language technology and basic research, allowing contributors to read, record and upload speech materials that are then processed, documented and published.

Finally, many open source speech corpora feature large amounts of high quality recordings, although they have not been prepared for speech synthesis training, for example LibriVox[11] that provides free public domain audiobooks.

4. The WikiSpeech System

In addition to the basic requirements governing all MediaWiki extensions (e.g. free licence, multi-language support), the project strives for a framework that is extensible and flexible enough to make use of as much as possible of the many open source efforts within text-to-speech; that provides easy editing (a requirement that is far from trivial in the text-only case, and considerably more complex when it comes to text-to-speech); that provides as broad access as possible; and that avoids hampering regular MediaWiki functionality. Steps towards these goals include ensuring that the platform is modular, engine agnostic, and cloud-based; that the client-side is easy to use; and that special attention goes into developing and tuning crowdsourcing and evaluation methods to make use of the possibilities given by the MediaWiki environment and its users.

4.1. Architecture

The WikiSpeech text-to-speech functionality will be deployed as a cloud service running on the Wikimedia server platform. It will be an open cloud-based text-to-speech platform that may be used by anyone for any purpose at any time.

While the full details of the architecture have not been pinned down at the time of the writing of this article, a number of basic principles have been agreed upon.

Modular: The WikiSpeech architecture will be centered around a set of core modules with well-defined standardized interfaces. The motivation for this is that we want to make use of the vast quantity of resources available in this area, whilst being able to deal with the fact that these are very often not standardised. The alternative would be to adhere to an existing standard. This would come at the cost of being unable to make use of most pre-existing resources, and in turn WikiSpeech would develop at a much slower pace.

The minimal set of modules that are expected to be present in any instantiation is a *lexicon*, a *text processor* and a *waveform synthesizer*. The text processor will take text (with optional markup) and output a phonetically transcribed, richly marked-up synthesis specification. This specification is passed on to the waveform synthesizer module, which outputs a speech waveform. A *JSON*-based markup format will be used both as the input and as the interface to pass information between the modules. The markup system will support (a subset of) the tags used in SSML. The *JSON*-based format will (or may, depending on the engine used, see below) also be used internally between different components of the text processor module. The modular design serves two main goals. Firstly, it will make it easier to re-use functionality and to facilitate community supported development and system improvement. Secondly, it will allow modules to be used independently for other purposes via cloud based API:s (see below). One requirement for this to work is interoperability of phonetic transcriptions. Rather than enforcing one transcription alphabet to be used throughout the WikiSpeech system, a module can use any transcription alphabet as long as it is uniquely named and defined in terms of a mapping to the *International Phonetic Alphabet* (*IPA*)

[1] http://freetts.sourceforge.net/docs/index.php
[2] http://espeak.sourceforge.net/
[3] https://www.gnu.org/software/gnuspeech/
[4] http://hts.sp.nitech.ac.jp/
[5] http://mary.dfki.de/
[6] http:\\speect.sourceforge.net
[7] https://sourceforge.net/p/kaldi/code/HEAD/tree/sandbox/idlak/
[8] http://www.festvox.org/cmu_arctic/
[9] http://www.cstr.ed.ac.uk/projects/blizzard/
[10] https://webapp.phonetik.uni-muenchen.de/wikispeech/
[11] https://librivox.org/

standard. This will simplify incorporation of existing language resources and software modules. The system will host a transcription mapping service where modules can register new mapping tables to allow seamless interoperability.

Engine Agnostic: The second basic architectural principle is that the WikiSpeech architecture is not tied to any particular text-to-speech system. Rather, it will allow – through internal API:s and wrappers – in principle any open source text-to-speech system to be integrated. However, to benefit from the modular architecture, the integration should be done by decomposing the text-to-speech system into at least the three core modules lexicon, text processor and waveform synthesizer, and ensuring that these modules read/write the common *JSON*-based format. Integration of a text-to-speech system thus typically entails more than mapping the input text and output waveform to the corresponding API calls, although, such high-level integration may also be allowed e.g. if it is required to quickly add support for a new language. The engine-agnostic approach will ensure that the WikiSpeech platform will be able to take full advantage of developments in the open source text-to-speech community, simplify support for new languages, and offer a broad choice for users to choose voice according to preference[1].

Cloud based: We have selected a server solution because we want to achieve the highest accessibility possible. Instead of relying on readers to have their own applications installed, we ensure that everyone can benefit from the results. The alternative is to develop a client which can be installed by the reader, which we view as too much of an extra hurdle for readers. A particular issue here is that the use of installed clients is often not possible when using computers with limited administration rights, as for example in a library or an Internet café. Hence, this solution risks cutting out people who cannot be assumed to own a computer or smartphone – in other words, a sizeable proportion of a very large target group of the project. Additionally, it requires the reader to be aware of the existence of the client. Finally, a server solution means that the speech synthesis can be used by third parties.

The WikiSpeech system will run as a cloud-based service accessible through a *REST* API. This is a well-established design that works well with both browser-based and standalone clients, and it is well in line with how other Wikimedia services are implemented and accessed. The *REST* API will provide access to end-to-end text-to-speech conversion services in different languages and with different voices – this is the most common use scenario in the Wikipedia setting. It will also be possible to access individual modules via the API in order to perform tasks such as transcribing a block of text phonetically, looking up individual lexicon entries or converting between different transcription alphabets. Although the primary raison d'être for the WikiSpeech system is to perform synthesis of Wikipedia articles, the cloud service itself will not be restricted to any particular usage, rather it will be a free and open service available for anyone at any time.

Easy-to-use client: On the client side, the primary goal is ease of use. The interface must be intuitive and integrate well with the Wikipedia structure. The GUI(s) will include a number of separate pieces where the most conspicuous to the end user is the audio player. The second most important piece will be the interface with which improvements are made. Wikimedia Foundation has already developed a design library with all the components well described (what they look like and how they will be used), so design on that level is not necessary. Rather it is the choice of layout and UX that will be important.

The markup of sounds must meet the specific requirement to be easily editable for anyone who edits on Wikipedia. A few different ways to store this has been investigated, and the proposed solution is to avoid markup directly in the traditional wiki text, and instead use a technique similar to Parsoid, which is used in the new editor for MediaWiki (VisualEditor). The solution prevents (sometime substantial amounts of) markup to obstruct those editing articles.

4.2. Crowdsourcing

The unparalleled opportunity for crowdsourcing text-to-speech resources provided by Wikipedia's millions of users and tens of thousands of active contributors can be harnessed in many ways, ranging from the simple to the complex and from the proven to the experimental. Initially, we will look to methods that are as close to the well-known methods used for the Wikipedia texts, but in the long run, the opportunities for more text-to-speech specific crowdsourcing are enormous. In order to benefit maximally from the contributors, any editors and crowdsourcing must work in multiple web browsers.

Proofing, editing, improving. A first step will be to make it possible for user to report bad synthesis, either by manually selecting bad passages or by pressing certain keys while listening to the synthesis. In the simplest case the user will not have to provide the correct pronunciation, instead the report ends up in a queue where someone else can correct the problem.

The second step is to allow trusted users to correct transcriptions. Tools to make this as easy and transparent as possible will be developed, for example automatic validation of the transcription and allowing the user to listen to their transcriptions. In addition to allowing users to provide transcriptions, we will look at other possibilities to extract pronunciations from users, for example *rhymes with*, where the contributor provides a rhyming word, or *re-readings*, where the contributor record the word as it should be pronounced (such recordings can also be used in other projects, e.g. Wikidata, a free, linked database that can be read and edited by both humans and machines; and Wikitionary, a free-content multilingual dictionary available in 170 languages).

As is the case with standard text edits, methods to ensure the quality of the transcriptions are crucial. Corrections from normal users end up in a queue awaiting approval, and after approval the correction is either made global, or applied locally only (in a specific article). The approval may be done by power users or by using specific crowdsourcing methods. Methods to detect vandalism and beginner mistakes also play an important role. As a final step, those who reported an error are notified when it has been corrected. We will also look into the possibilities of allowing hearers to contribute to the other mark-up in the synthesis specification (in addition to the transcriptions). Here, we are likely to run into more problems

[1] Note that any software that is to be deployed on the Wikimedia servers will have to be code-reviewed for security and stability by the Wikimedia Foundation team.

(e.g. complex text-to-speech system dependencies). The goal is generic methods that can be used to edit specific mark-up.

The results of edits to the text-to-speech resources can (but need not) be used directly to change the speech output, but a difference from standard edits on Wikipedia is that the user supplied data may also be used to retrain models used by the system, for example a G2P engine.

Expanding, adding. In addition to contributions that correct specific problems in the synthesis, there is scope for contributions with a larger scope, that aim more at general improvements, such as the creation of new voices or training of new models. For example, contributions of readings of paragraphs or entire articles can be used in a number of ways. They can be read out as is if no synthesis exists for the language, or if it encounters problems with the article in question.

Provided that we have a forced aligner (this is mainly a target for the second step of the project), the readings can be aligned, and the results can for example be read as is (but now with navigation in place), used to create new or better voices, or used to build prosodic models and the like. Read and analysed passages can be resynthesized and replayed to the contributor, a process that validates both the original recording and the analysis [18].

Regular crowdsourcing and human computation. In addition to methods that relate to Wikipedia editing, the project aims to support more conventional crowdsourcing techniques. Tools for performing standard BLIZZARD style listening tests [17] as well as crowdsourced labelling will be tested, both for evaluation and data processing.

Taken together, we hope that this will provide a basis for refined and high-quality text-to-speech even for more obscure topics in languages which previously had no working text-to-speech solutions.

4.3. The process for adding a new language

The aim of the project is not to develop the system for three languages only, but make it possible to add synthesis in any of the 293 Wikipedia languages. This requires a process and committed Wikipedia volunteers. The addition of a new language will involve the following steps:

1. Expressed interest for the activation of a new language
   - Communication about the interest e.g. through a wiki page
2. Identification of existing TTS components
   - Identification of existing resources (i.e. text processing, lexicon, audio corpus) and analysis of coverage and maturity
3. Possible API adaptations (to be included in the "wrapper")
   - Adaptations of the component APIs for the new components
4. Development of missing or bad components
   - This is decided on a case-by-case basis.
     - Creation or improvements of bad components
     - Development of simple lexicon tools (e.g. imports)
     - Training of prosody models using existing and newly recorded material
     - Development of the lexicon with the help of gamified tools (cf. Wikidata-game[1])
5. Installation
   - Manual installation by developers
6. Local configuration
   - Manual server-side configurations by developers
   - Manual configuration by the community on the wiki (e.g. the possibility of local (re)naming of different technical messages for the specific wiki and styling)

Already in the wikispeech project this process will be tested for three languages: Swedish, English and a right-to-left language (Arabic).

For Swedish we are currently in the second step in this process. The Swedish Wikipedia is the second largest in the world with ca 3 million articles. The first part of the process is to extract the actual text from the article and tagging it with part-of-speech tags. We are currently developing open source tools for extracting the text and tagging the Swedish Wikipedia. In Sweden there is a national center, Språkbanken[2] (the Swedish Language Bank) that has the task of collecting, developing, and storing (Swedish) text corpora, and to make linguistic data extracted from the corpora available to researchers and to the public. They regularly download the Swedish Wikipedia pages and turn them into a corpus in XML format, where the articles are divided into paragraphs and sentences, and each word is tagged with part-of-speech tags, lemmatisation, sense and dependency relations. As a quick starting point we have investigated their cleaned and tagged Wikipedia corpus from the fall of 2015 that has about 60 million sentences (including headlines) and 180 million words.

There are several pronunciation dictionaries for Swedish available for Swedish both from KTH and the NST Lexical database for Swedish[3]. We have made the first dictionary for the words in the 2015 Swedish Wikipedia corpus that includes ca 3.7 million entries. When running these through KTH dictionary or the transcribed SWETWOL morphology analyser for Swedish [19] half of the words are missing. This indicates that these are mostly proper names and foreign words, that will have to be dealt with [20,21]. In some cases, the local context (i.e. the article they appear in) will give information about the origin of the unknown word. Another option is to look the word up in Wikipedias in other languages, and apply heuristic rules to deduce the most probable origin. Språkbanken have a number of open resources that could be useful for text-to-speech synthesis: i.e. the Semantic and morphological lexicon Saldo, a Swedish FrameNet and SweSaurus a Swedish WordNet. For POS tagging we have identified Stagger[4] [22] that is a Swedish part-of-speech tagger based on Collins [23] averaged perceptron, that has per-token accuracy of about 96.6 percent. Finally, we have identified an open source audio corpus that was designed for Swedish speech synthesis: NST has an open source acoustic database for Swedish speech synthesis (5000 sentences, 44 kHz).

4.4. Evaluation

The Blizzard Challenge[5] ([17]; see [24] for an overview) has become the go-to evaluation for comparing voices and voice creation methods. Blizzard uses listening tests where listeners first listen to a reading, then judge it on one or more dimensions. Building this test methodology into the WikiSpeech system would give access to WikiSpeech users for this type of evaluation. WikiSpeech will, for the most part, deliver a rather

[1] https://www.wikidata.org/wiki/Wikidata:The_Game

[2] https://spraakbanken.gu.se/

[3] https://www.nb.no/sprakbanken/repositorium#ticketsfrom?lang=en&query=alle&tokens=swedish&from=1&size=12&collection=sbr

[4] http://www.ling.su.se/english/nlp/tools/stagger/stagger-the-stockholm-tagger-1.98986

[5] http://www.synsig.org/index.php/Blizzard_Challenge

special type of speech: long, continuous readings of complex texts. This speaking style is not the most common in applications using text-to-speech, and it presents special challenges that are not present in for example a standard, transactional spoken dialogue system like an airplane booking system, such as listener fatigue. The perhaps best, large-scale match to reading Wikipedia pages out loud (in addition to services that actually read Wikipedia pages out loud, such as Pediaphon[1] and Readspeaker[2]), is synthesized audio books. Although audio books in general are still read by human voice professionals, some – for example academic literature provided to reading-impaired students – is largely produced using speech synthesis, at least in some countries (e.g. Sweden, Norway).

The 2012 Blizzard Challenge [25] included a special task for developing new evaluation methods suitable for audio book evaluation. Only one contribution was entered [26], dealing with the automatic prediction of human judgement using acoustic features. This type of instrumental evaluation of speech synthesis is highly interesting in WikiSpeech, in particular at a stage where new voices are developed for the project, and (semi-)automated quality assurance is required.

The Blizzard 2012 rally for new evaluation methods suitable for audiobook evaluation did not result in any new methods for listener evaluation. Still, the special challenges involved (e.g. listener fatigue, large portions of names, complex relations) are not captured well by current methods. As WikiSpeech has the opportunity to gather listener responses with a uniquely high ecological validity – responses from actual listeners in their actual listening environment – we will make a focussed effort to develop a suitable evaluation strategy that captures these aspects.

The WikiSpeech project have the following requirements on the evaluation methods to be used:

Web delivery. In order to minimize the effect of unknown parameters, we will run the initial experiments with these methods in a controlled laboratory environment. However, it is obviously a strong requirement that the final evaluation methods are suitable for web delivery.

Crowdsourcing. The line between labelling training data for TTS, performing experiments on for example human perception of TTS, and evaluating TTS is quite blurred. If treated properly, the judgements that are the result of an evaluation can double as training data for the next generation of the TTS. Thus, the goal is the find evaluation methods that lends themselves to doubling as crowdsourced data collections.

Task appropriateness. As mentioned, the TTS task for WikiSpeech is dissimilar to that of many other TTS systems – as an example, TTS is often used in spoken dialogue systems and for announcements (with well-defined tasks and single utterances) or for reading simple texts for entertainment (audio books). In the Wikipedia setting the texts are often long and complex and the task is to learn something.

Ecological validity. An ecologically valid evaluation is one in which the listeners are similar to the listeners that are the target group of the TTS application, and where they listen in an environment that is representative for a real-world listening situation and a text that is equally representative for what they would listen to. Ideally, their motivations to listen should be the same, or similar, as well. To ensure ecological validity, we will involve end-users and end-user organizations in the user trials.

Apart from the standard evaluation test used in Blizzard we will explore some new methods:

KTH Speech, Music and Hearing has adapter the so-called Audience Response System tests used by the entertainment industry to evaluate movies and series to a system for evaluating long, continuous stretches of streaming speech technology material, such as TTS [27]. The basic idea is that listeners listen to continuous TTS and press a single button whenever they dislike something. The clicks represent time series that are then weighed together. The resulting data shows in an efficient manner where there are problems in the TTS (but not the nature of these problems), and the overall number of clicks correlate negatively with perceived quality as well as with automatic, objective quality measures. The method has clear advantages in the WikiSpeech context in that it allows listeners to judge whole articles in continuous listening, and the judgement causes very little cognitive load (as opposed to repeatedly stopping the playback to ask complicated questions). The method maintains a high ecological validity by avoiding the need to cut up the reading in short bursts interspersed pauses for filling in judgement questionnaires, and works well in web based applications.

We will also explore reaction-time based methods. In order to get a measure that represents how easy a TTS is to comprehend (as opposed to how easy it is to perceive individual words) and how much cognitive load listening to the TTS causes over time, we will design an evaluation method in which listeners are told to go on listening until they have the answer to a question they are presented with in advance. Once they know the answer, they are told to hit a button as quickly as possible and write the answer down. They then receive a new question, and listening continues. The method uses reaction time as its main metric, which also allows us to investigate listener fatigue – a TTS that causes more fatigue will result in continuously increasing reaction times as the listener grows increasingly tired.

5. Conclusion

We have presented WikiSpeech, a newly inaugurated project in which professionals within MediaWiki/Wikipedia, the speech technology industry, and speech technology academia collaborate to establish a starting point for delivering open source, multilingual text-to-speech through Wikimedia Foundation's server architecture.

As a result, it will be possible to harness the power of the MediaWIKI's/Wikipedia's enormous user base, as well as the and the considerable skills in community development that this user base possesses, for text-to-speech development in of existing systems and for the creation of new.

In the long term, the system has the potential to reach many millions of people in need of information and education and to boost the already considerable impact of Wikipedia. In addition to increasing the accessibility of one of the most widely visited websites around, all other platforms using MediaWiki will be able to make use of the technical solutions which are developed during the project. That means is several thousand websites which can quickly and easily activate text-to-speech.

The initial project will set up working instances of text-to-speech in three languages, English and Swedish (the languages of the two largest Wikipedia collections), and Arabic (in order to test right-to-left writing).

[1] http://www.pediaphon.org/~bischoff/radiopedia/

[2] http://www.readspeaker.com/listen-to-wikipedia/

6. Acknowledgements

The Wikispeech project is backed by The Swedish Post and Telecom Authority (PTS).

7. References

[1] Dunn, R., & Dunn, K. J. (1979). ”Learning styles/teaching styles: Should they, can they be matched?” Educational Leadership, 36, 238–244.

[2] Dutoit, F. Bataille, V. Pagel, O. Pierret, and O. Van der Vreken. (1996). “The MBROLA project: Towards a set of high-quality speech synthesizers free of use for noncommercial purposes.” In Proc. ICSLP, Philadelphia, USA.

[3] Taylor, P., Black, A. W., & Caley, R. (1998). “The architecture of the Festival speech synthesis system”

[4] Black, B. and Lenzo. K. (2007). “Festvox: Building synthetic voices”, Version 2.1. http://www.festvox.org/bsv/. (accessed March 2010).

[5] Black, A. W., & Lenzo, K. A. (2001). “Flite: a small fast run-time synthesis engine” In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis.

[6] Walker, W., Lamere, P., & Kwok, P. (2002). “FreeTTS: a performance case study.”

[7] Duddington. (2010). “eSpeak text to speech Version” 1.43.12. http://espeak.sourceforge.net/.

[8] Hill. D. (2008). Gnuspeech: Articulatory Speech Synthesis. http://www.gnu.org/software/gnuspeech/. (accessed March 2010).

[9] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007). “The HMM-based speech synthesis system (HTS) version 2.0”. In SSW (pp. 294-299))

[10] Schröder, M., & Trouvain, J. (2003). “The German text-to-speech synthesis system MARY: A tool for research, development and teaching” International Journal of Speech Technology, 6(4), 365-377

[11] Pammi, S., Charfuelan, M., & Schröder, M. (2010). “Multilingual Voice Creation Toolkit for the MARY TTS Platform” In proceedings of LREC 2010.

[12] Louw, J. A., Van Niekerk, D. R., & Schlünz, G. I. (2010). “Introducing the Speect speech synthesis platform”. In Blizzard Challenge Workshop 2010.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, (2005). “The HTK Book” (for HTK Version 3.3). Cambridge University Engineering Department, 2005

[14] Aylett, M., Dall, R., Ghoshal, A., Eje Henter, G. and Merritt, T. (2014) ”A flexible front-end for HTS”. In Proc. Interspeech, pages 1283-1287, September 2014.

[15] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., & Silovsky, J. (2011). “The Kaldi speech recognition toolkit” In IEEE 2011 workshop on automatic speech recognition and understanding

[16] Kominek, J. & Black, A. (2003) “The CMU ARCTIC speech databases for speech synthesis research,” Tech. Rep. CMULTI-03-177 http://festvox.org/cmu arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[17] Black, A., & Tokuda, K. (2005). “The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases.” In Proceedings of Interspeech 2005.

[18] Gustafson, J., & Edlund, J. (2008) “Expros: a toolkit for exploratory experimentation with prosody in customized diphone voices" In Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems. Kloster Irsee, Germany.

[19] Magnuson, T., Granström, B., Carlson, R., & Karlsson, F. (1990). “Phonetic transcription of a Swedish morphological analyzer.” In Proceedings of the of Fonetik-90, Reports from the Department of Phonetics University of Umeå, Phonum. Umeå.

[20] Gustafson, J. (1995). "Transcribing names with foreign origin in the ONOMASTICA project", in proceedings of ICPhS'95 in Stockholm, August 13-19.

[21] Eklund, R., & Lindström, A. (2001). Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. Speech Communication, 35(1), 81-102.

[22] Östling, R (2013) “Stagger: an Open-Source Part of Speech Tagger for Swedish”. Northern European Journal of Language Technology, 2013, Vol. 3, pp 1–18

[23] Collins, Michael. (2002). “Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms”. In Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, pages 18. Philadelphia, PA, USA.

[24] King, S. (2014). “Measuring a decade of progress in Text-to-Speech”. Loquens, 1(1).

[25] King, S. and Karaiskos, V. (2012). “The Blizzard Challenge 2012”, In Procedings Blizzard Workshop, 2012.

[26] Norrenbrock, C. R., Hinterleitner, F., Heute, U., and Möller, S. (2012). “Towards Perceptual Quality Modeling of Synthesized Audiobooks – Blizzard Challenge 2012
”, In Procedings Blizzard Workshop, 2012.

[27] Edlund, J., Tånnander, C. and Gustafson, J. (2015). "Audience response system-based assessment for analysis-by-synthesis", In proceedings of ICPhS 2015, Glasgow, UK