Open source voice creation toolkit for the MARY TTS Platform

Marc Schröder,[1] *Marcela Charfuelan,*[1] *Sathish Pammi,*[1] *Ingmar Steiner*[2]

[1]DFKI GmbH, Language Technology Lab
Saarbrücken and Berlin, Germany
[2]INRIA/LORIA Speech Group, Nancy, France
firstname.lastname@{dfki.de|inria.fr}

Abstract

This paper describes an open source voice creation toolkit that supports the creation of unit selection and HMM-based voices, for the MARY (Modular Architecture for Research on speech Synthesis) TTS platform. The toolkit can be easily employed to create voices in the languages already supported by MARY TTS, but also provides the tools and generic reusable run-time system modules to add new languages. The voice creation toolkit is mainly intended to be used by research groups on speech technology throughout the world, notably those who do not have their own pre-existing technology yet. We try to provide them with a reusable technology that lowers the entrance barrier for them, making it easier to get started. The toolkit is developed in Java and includes an intuitive Graphical User Interface (GUI) for most of the common tasks in the creation of a synthetic voice. We present the toolkit and discuss a number of interoperability issues.

Index Terms: Speech synthesis, multilingual, unit selection, HMM-based synthesis

1. Introduction

The task of building synthetic voices requires not only a large number of steps but also patience and care, as advised by the developers of Festvox [1], one of the most important and popular tools for creating synthetic voices. Experience with creating synthetic voices has shown that going from one step to another is not always a straightforward task, especially for users who do not have expert knowledge of speech synthesis or when a voice should be created from scratch. In order to simplify the task of building new voices we have created a toolkit aimed to streamline the task of building a new synthesis voice from text and audio recordings. This toolkit is included in the latest version of MARY TTS[1] (version 4.3) and supports the creation of unit selection and HMM-based voices on the languages already supported by MARY: US English, British English, German, Turkish, Russian, and Telugu.

Among the open source voice creation toolkits available nowadays by far the most used system is Festival and its sister project Festvox [1], which offers a free, portable, language independent, run-time speech synthesis engine for various platforms under various APIs with detailed documentation for building new synthetic voices. Another popular speech synthesis tool, free for non-commercial use but not available as open source, is the MBROLA system, a speech synthesiser based on the concatenation of diphones. For creating a new voice, a diphone database should be provided to the MBROLA team, who will process and adapt it to the MBROLA format for free. The resulting MBROLA diphone database is made available for non-commercial, non-military use as part of the MBROLA project [2]. One of the latest open source systems for developing statistic fully parametric speech synthesis is the HMM-based Speech Synthesis System (HTS), first version released in 2002. HTS and hts_engine API do not include any text analyser, but the Festival system, the MARY TTS system, or any equivalent text analyser can be used with HTS. For the MARY TTS system the hts_engine has been ported to Java [3]. The development of a new voice with the HTS system is fully documented including training demos [4].

Most of the systems described above provide guidelines and recipe lists of the steps to follow in order to create a new voice. In most of the cases, it is expected that the developer has a basic knowledge of the TTS system and speech signal processing and in some cases minimal programming skills. In MARY TTS, we have developed, in Java, a voice creation toolkit [5] that includes graphical user interfaces (GUIs) for most of the common tasks in the creation of a synthetic voice. This aims to facilitate the understanding of the whole process and to streamline the steps. All the steps are documented in the MARY wiki pages (http://mary.opendfki.de/wiki) and there is also the possibility to get support via the MARY mailing lists. The toolkit supports the creation of new voices for existing languages, as well as a set of tools and generic run-time system modules for adding support for a new language to MARY TTS.

The paper is organised as follows: in Section 2 we describe the MARY multilingual voice creation toolkit, explaining briefly the support for the creation of a new language and the voice building process. We report on some experience with the toolkit in Section 3. We discuss a number of interoperability issues in Section 4.

2. Toolkit workflow

The steps required to add support for a new language from scratch are illustrated in Figure 1. Two main tasks can be distinguished: (i) building at least a basic set of natural language processing (NLP) components for the new language, carrying out tasks such as tokenisation and phonemic transcription (left branch in Figure 1); and (ii) the creation of a voice in the new language (right branch in Figure 1). Whereas high-quality support of a language will usually require language-specific processing components, it is often possible to reach at least a basic support for a language using generic methods [1]. In the following two sections we briefly explain these two branches.

[1] http://mary.dfki.de/

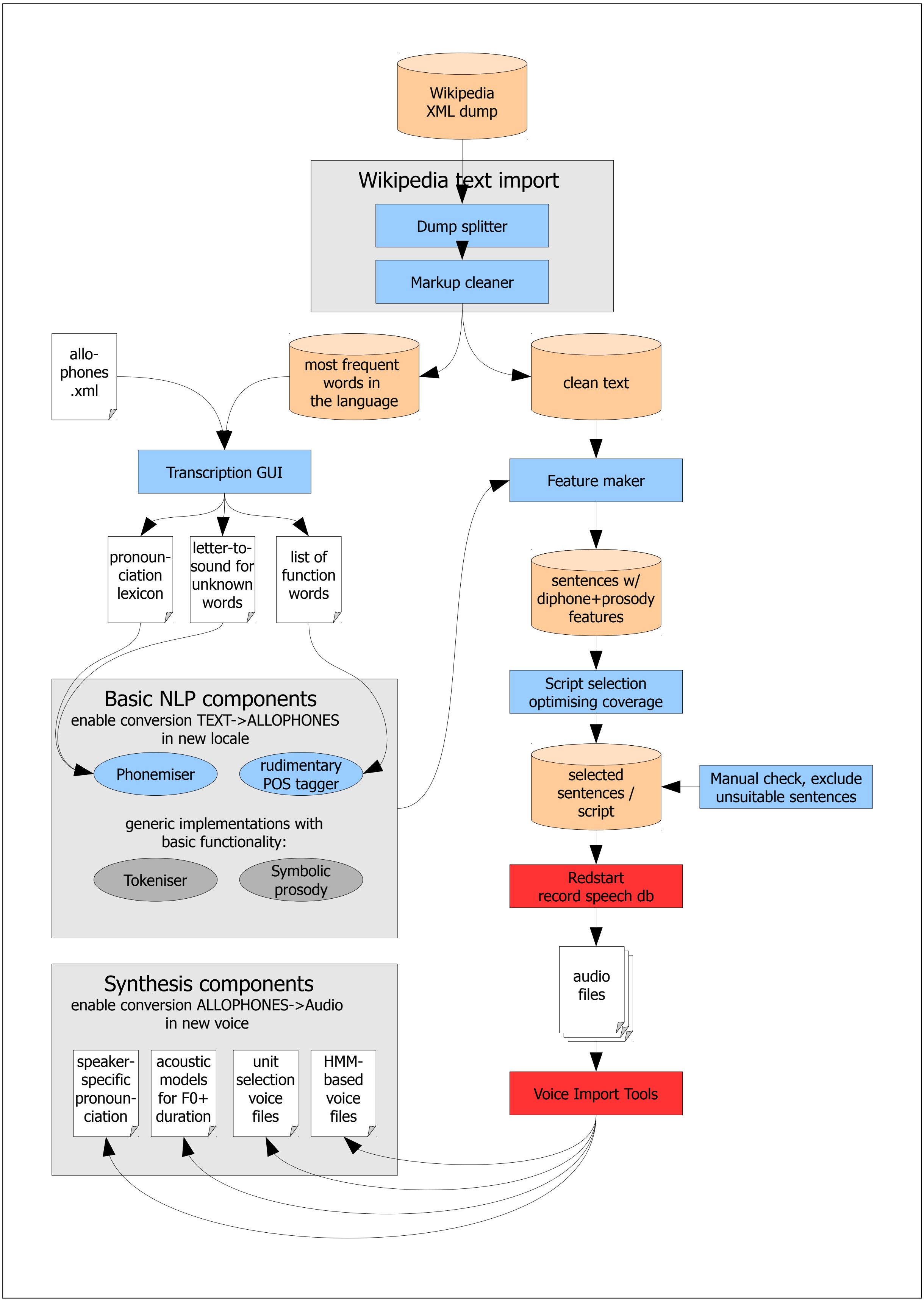
Wikipedia XML dump
Wikipedia text import
Dump splitter
Markup cleaner
allo-phones .xml
most frequent words in the language
clean text
Transcription GUI
Feature maker
pronoun-ciation lexicon
letter-to-sound for unknown words
list of function words
sentences w/ diphone+prosody features
Basic NLP components
enable conversion TEXT->ALLOPHONES in new locale
Phonemiser
rudimentary POS tagger
generic implementations with basic functionality:
Tokeniser
Symbolic prosody
Script selection optimising coverage
selected sentences / script
Manual check, exclude unsuitable sentences
Redstart record speech db
audio files
Voice Import Tools
Synthesis components
enable conversion ALLOPHONES->Audio in new voice
speaker-specific pronoun-ciation
acoustic models for F0+ duration
unit selection voice files
HMM-based voice files

Figure 1: *Workflow for multilingual voice creation in MARY TTS, more information about this tool can be found in: http://mary.opendfki.de/wiki/VoiceImportToolsTutorial*

2.1. New language support

Our workflow starts with a substantial body of Unicode text in the target language, such as a dump of Wikipedia in that language. After automatically extracting text without markup, as well as the most frequent words, the first step is to build up a pronunciation lexicon, letter-to-sound rules for unknown words and a list of function words.

Our toolkit provides a *Transcription GUI*, which a language expert can use to generate a pronunciation dictionary. An `allophones.xml` file defines the allophones of the target language that can be used for transcription, and it characterises them using a set of phonetic features. The features include length, height, frontness and lip rounding for vowels, as well as type, place of articulation and voicing for consonants. If other features are distinctive in a given language, additional features can be added without any problems. The allophones file has to be prepared manually by a language expert.

Once the allophones file for a target language is available, a language expert or at least a native speaker can use the Transcription GUI to transcribe as many of the most frequent words as possible using the allophone inventory. The tool supports this task by training, on the available data, a letter-to-sound predictor which can propose candidate transcriptions for untranscribed words. Furthermore, it is possible to mark function words in the list in order to enable a simplistic POS tagger, which works based on simple context-free string matching. Where a better quality POS tagger or morphological analysis is required, a custom TTS module needs to be implemented. This is unproblematic due to the modular architecture of the MARY TTS system.

With this minimal manual input for a new language, a simple NLP system can be built, using a generic tokeniser and a rule-based prediction of symbolic prosody.

2.2. Voice-building process

Once the NLP component has been developed, the task of creating a voice can be pursued (right branch in Figure 1).

First, a recording script providing good diphone and prosodic coverage is selected from the text collection [6]. Using the NLP components a *feature maker* component annotates each sentence in the text database with diphone and prosody features to be used in a greedy selection. The resulting collection of sentences can be used as the recording script for voice recordings with our tool *Redstart*. The recorded audio files can then be processed by our voice import tools which generate a unit selection and/or an HMM-based voice, as well as speaker-specific prediction components for acoustic parameters. If, during the voice-building process, force-aligned transcriptions were manually corrected, it is also possible to predict *speaker-specific pronunciations*. In the following these steps are explained in more detail.

2.2.1. Optimal text selection

Creating a recording script that provides a good diphone and prosodic coverage is not a trivial task. In the MARY voice creation toolkit a greedy algorithm is used for selecting sentences to optimise coverage. Three parameters are taken into account: the units, coverage definition and stop criteria. Units are defined as vectors consisting of three features: phone, next phone and prosody property. The definition of coverage fixes which kind of units are wanted in the final set; in the current version all diphones and their prosodic variation are used. Other aspects like frequency weights, sentence length, features weight, etc. can be set for optimising the coverage. The stop criteria are a combination of number of sentences, maximum diphone coverage and maximum prosody coverage [6]. The selected sentences then need to be manually checked in order to discard any problematic sentences – e.g., sentences that are too long or that contain words that might be too difficult to pronounce fluently.

If the aim is to support a specific domain, it is possible to either use domain-specific material instead of general-domain text as the basis for selection, or, if the domain is small enough, to manually design a representative set of sentences.

2.2.2. Voice recording

MARY comes with a tool called Redstart to assist the user in the process of voice recording. The tool displays sentences one by one, and records each sentence into a separate wave file. An estimate of recording time is used to pace the recordings; beep sounds indicate when the microphone is opened and closed. Checks for temporal and amplitude clipping are automatically performed; if in doubt regarding the quality of a recording, the user can play the recorded waveform, display the speech signal and the corresponding spectrogram, pitch, and energy contours, and of course re-record the sentence. No files are overwritten, a history of attempts to utter a given sentence is kept. This way, it is possible to revert to the best recording achieved rather than having to try until a perfect version is produced.

2.2.3. Voice import components

The voice import tool combines an extensible list of components in a simple GUI, designed primarily to facilitate the creation of new voices by users without expert knowledge of speech synthesis. Several voice import components execute high quality, freely available components specialised for particular tasks, for example, for automatic labelling we can use Festvox EHMM, or for training HMM models we use the scripts provided by HTS adapted to the MARY TTS architecture. Our toolkit provides reasonable baseline configuration settings to external tools to allow non-expert users to execute the tools in a default setting; experts are given the option to configure many aspects if needed.

3. Experience with the toolkit

The toolkit has been successfully applied to the creation, from scratch, of Turkish, Mandarin Chinese, British English and Telugu text-to-speech systems at DFKI.

Members of the open source community have contributed support for Russian, and are working on a number of additional languages, including Italian, Spanish, Swedish, French, and Greek.

To give an example, for the creation of the Turkish voice, the selection of sentences (1170 sentences, approximately 1 hour and 38 minutes of recording) and the semi-automatic transcription of word pronunciations (approximately 500 words) took around 2 weeks, the recordings were done in two days and the creation of a Turkish unit selection voice took between one and two days for a member of the MARY team. The creation of a Turkish HMM-based voice took approximately two days, also for a member of the MARY team.

The unit selection and HMM-based voices available in MARY 4.3 are presented in Table 1.

Table 1: *Voices freely available in the MARY TTS system version 4.3. Gen.: gender (M) male, (F) female. The GB English voices are expressive voices built for the SEMAINE project; dfki-pavoque-styles is a multi-style voice created for the PAVOQUE project. All other voices are non-expressive.*

Language	Gen.	Name	Unit selection	HMM-based
German	M	bits3	X	X
	F	bits1		X
	M	dfki-pavoque-neutral	X	X
	M	dfki-pavoque-styles	X	
US English	F	cmu-slt	X	X
GB English	M	dfki-obadiah	X	X
	F	dfki-poppy	X	X
	F	dfki-prudence	X	X
	M	dfki-spike	X	X
Turkish	M	dfki-ot	X	X
Telugu	F	cmu-nk	X	X
Russian	M	voxforge-nsh	X	

4. Interoperability issues

Interoperability is relevant at two levels. On the one hand, the MARY voice import toolkit includes of a number of steps for which third-party software is used: EHMM or Sphinx for force-alignment; Wavesurfer or Emu-label for manually correcting labels; Praat or Snack for pitch marking; wagon from the Edinburgh Speech Tools or Weka for training decision trees. Each of these tools has custom requirements to the respective input data and produces output data in a custom format. Even where de facto standards exist, such as the XWaves file format for label files, tools turn out to show minor differences. Every time a different tool is to be tried out, substantial effort needs to go into interfacing. Work would be substantially simpler if standard file formats could be agreed on, or if standard programming interfaces were available for interacting with tools implementing a certain functionality.

The other level where interoperability is of relevance to MARY TTS is the use of the speech synthesis itself. Here, the TTS is the tool, to be used by others. The same desideratum applies here: users would benefit if they could use MARY through the same interface as other TTS engines. Platform-specific interfaces exist, such as the Microsoft SAPI, the Macintosh Speech Synthesis API, or Gnome Speech and Speech Dispatcher for Linux; however, it is difficult to justify investing effort into supporting these interfaces even as a side activity in a research project. For this reason, the MARY TTS system currently only supports its own custom API and HTTP protocol.

5. Conclusions

We have presented a multilingual voice creation toolkit that supports the user in building voices for the open source MARY TTS platform, for two state-of-the-art speech synthesis technologies: unit selection and HMM-based synthesis. For languages not yet supported by MARY TTS, the toolkit provides the necessary tools and generic reusable run-time system modules for adding support for a new language.

The toolkit is mainly intended to be used by research groups in speech technology throughout the world, notably those who do not have their own pre-existing technology yet. We try to provide them with a reusable technology that lowers the entrance barrier for them, making it easier to get started. The whole process of creating a synthetic voice is fully documented in the MARY wiki pages and there is also the possibility to get support by subscribing to the MARY mailing lists. Our experience with the toolkit has demonstrated that it enables rapid development of new voices with good quality.

We have pointed out the value we see in improving interoperability between tools by defining standard file formats and API interfaces. However, we are aware that it is difficult to find funding for efforts in this direction.

Future improvements to the MARY TTS voice building toolkit and synthesis framework are planned on the level of error handling, robustness and ease of use, as well as on individual component technologies. For example, signal processing techniques can be used to reduce the amount of audible concatenation artifacts in unit selection voices. More sophisticated prosody models may be able to improve the naturalness of the synthetic speech, especially for HMM-based voices. As core technologies, these improvements will improve the quality of MARY voices independently of their language.

6. Acknowledgements

This work reported in this paper has received funding from the DFG project PAVOQUE and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements no. 211486 (SEMAINE), 248116 (ALIZ-E) and 231287 (SSPNet).

7. References

[1] A. W. Black and K. Lenzo, "Festvox: Building synthetic voices, Version 2.1," http://www.festvox.org/bsv/, 2007, (accessed March 2010).

[2] T. Dutoit, F. Bataille, V. Pagel, O. Pierret, and O. Van der Vreken, "The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," in *Proc. ICSLP*, Philadelphia, USA, 1996.

[3] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard Challenge 2008," in *Proc. of the Blizzard Challenge 2008*, 2008.

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) Version 2.0," in *The 6th International Workshop on Speech Synthesis*, 2006.

[5] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual voice creation toolkit for the MARY TTS platform," in *Proc. LREC*. Valettea, Malta: ELRA, 2010. [Online]. Available: http://www.dfki.de/lt/publication_show.php?id=4878

[6] A. Hunecke, "Optimal design of a speech database for unit selection synthesis," Master's thesis, Fachrichtung 4.7 Allgemeine Linguistik, Universität des Saarlandes, 2007.