Creation and evaluation of MaryTTS speech synthesis for Polish

Jolanta Bachan, Marceli Tokarski

Institute of Linguistics
Adam Mickiewicz University in Poznań
al. Niepodległości 4, Poznań, Poland
jolabachan@gmail.com, marcel530@gmail.com

Abstract

The present paper describes creation and evaluation of a Polish synthetic voice for MaryTTS speech synthesis system. The method applied for the synthesis was unit-selection. The Polish voice was created from the available Polish text and audio resources, and with the use of already existing automatic grapheme-to-phoneme converter and a set of Python and Praat scripts for text data manipulations. The result of the work was a text-to-speech synthesis system for Polish in MaryTTS. The Polish MaryTTS synthetic speech output was evaluated in speech perception tests by 12 people. The result of the word recognition test was over 80%, whereas the Mean Opinion Score (MOS) of speech quality was 2.74 in the 5-point rating scale.

Keywords: speech synthesis, MaryTTS, Polish, NLP, speech perception tests

1. Introduction

A text-to-speech (TTS) synthesiser is a computer system which converts written text into human-like speech. Ideally, a TTS system should be able to read any text which is input to the system, but in practice it is not easy to achieve and TTS systems still need improvements (Dutoit, 1997).

There are several different methods to synthesise speech. These methods may be classified into three groups (1) parametric synthesis – synthesis by rule, e.g. formant synthesis, articulatory synthesis, (2) concatenative synthesis, e.g. diphone synthesis, unit selection synthesis, (3) statistical parametric synthesis – HMM-based synthesis.

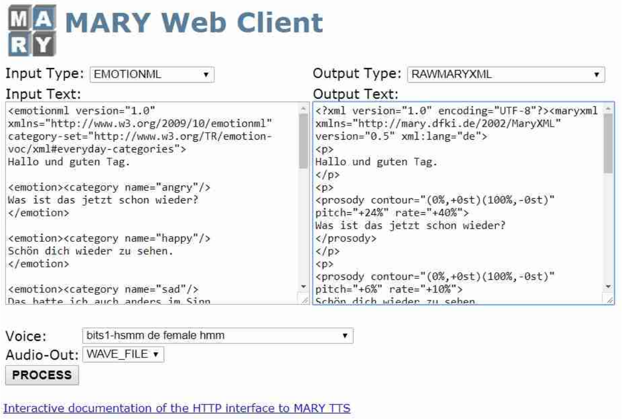
In the present paper a creation and evaluation of unit-selection speech synthesis for Polish in MaryTTS system is described. The Polish language component of MaryTTS was created using the available Polish resources. The generated synthetic speech was evaluated in perception tests for correct word and sentence recognition and speech quality.

The motivation for creating the Polish synthetic voice is the ongoing research on phonetic convergence between humans and in human-computer interaction. One of the project goals is to create phonetic convergence models applicable in dialogue systems (Demenko and Bachan, 2017; Bachan *et al.,* 2017).

2. MaryTTS

MaryTTS[1] (Modular Architecture for Research on speech sYnthesis) is a multilingual open-source text-to-speech platform written in Java (Schröder *et al.*, 2011; Steiner *et al.,* 2017). Up till now languages subject to synthesis were German, British and American English, French, Italian, Luxembourgish, Russian, Swedish, Telugu, and Turkish. The platform may use three speech synthesis techniques: diphone concatenation, unit-selection and HMM-based synthesis. For the first attempt to build the synthetic voice for Polish in MaryTTS, the unit-selection method was used.

Input to MaryTTS can be represented in the form of text, phonemes, XML format with intonation or emotion markers and many others. For the phonetic convergence study, the ability of modifying intonation matters most. Apart from text-to-speech synthesis, the system has other functionalities such as generating a list of allophones, tokens, part of speech tags, intonation in the XML format or TextGrid Praat format for phones for selected languages. Audio output is available in three formats: WAV, Au and AIFF. Fig. 1 shows a MaryTTS web interface with an exemplar input in EmotionML format (Charfuelan and Steiner, 2013) and RawMaryXML format output and female German voice selected for HMM-based speech synthesis.



Input Type: EMOTIONML

Input Text:

```
<emotionml version="1.0"
xmlns="http://www.w3.org/2009/10/emotionml"
category-set="http://www.w3.org/TR/emotion-
voc/xml#everyday-categories">
Hallo und guten Tag.

<emotion><category name="angry"/>
Was ist das jetzt schon wieder?
</emotion>

<emotion><category name="happy"/>
Schön dich wieder zu sehen.
</emotion>

<emotion><category name="sad"/>
Das hatte ich auch anders im Sinn
```

Output Type: RAWMARYXML

Output Text:

```
<?xml version="1.0" encoding="UTF-8"?><maryxml
xmlns="http://mary.dfki.de/2002/MaryXML"
version="0.5" xml:lang="de">
<p>
Hallo und guten Tag.
</p>
<p>
<prosody contour="(0%,+0st)(100%,-0st)"
pitch="+24%" rate="+40%">
Was ist das jetzt schon wieder?
</prosody>
</p>
<p>
<prosody contour="(0%,+0st)(100%,-0st)"
pitch="+6%" rate="+10%">
Schön dich wieder zu sehen
```

Voice: bits1-hsmm de female hmm

Audio-Out: WAVE_FILE

Fig. 1: MaryTTS Web Client interface

3. Polish data for MaryTTS

The NLP components for MaryTTS for Polish were created from existing Polish text and audio resources: an online newspaper text corpus and an excerpt of a speech corpus created for BOSS synthesis system (Bonn Open Synthesis System) (Klabbers *et al.*, 2001). To prepare the text data for MaryTTS needs, a set of already existing Python and Praat scripts were used. The scripts were used

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

to adopt the BOSS annotation BLF file format to LAB format required by MaryTTS. The scripts were created in the previous work on data format conversion (Bachan, 2011). Finally, for automatic transcription of text, Polphone, a grapheme-to-phoneme converter for Polish, was applied. Automatic transcription allowed to build a pronunciation lexicon which was used to train the letter-to-sound rules in MaryTTS.

Fig. 2 shows the data flow for Polish synthetic voice creation in MaryTTS from existing text and audio resources. The data conversion steps are described in the following sections.

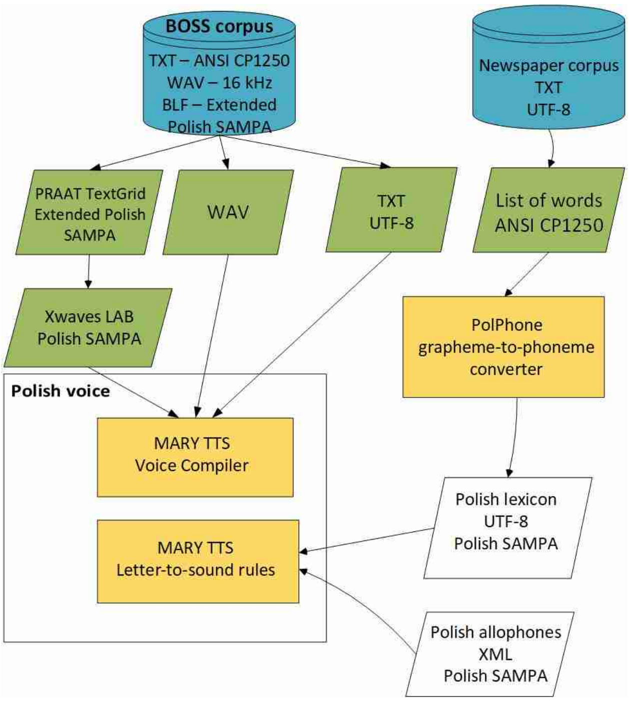


Fig. 2: Data flow for Polish synthetic voice creation in MaryTTS from existing resources

3.1 Text resources

For creating a Polish lexicon, a small newspaper corpus in TXT format was used. The corpus was created from online news articles in 2010 and was to serve as input to Phonetically Rich Diphone Extractor software. The aim of the software was to select the smallest possible set of sentences from a text corpus which would contain the largest number of diphones (Bachan, 2010). The corpus contained 16,381 of different word types (tokens without duplications) which included 1,100 diphones (Bachan, 2011). The Polish SAMPA contains 37 labels (Wells, 1996). Adding the pause to it, it gives a maximum of 38*38=1444 diphones. Some of the diphones do not exist in the spoken language, so the number of 1,100 diphones was estimated as a fair diphone coverage in Polish.

3.2. Polish pronunciation lexicon for MaryTTS

The newspaper corpus was encoded in UTF-8, but had to be converted to ANSI "cp1250" as that was the encoding accepted by the Polish automatic transcription program PolPhone (Wypych *et al.*, 2002; Szymański and Grocholewski, 2005). Then a list of 16,381 unique words was created and input to PolPhone. The grapheme-to-phoneme conversion program generated 18,644 entries as one word could have a few variants of pronunciation. Such prepared data was encoded back to UTF-8 and compiled by the MaryTTS tools to create the Polish pronunciation lexicon and to train letter-to-sound rules. Because some of the entries had to be removed, the final list of words in the Polish lexicon for MaryTTS was 16,851. The format of the lexicon entries is presented below. The first row are the words written orthographically, next the phonemes are written one by one, separated by a space. The dot "." stands for the beginning of a syllable and a double vowel marks the lexical stress on a syllables in a word.

```
kazał k aa . z a w
kazała k a . z aa . w a
krzyczeć k Sz yy . tSz e tsi
ksiądz k si oo n dz
nieoszczędny ni e . o Sz . tSz ee n . d n y
prawomocnie p r a . v o . m oo . ts ni e
przed p Sz e t
przed p Sz e d
reporter r e . p oo r . t e r
```

Apart from the pronunciation dictionary, a list of allophones for Polish had to be created in an XML file format. The list of allophones contained 37 allophones and was consistent with Polish SAMPA (not the Extended-Polish SAMPA which contains 40 phonemes, (Demenko *et al.*, 2003, Bachan, 2007))

The XML allophones file describes a few sound features: vowel height, vowel frontness, vowel roundness, consonant type, place of consonant articulation, consonant vocalisation. An excerpt of the XML file with the Polish allophones is presented below.

```
<allophones name="sampa" xml:lang="pl"
features="vheight vfront vrnd ctype cplace cvox">
<silence ph="_"/>
<vowel ph="a" vheight="3" vfront="3" vrnd="-"/>
<vowel ph="e" vheight="2" vfront="1" vrnd="-"/>
<vowel ph="i" vheight="1" vfront="1" vrnd="-"/>
<vowel ph="o" vheight="2" vfront="3" vrnd="+"/>
<consonant ph="p" ctype="s" cplace="l" cvox="-"/>
<consonant ph="b" ctype="s" cplace="l" cvox="+"/>
<consonant ph="ts" ctype="a" cplace="a" cvox="-"/>
<consonant ph="dz" ctype="a" cplace="a" cvox="+"/>
<consonant ph="s`" ctype="f" cplace="p" cvox="-"/>
<consonant ph="z`" ctype="f" cplace="p" cvox="+"/>
</allophones>
```

3.3. Audio resources

To create the Polish synthetic voice for MaryTTS, an existing BOSS (Bonn Open Synthesis System) corpus was used. The corpus consists of approximately 3,240 utterances, read by a professional male speaker in a professional recording studio. The sampling rate of the data is 16kHz and the recordings are saved in the standard WAV format (Demenko *et al.*; 2007). The BOSS corpus is divided into 5 sets. Each of the sets was created for different purposes. For the present work, only two sets were used:

- Base A – 289 sentences with the most frequent Polish consonant clusters, duration: 14min 42sec
- Base B – 109 meaningless sentences which aimed to contain all Polish diphones, duration: 3min 27sec

The BOSS corpus, along with the recording in the WAV format for each of the sentences, provides the sentence annotation in the BLF (BOSS label file) format on the phone level and its orthographic text in a TXT file.

Because the MaryTTS voice compiler requires Xwaves LAB format, the BLF format had to be converted to the LAB format. Moreover, the phoneme set in BOSS has 40 allophones and is compatible with Extended-Polish SAMPA, but the pronunciation dictionary and the set of allophones in MaryTTS was created for 37 sounds. Therefore, the mismatches had to be removed and the phoneme sets had to be unified to Polish SAMPA with 37 labels. This was done by, first, converting BLF files to Praat TextGrid file format (Boersma and Weenink, 2001) using a Python script. During this step, the additional prosody markers and special characters were removed from the annotation. Second, the TextGrid files were converted to the LAB Xwaves format using a Praat script while the phoneme labels were normalised to Polish SAMPA. Last but not least, the TXT files with orthographic texts were automatically converted from ANSI “cp1250” to UTF-8 encoding. Such a triplet (a WAV file, a LAB file and a TXT file) for each of the sentences was input to MaryTTS unit-selection voice compiler and a Polish synthetic voice was created.

4. Evaluation

In the course of the preliminary evaluation of the Polish speech synthesis in MaryTTS the following was observed:

- The synthesized speech was understandable after having listened to it once or twice.
- Some interruptions in the speech signal occurred, but they did not make the synthetic speech difficult to understand.

These promising results fed into carrying out the speech quality assessment tests. The design of the tests was corresponding to the tests of a Polish male MBROLA synthetic voice (Bachan, 2007) and met the EAGLES standards (Gibbon *et al*., 1997).

The speech output assessment tests were carried out on 12 Polish subjects (6 males and 6 females) and are described in the following sections.

4.1. Sentence and word recognition test

Method: Meaningful and Meaningless synthesised sentences were presented to the subjects. The subjects were asked to write down what they heard in an answer sheet. The set of meaningless sentences, i.e. semantically unpredictable sentences, was used to eliminate the influence of the top-down processing (Clark & Yallop 1995: 312, Ryalls 1996: 94).

Material: 10 meaningful and 10 meaningless sentences.

Instructions: In a moment you will hear 20 sentences. Your task is to write down the sentences. After each sentence, there is a few-second pause. This is the time for you to write down the sentence. You can play the sound once or twice.

The results of the word recognition tests are presented in Table 1. In the sentences, there were 126 words (in the meaningful sentences: 75, in the semantically unpredictable sentences: 51). The word recognition was quite high: 81% for female and 88% for male subjects.

Subjects	N	Words (absolute)	Words (%)
Polish male	6	110.33	88
Polish female	6	101.67	81
Polish overall	12	106	84

Table 1. Average correctly recognised words in all sentences. N stands for the number of subjects

Table 2 presents the comparison of separate results for the semantically predictable (meaningful) and unpredictable sentences (meaningless). The sentence recognition rate was low (around 54% for meaningful sentences), because the sentence was counted as unrecognised if at least one word in the sentence was incorrectly recognised. The difference between the word recognition of meaningful and meaningless sentences equals 6% points. This suggests that the semantically predictable structure of the sentences could help in recognising words. When the top-down component was eliminated from the speech perception process, the recognition of single words was worse.

In the same test, the Polish MBROLA female voice scored 96.28% in semantically predictable sentences and 81.53% in semantically unpredictable sentences (Bachan, 2007).

Units	N	Meaningful (%)	N	Meaningless (%)
Sentences	10	54.17	10	42.50
Words	75	86.56	51	80.56

Table 2. Sentence and word recognition rates for meaningful vs. meaningless stimuli

4.2. Subjective speech quality test

Method: The subjects were asked to evaluate the quality of isolated long (multiple) sentences at 5-point Mean Opinion Score (MOS) scale: Excellent – Good – Fair – Poor – Bad, where 1 was the lowest grade and 5 was the highest.

Material: 9 different compound sentences, the sentences were played in a random order.

Instructions: In a moment you will hear 9 long sentences. Your task is to evaluate the quality of the speech. After each sentence, there is a few-second pause. This is the time for you to decide which of the five grades you would give to the utterance: Excellent – Good – Fair – Poor – Bad.

Table 3 shows the test results for Polish male and female subjects, and their average. In the overall score the MaryTTS male voice received 2.74 points. This result is comparable with MBROLA speech synthesis for Polish female voice which received 2.72 points when the same sentences were used and synthesised using the Close Copy Speech Synthesis method (Bachan, 2007).

Subjects	N	MOS score
Polish male	6	2.83
Polish female	6	2.65
Polish overall	12	2.74

Table 3. Judgement quality test results

5. Conclusions

In the present paper, the creation and evaluation of the Polish voice for MaryTTS speech synthesis system was presented. The voice was created using the available Polish resources: text and speech corpora and automatic tools and scripts. The text data and automatic transcription program PolPhone were used to build a Polish pronunciation dictionary to train the letter-to-sound rules for MaryTTS. This made it possible to create a full text-to-speech NLP component for Polish.

The unit-selection speech synthesis for Polish was built on only 18min 9sec of speech. The speech perception tests of word recognition were promising and the score of 2.74 in the MOS scale was comparable with similar systems. However, the speech still needs some improvement in the naturalness of voice and eliminating the cracks.

In the future, a development of HMM-based speech synthesis in MaryTTS is planned which will allow for more modifications of prosody and testing the models of phonetic convergence in human-computer communication.

6. Acknowledgements

The present study was supported by the Polish National Science Centre, project no.: 2014/14/M/HS2/00631, “Automatic analysis of phonetic convergence in speech technology systems” and was conducted in cooperation with the project partner Prof. Bernd Möbius who is the leader of a project “Phonetic convergence in Human-Machine Communication.”

The authors would like to particularly thank Dr. Ingmar Steiner for his help in compiling the Polish lexicon and creating the NLP component code for MaryTTS and for building the unit-selection voice.

References

Bachan, J. (2007). *Close Copy Speech Synthesis for Perception Testing and Annotation Validation.* M.A. Thesis. Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland

Bachan, J. (2009/2010). Tools for automatisation of voice creation for diphone based speech synthesis. In: Demenko, G. and Wagner, A. (Eds.) *Speech and Language Technology*. Volume 12/13. Poznań (Eds). Polish Phonetic Association, pp. 229-237

Bachan, J. (2011). *Communicative alignment of synthetic speech.* Ph.D. Thesis. Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland

Bachan, J., Owsianny, M. and Demenko, G. (2017, current proceedings) Creation of a Dialogue Corpus for Automatic Analysis of Phonetic Convergence. In: Vetulani, Z. and Paroubek, P. (Eds.) *Proceeding of 8th Language & Technology Conference.* 17-19 November 2017, Poznań, Poland

Boersma, P. & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. In: *Glot International* 5(9/10), pp. 341-345

Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In: *INTERSPEECH-2013*, Lyon, France, pp. 1564-1568

Clark, J. and Yallop, C. (1995). *An introduction to phonetics and phonology*. 2nd edition. Oxford: Blackwell

Demenko, G., Wypych, M. and Baranowska, E. (2003). Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: Demenko, G. and Karpiński, M. (Eds.) *Speech and Language Technology*, Vol. 7. Poznań: Polish Phonetic Association, pp. 79-95

Demenko, G., Klessa, K. Szymański, M. and Bachan, J. (2007). The design of Polish speech corpora for speech synthesis in BOSS system. In: *Preceedings of XII SympozjumPodstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki* (PPEEm2007). Wisła, Poland, pp. 253-258

Demenko, G. and Bachan, J. (2017). Annotation specifications of a dialogue corpus for modelling phonetic convergence in technical systems. In: Trouvain, J. Steiner, I. and Möbius, B. (Eds.): *Proceedings of 28th Conference on Electronic Speech Signal Processing (ESSV)*. 15–17 March 2017, Saarbrücken, Germany, pp. 75-82

Dutoit, T. (1997). *An Introduction To Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Dutoit, T. (2005). The MBROLA project. Access date: September 30, 2017

Gibbon, D., Moore, R. and Winski, R. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter

Klabbers, E., Stöber, K., Veldhuis, R, Wagner, P. and Breuer, S. (2001). Speech Synthesis Development Made Easy: the Bonn Open Synthesis System. In: *Eurospeech-2001*, pp. 521-525

Ryalls, J. (1996). *A Basic Introduction to Speech Perception*. San Diego, California: Singular Publishing Group, Inc., and London: Singular Publishing Ltd

Schröder, M., Charfuelan, M., Pammi, S. and Steiner, I. (2011). Open source voice creation toolkit for the MARY TTS platform. In: *INTERSPEECH-2011*, pp. 3253-3256

Steiner, I., Le Maguer, S., Manzoni, J., Gilles P. and Trouvain, J. (2017). Developing new language tools for MaryTTS: the case of Luxembourgish. In: Trouvain, J. Steiner, I. and Möbius, B. (Eds.): *Proceedings of 28th Conference on Electronic Speech Signal Processing (ESSV)*. 15–17 March 2017, Saarbrücken, Germany, pp. 186-192

Szymański, M. and Grocholewski, S. (2005). Transcription-based automatic segmentation of speech. In: *Proceedings of 2nd Language and Technology Conference*, Poznań, pp. 11–14

The MARY Text-to-Speech System (MaryTTS). http://mary.dfki.de/. Access date: September 30, 2017.

Wells, J.C. (1996). SAMPA - computer readable phonetic alphabet. Polish. http://www.phon.ucl.ac.uk/home/sampa/polish.htm. Access date: September 30, 2017

Wypych. M., Demenko, G., Baranowska, E. and Szymański, M. (2002-2006). PolPhone – Polish phonetizing filter. Based on phonetization rules from M.S. Batóg and W. Jassem