Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts
der Philosophischen Fakultät der Universität Zürich

Improving comprehensibility of rule-based text-to-speech output: eSpeak-NG and Polish

Verfasst von: Victor Bielawski
Matrikel-Nr: 10-741-940

Betreut von: Dieter Andreas Studer

Institut für Computerlinguistik

Abgabedatum: 01.07.2022

Abstract

eSpeak NG is a freely available additive rule-based speech synthesis system with support for many languages. Although eSpeak NG translates Polish text into phonemes nearly ideally, given its constraints, there is room for improvement in its realization of the language's phonemes. This paper describes a set of changes to the system's Polish phoneme definitions and voice paramaters to improve the comprehensibility of its output.

Zusammenfassung

eSpeak NG ist ein frei verfügbares additives und regelbasiertes Sprachsynthesesystem mit Unterstützung für viele Sprachen. Obwohl eSpeak NG im Rahmen seiner Möglichkeiten polnische Texte nahezu ideal in Phoneme übersetzt, gibt es bei der Realisierung dieser Phoneme Verbesserungspotenzial. In dieser Arbeit wird eine Reihe von Änderungen an den polnischen Phonemdefinitionen und Stimmparametern beschrieben mit dem Ziel, die Verständlichkeit der Ausgaben zu erhöhen.

Contents

Abbreviations

ALSA	Advanced Linux Sound Architecture
F	formant
HMM	hidden Markov model
Hz	hertz
ms	milliseconds
IPA	International Phonetic Alphabet
OSS	Open Sound System
RNN	recurrent neural network
s	seconds
TTS	text-to-speech

Glossing abbreviations

2	second person
ACC	accusative case
DAT	dative case
GEN	genitive case
INS	instrumental case
ITER	iterative aspect
LOC	locative case
MASC	masculine gender
NOM	nominative case
PFV	perfective aspect
PL	plural
SG	singular
VIR	virile (gender and animacy)

1 Introduction

In modern times, there is increasing demand for speech produced by machines (text-to-speech, speech synthesis, synthesized speech). On the one hand, automation of tasks traditionally performed by humans often leads to interfaces between human and machine that rely on speech; on the other hand, speech has largely replaced tactile writing as a method of making text accessible to the blind. With people being exposed to increasing amounts of synthesized speech in daily life, it is desirable to minimize negative impacts of synthesized as opposed to natural speech, such as cognitive load due to insufficent comprehensibility, or distraction due to unnatural or aesthetically displeasing qualities of synthetic voices.

Speech synthesis has a long history and can be done in various ways. The most widespread text-to-speech systems, especially for personal use, function by applying rules devised by humans to transform written text into a phonetic representation, which is then converted into audible speech using recordings, models, or both, of human voice. One such rule-based text-to-speech system is eSpeak NG.

In this paper, I describe the design of eSpeak NG, examine its support for the Polish language, identify areas in which comprehensibility and naturalness of its output can be improved, apply concrete improvements to the system, and evaluate their effect through annotation of pre- and post-modification output by Polish speakers.

2 Overview of text-to-speech technology

Text-to-speech (TTS) systems are designed to mimic the output of a human speaker reading a written text out loud. To do this apparently simple task, a TTS system must deal with ambiguities and irregularities on many levels: when splitting a text into sentences and words, when choosing which spoken segments a written word represents, and when deciding with which prosody a word or sequence of words is to be spoken. It must also transform a discrete representation of segments and suprasegmental features into an approximation of a human voice [Rashad et al., 2010]. There are various approaches to solving this problem, each with its advantages and drawbacks.

2.1 Precursors of TTS

There has long been a need for technology that produces speech without human effort. Early steps on the way to current TTS technology include systems that concatenate recordings of sentences, words or syllables, as well as others that recreate parts of the human vocal tract [Van Santen, 2005]. A theme common among these precursor systems is that they do not include processing of a language's orthography; some of them even predate the practice of representing text digitally. Instead, input is given by persons familiar with the system in a format suited for the process that leads to the output.

2.2 Rule-based TTS

The earliest systems that can be called text-to-speech – that is, systems that accept language in its standard written form and convert it to speech – function mechanistically. Some modern TTS systems are rule-based as well. Each step on the way from text to speech is governed by rules defined by the system's authors [Rashad et al., 2010]:

1. Normalization and tokenization: splitting of a text into sentences, phrases and words

2. Grapheme-to-phoneme conversion: conversion of written words into sequences of segments
3. Prosodic analysis: inference of prosody from sentence structure and function words
4. Synthesis: conversion of phonemes and prosody into speech sounds

The synthesis step can use various techniques [Rashad et al., 2010]:

- Concatenation of single phones. Audio recordings of single phones pronounced by a human speaker are played in sequence. They may be adjusted in speed or pitch to suit the calculated prosody.
- Concatenation of diphones or triphones. Instead of using recordings of single phones, each recording runs from the midpoint of one phone to the midpoint of another, possibly including a third phone in between. Unlike concatenation of single phones, this approach takes coarticulation into account.
- Articulatory synthesis. Rather than assembling speech from human-recorded segments, it is constructed de novo using a model of the human speech apparatus. A basic voice, corresponding to the vibrations of the larynx, is passed through a series of filters corresponding to the articulatory gestures modelled for each phone.
- Additive synthesis. Vowels are synthesized de novo as a combination of formant frequencies, while consonants are played from recordings of human speech. This strategy is a compromise that avoids the difficult task of accurately modelling consonant articulation.

2.3 Other approaches

Since approximately the turn of the century, advances in computational power have led to increased use of neural networks and machine learning for tasks previously done by human-written code; text-to-speech is among these. With this approach, the system generates an algorithm based only on examples of input and corresponding output. The algorithms resulting from training such systems are opaque to humans; they may only be influenced by manipulating the input or adding further training data.

2.3.1 Early use of machine learning

The first uses of statistical methods in text-to-speech, including the first ones that could be termed machine learning, replaced only individual parts of the speech synthesis process. For example, models could be trained to determine sentence breaks in written text, to translate graphemes to phonemes, or to produce audio from phonemes. An early example of the latter can be found in the work of Tokuda et al. [2000], which uses a hidden Markov model trained on human speech and transcriptions to, in a sense, produce spectrograms by reversing speech recognition, by predicting the hidden state of formants and spectral parameters from the observed state of phonemes produced by a rule-based algorithm.

2.3.2 End-to-end synthesis by machine learning

When input passes through multiple machine learning steps, they communicate only by their respective output and input; the hidden state is lost. This can result in errors propagating and multiplying at each step. [Ning et al., 2019] To counteract this, models could be trained to perform the entire transformation from text to speech. For example, Tacotron [Wang et al., 2017], a bidirectional recurrent neural network based on seq2seq, can be trained to produce spectrograms directly from text, from which audio samples can be generated using another model, such as WaveRNN [Leviathan and Matias, 2018]. A later refinement to Tacotron [Weiss et al., 2020] generates audio directly and surpasses rule-based systems in naturalness.

3 Polish phonology and orthography

Polish is a West Slavic language spoken by approximately 40 900 000 people [Simons and Fennig, 2017]. It is the official language of Poland and is also traditionally spoken in parts of neighbouring Lithuania, Ukraine and Czechia. This overview of Polish is limited to matters relevant to the project described in this paper: the phonology as well as inconsistencies between phonology and orthography.

Polish is syllable-timed and has a fixed stress on the penultimate syllable. Depending on the analysis, it has five to eight vowels and 28 to 33 consonants, which appear in both onsets and codas. It shows voicing assimilation and final devoicing. Polish orthography is essentially phonemic, but contains some relics of former distinctions that are now merged.

Internal developments as well as intake of loanwords are driving several changes in this phonology.

3.1 Vowels

Polish distinguishes five or six oral vowels:

	i	
		u
ɛ <e>	(ɨ) <y>	ɔ <o>
	a	

Most analyses also posit nasal vowels: /ɔ̃/ and sometimes /ɛ̃/. The vowels [ɨ] and [i] are sometimes described as allophones of a single phoneme, and the consonant inventory is extended to allow this [Jassem, 1958]. There is no distinctive vowel length. Diphthongs are analyzed as a vowel adjacent to an approximant [Wells, 2006].

In this paper, I treat /ɨ/ and /ɔ̃/ as phonemes, to keep the number of consonant phonemes maximally small.

The vowels /a ɨ i/ are each represented by a single grapheme; the others have two:

- /u/ may be represented by either <u> or <ó>

(3.1) <buk> /buk/ ”beech”

(3.2) <Bóg> /bug/ ”God”

- /ɛ/ may be represented by either <e> or <ę>

(3.3) <chwile> /xvilɛ/ ”moments-NOM/ACC”

(3.4) <chwilę> /xvilɛ/ ”moment-ACC”

- /ɔ/ may be represented by either <o> or <ą>.

(3.5) <fioł> /fʲɔw/ ”frenzy, insanity”

(3.6) <wziął> /vʑɔw/ ”took-MASC.SG”

3.2 Consonants

Polish distinguishes at least 28 consonants:

p	b	m	f	v
t	d	n	s	z
t͡s <c>	d͡z <dz>			
t͡ɕ <ć>	d͡ʑ <dź>	ɲ <ń>	ɕ <ś>	ʑ <ź>
t͡ʂ <cz>	d͡ʐ <dż>		ʂ <sz>	ʐ <ż>
k	g		x <ch>	
r	l	w <ł>	j	

Five of these consonants, /t͡ɕ d͡ʑ ɲ ɕ ʑ/, are palatalized variants of /t d n s z/; they transparently alternate with each other in different inflected forms of a word. Some analyses of the language posit phonemic palatalized versions of most other consonants (which also allows the vowels [ɨ] and [i] to be seen as allophones); these can alternatively be described as clusters of said consonants plus /j/. The affricates must be analyzed as phonemes, because they contrast with the corresponding stop-fricative sequences [Wells, 2006].

In this paper, I do not make assumptions about the underlying structure of palatalized consonants not shown in the above table, and represent such palatalization with <ʲ>.

The orthography generally has a one-to-one mapping between consonants and their graphemes, with the following exceptions:

- /m n ɲ/ in codas, word-internally after /ɛ/ and /ɔ/, can be represented either separately, or together with the vowel using the single letters <ę> and <ą>.

(3.7) <kolęda> /kɔlɛnda/ ”carol, carolling”

(3.8) <legenda> /lɛgɛnda/ ”legend”

(3.9) <kąt> /kɔnt/ ”angle”

(3.10) <front> /frɔnt/ ”front(line)”

- /t͡ɕ d͡ʑ ɕ ʑ/ immediately before vowels are represented by the grapheme of the corresponding unpalatalized consonant followed by <i>. This is unproblematic in native words, because the unpalatalized consonants never occur in such contexts, but results in ambiguity in loanwords in which this is permitted.

  (3.11) <siny> /ɕiniɨ/ ”blue-grey”

  (3.12) <Sindbad> /sindbad/ ”S.”

- /ʐ̩/ can be represented by both <ż> and <rz>.

  (3.13) <wierzy> /wʲɛʐ̩ɨ/ ”believes”

  (3.14) <wieży> /wʲɛʐ̩ɨ/ ”tower-GEN/DAT/LOC”

- <rz> can represent both /ʐ̩/ and /rz/.

  (3.15) <marzenie> /maʐ̩ɛɲɛ/ ”dream”

  (3.16) <marznie> /marzɲɛ/ ”freezes”

- /x/ can be represented by both <h> and <ch>.

  (3.17) <huk> /xuk/ ”boom”

  (3.18) <chór> /xur/ ”choir”

3.3 Phonotactics

Each Polish syllable contains a vowel as a nucleus, and may have an onset of up to four consonants and a coda of up to three consonants.

Onsets, excluding any final sonorants, generally assimilate in voicing to their last member. Codas, excluding initial sonorants, assimilate in voicing to the following onset. Sentence-final codas, and depending on dialect, word-final codas, are unvoiced.

Adjacent consonants within an onset or coda also assimilate regressively in palatalization. Across consonants without palatal variants, this kind of assimilation is found in native words,

but is no longer a phonological rule.

Consonant clusters do not necessarily follow the sonority hierarchy. Sonorant consonants not adjacent to the nucleus, sometimes called *trapped sonorants*, behave anomalously in the face of voicing assimilation: they are invisible to adjacent consonants and may or may not be subject to devoicing. Trapped sonorants, regardless of how they are realized[1], never take stress or otherwise affect stress placement. [Pawelec, 2012]

Many alternations that originate from vowel length or palatalization events, but no longer have a phonological motivation, remain productive in inflection and partially in derivation.

3.4 Prosody

Polish prosody is syllable-timed: stressed and unstressed syllables have similar lengths.

Stress is regularly on the second-last syllable of each morphological word, but stress may be seen one syllable earlier or later in loanwords, and one syllable earlier in compound numerals [Swan, 2002]. Such anomalous stress is characteristic of higher register.

(3.19) <gramatyka> /graˈmatɨka/ ”grammar”

(3.20) <winegret> /vinɛˈgrɛt/ ”vinaigrette”

(3.21) <pięciuset> /ˈpʲɛɲt͡ɕusɛt/ ”500.GEN/DAT/LOC”

Stress involving clitics may appear unstable; this is because prepositions and auxiliaries, normally described as clitics, are in certain cases affixes. For prepositions, this occurs when their direct argument is a pronoun, or one of a closed set of nouns [Swan, 2002]. The clitic and affix forms of the past auxiliary are in seeming free variation, with the affix forms slowly becoming more common [Abramowicz, 2008]. The verb negator <nie> /ɲɛ/ is also an affix, despite being written separately [Swan, 2002].

(3.22) <widzieliście> /viˈd͡ʑɛli ɕt͡ɕɛ/ ~ /vid͡ʑɛˈliɕt͡ɕɛ/ ”saw-VIR.PL-2PL”

(3.23) <koło nas> /kɔˈwonas/ ~ /ˈkɔwo nas/ ”around us”

(3.24) <nie wie> /ˈɲɛvʲɛ/ ”does not know”

[1] There is a wide variety. In slow, careful speech, trapped sonorants may be realized (almost) syllabically, but not perceived as such. In fast speech, they tend to assimilation and reduction.

3.5 Changes in progress

Polish is known as the only codified Slavic language to retain phonemic nasal vowels [Długosz-Kurczabowa and Dubisz, 2006]. This status is, however, uncertain in the current spoken language. Orthographic <ę>, sometimes analyzed as /ɛ̃/, no longer contrasts with /ɛ/ alone, or with /ɛ/ followed by a nasal consonant, in any position. Orthographic <ą>, analyzed as /ɔ̃/, is still distinct from <o> /ɔ/ word-finally in some speakers, but has merged with /ɔ/, or /ɔ/ plus a consonant homorganic to the following onset, in other positions[2].

Additionally, and more related to the aim of this paper, the above merger of <e> with <ę> as well as the intake of loanwords have resulted in a loosening of the phonotactics regarding palatalized consonants and the following vowels. Namely, /k/ and /g/, which were historically always palatalized before /ɛ/, may now appear in both forms [Długosz-Kurczabowa and Dubisz, 2006].

(3.27) <ankieta> /ankʲɛta/ ”survey”, borrowed before this change

(3.28) <keton> /kɛtɔn/ ”ketone”, borrowed after it

Moreover, as hinted above, unpalatalized /t d s z/, as well as /t͡s d͡z t͡ʂ d͡ʐ ʂ ʐ/, which do not have palatalized forms, have started to appear before /i/, which historically has only appeared after palatalized consonants. The latter two series of sibilants may be described as carrying a weak form of palatalization in this environment, but they remain distinct from the native alveolo-palatal series [Rochon and Pompino-Marschall, 1999].

[2] As shown indirectly, for example, by the applicability of ablaut ([ɔ] > [a], historically restricted to /ɔ/) to historical /ɔ̃/ when deriving secondary imperfective verbs [Doroszewski, 1960].

(3.25) <ochronić> /ɔxrɔɲit͡ɕ/ ”protect.PFV”, <ochraniać> /ɔxraɲat͡ɕ/ ”protect.ITER”

(3.26) <dołączyć> /dɔwɔnt͡ʂɨt͡ɕ/ ”join.PFV”, <dołączać> often /dɔwant͡ʂat͡ɕ/ ”join.ITER”

4 eSpeak-NG TTS system

The work shown in this paper is based on eSpeak NG, which is a rule-based text-to-speech system originally developed by Jonathan Duddington [Duddington, 2010] and currently maintained by Reece H. Dunn [Dunn et al., 2022].

eSpeak NG transforms written language into a phonemic representation using rules of two types: ”rules”, which translate letters or strings of letters within tokenized words into phonemes, possibly under limited conditions, and ”lists”, which for individual words either set some modifying attributes (such as stress placement) or override the entire pronunciation with a custom string of phonemes. A third set of rules defines how the phonemes are realized. This is done using the additive strategy: vowels and semivowels are realized using a formant synthesizer, while others are realized using recordings of human speech, which are played in addition to synthesized formants if voiced, or alone if unvoiced.

4.1 Structure of language definitions

A language definition consists of the following files:

tr_languages.c
: This single source file hard-codes certain parameters for each language, such as alphabet range, structure of numerals, default stress and assimilation.

ph_*
: Definitions of phonemes used in the language, if they are not defined by default or if the language's definition differs from the default.

  A phoneme definition (`phoneme ... endphoneme`) may include:

  - formant frequencies and transitions between them (`FMT(file)`, where `file` is a binary file generated by `espeakedit`)
  - length (`length len`, where `len` is the length in ms)
  - recorded sounds (`WAV(file, amp)`, where `file` is a WAV file and `amp` is op-

tionally the amplitude as a percentage)

- effects on neighbouring vowels (`Vowelin` and `Vowelout`, with various formant-specific parameters)
- IPA representation (`ipa`)
- links to voiced or unvoiced counterparts (`voicingswitch`).

All of these attributes may be conditioned (`IF ... THEN ... ENDIF`) on the phoneme's immediate environment, such as neighbouring phonemes (`prevPh(...)`, `nextPh(...)`, position within the word (`isWordStart`, `isWordEnd`), and stress position (`isStressed`, `isAfterStress`).

*_rules

General grapheme-to-phoneme translation rules. Each rule consists of: any number of previous graphemes, graphemes to translate, any number of following graphemes, and the corresponding phonemes. For example,

```
s (i S;
_) c (an k
```

means that <s> followed by <i> is pronounced /ɕ/ (which eSpeak represents as "S;"), and that word-initial <c> followed by <an> is pronounced /k/.

*_list

Dictionary of words with overridden pronunciations. Each rule consists of the word to override, and either or both of a phoneme string and a set of attributes (e.g. stress override, treatment as part of previous or following phonological word). For example,

```
amsterdam $1
ctrl k'OntROl
nie $u $combine
```

assigns initial stress to the word <amsterdam> and the pronunciation /ˈkɔntrəl/ to <ctrl>, and marks <nie> to be treated as a part of certain following words.[1]

This file also defines the pronunciation of numerals using a special syntax in the left-most column (`_` for units, `X` for tens, `C` for hundreds, `M`*n* for multiples of the *n*th power of 1000).

*_emoji

Dictionary of symbol pronunciations. Each rule consists of a symbol followed by its name written in the language's orthography.

[1] The behaviour of `$combine` and other flags is hard-coded in `tr_languages.c`.

4.2 State of Polish support

The support for Polish in eSpeak NG, developed c. 2007-2008 by an anonymous contributor, is subjectively solid. Words spelled in Polish orthography are converted into phonemes nearly flawlessly. Stress determination is similarly excellent, even where the orthography does not reflect word boundaries. There are, however, points where the synthesized speech does not match the expected pronunciation.

4.2.1 Inflection of numerals

eSpeak NG does not, as a rule, inflect numerals. This leads to incorrect, sometimes poorly understandable pronunciations. For example, the date

(4.1) *17.* *sierpnia*
17th-GEN August-GEN
17 August

is incorrectly expanded to <siedemnaście sierpnia>, as if the numeral were cardinal, whereas it is ordinal and should usually be expanded to <siedemnastego sierpnia>.

However, expansion of numerals is complicated not only by the distinction between ordinal and cardinal numbers, but also by case and gender inflection: in <przed 17. sierpnia> "before 17 August", the numeral, spelled identically as above, must be inflected for the instrumental case:

(4.2) *przed* *siedemnastym* *sierpnia*
before seventeenth-INS August-GEN
before 17 August

Agreement resolution may require examination not only of the surrounding words, but also further ones.

(4.3) *24* *pod* *szafą* *schowanych* *rachunków*
24-GEN under cupboard-INS hidden-GEN.PL bills-GEN
of 24 bills hidden under the cupboard

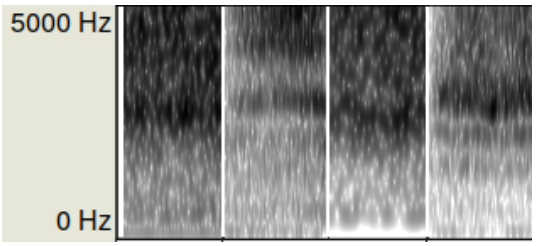
In this phrase, <24> agrees with <schowanych> and <rachunków>, but not with the intervening two words. In fact, a general solution to the numeral inflection problem requires a full parse of the sentence, since if <24> in the previous example were replaced with <25>, the case form would depend on lexical attributes of the head of the enclosing phrase, and in the case of a verb phrase with a verb of arity >1, on the verb's other arguments.

eSpeak NG has plans to, in the long term, allow for rules expressive enough to support numeral inflection. Such support would, however, require a nearly complete rewrite of the grapheme-to-phoneme translation mechanism, and as such is out of scope for this paper.

4.2.2 Distinguishability of sibilants

The Polish phoneme definitions in eSpeak NG are nearly complete: all vowels are distinguished, as are all consonants of the minimal analysis. Most of the definitions are copies or slight adjustments of existing ones that were developed with other languages in mind. As languages with a distinction between retroflex and alveolo-palatal sibilants are typologically uncommon, there is little pressure to ensure distinguishability of the two series, and the realizations of them produced by eSpeak NG are in fact very similar.

This contrasts with how these phonemes are realized by a Polish speaker. While the eSpeak realization of /ɕ/ is acceptably close to the native one, /ʂ/ is visibly different and is more distinct from /ɕ/ in the native realization. Both phonemes share formants around 3000, 4100 and 4500 Hz. Unlike /ɕ/ and eSpeak /ʂ/, native /ʂ/ has additional formants around 1500 and 2200 Hz and shows smaller amplitude of the 4500 Hz formant.

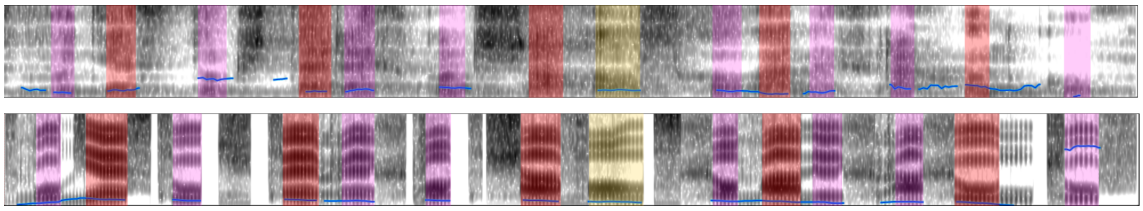


Praat spectrograms, from left to right: eSpeak /ɕ/, human /ɕ/, eSpeak /ʂ/, human /ʂ/.[2]

4.2.3 Prosody

Although eSpeak NG assigns stress to Polish utterances almost perfectly, the spoken output diverges in terms of prosody from human speech. Subjectively, the rhythm of eSpeak's output sounds more stress- than syllable-timed. This is to be expected, as the Polish voice file does not override the default stress lengths, which were developed for English.

[2]Human recordings spoken by the author.



Spectrograms of human (top) and eSpeak (bottom) realizations of a short Polish sentence with eight unstressed (pink) and five stressed (red) oral[3] vowels.

As seen in this short example, eSpeak realizes stressed vowels about 42% longer than unstressed vowels (stressed[4]: 110 ± 10.0 ms, unstressed: 77.1 ± 11.5 ms, whereas in natural speech, this value is closer to 26% (stressed: 80.4 ± 13.2 ms, unstressed: 64.0 ± 9.13 ms). Additionally, vowels in the human recording (70.3 ± 13.2 ms) are generally only about 78% as long as those produced by eSpeak (89.6 ± 19.5 ms).

The eSpeak realization, as shown by the white bands on the spectrogram, contains silences of up to 45 ms next to unvoiced consonants. These silences have the subjective effect of clearer enunciation; removing them manually results in the sentence sounding rushed.

[3]The sentence contains one unstressed nasal vowel (yellow), which is excluded from consideration since the length difference between oral and nasal vowels surpasses that between unstressed and stressed vowels.

[4]Vowel length varies in eSpeak output depending not only on stress, but also on surrounding consonants and position within the sentence.

5 Task: improve distinctions between phonemes in eSpeak output

Based on the state of eSpeak NG's Polish voice described in the previous chapter, and subject to the restrictions of the codebase itself and the limited time available for the project, I chose to first improve the distinguishability of phonemes that are currently pronounced similarly by adding and modifying phoneme definitions and possibly also grapheme-to-phoneme rules.

5.1 Main task

The main distinctions chosen for improvement are between the three series of alveolo-dental, alveolo-palatal and retroflex sibilants, namely:

t͡s	d͡z	s	z
t͡ɕ	d͡ʑ	ɕ	ʑ
t͡ʂ	d͡ʐ	ʂ	ʐ

Besides improving comprehensibility of eSpeak's output, bringing the realizations of these phonemes closer to the natural ones has the potential to reduce the subjective perception of the output as "foreign-accented", as this distinction, being typologically uncommon, serves as a sort of shibboleth for distinguishing native and non-native speakers of Polish.

The improved phoneme definitions may additionally find use in other languages with similar distinctions, such as Russian and Mandarin Chinese.

5.2 Additional tasks if time permits

If time permits, further aspects of the voice will be modified to sound more natural, including phonemes (e.g. /ɛ ɲ t/) and prosody.

5.3 Possible regressions and mitigation

Due to the phenomenon of secondary palatalization, described in chapter 2, adjusting the current realizations of the alveolo-dental and retroflex sibilants may result in a less natural realization in some loanwords. To mitigate this possible regression, these phonemes may need to be split and the conversion rules and dictionary adjusted to properly select the correct variation.

6 Implementation

6.1 Setup of eSpeak NG

The eSpeak NG source code is available as a Git repository hosted on GitHub.

The project is set up with GNU autotools, which in most cases simplifies the build process to the two commands, `./configure` and `make`. The project is portable; apart from a C standard library, it only requires one of the audio frameworks supported by its component `pcaudiolib`. On Unix-like systems, this may be ALSA, CoreAudio, OSS or PulseAudio.

Once built, the binaries may be either installed (`make install`) or run locally by setting the environment variables `ESPEAK_DATA_PATH` and `LD_LIBRARY_PATH` to subdirectories of the project directory. To simplify comparison between upstream and modified versions of eSpeak, the modified version was run locally and the upstream version was kept installed.

Further notes on setup can be found in Appendix A.

6.2 Definition of basic consonant phonemes

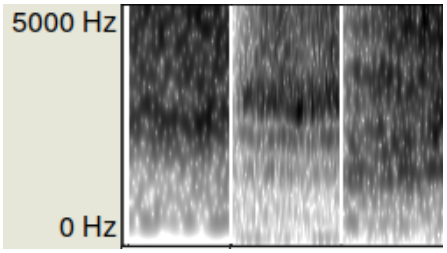
Of the sibilants mentioned above, the `ph_polish` file in eSpeak NG mentions only /t͡s d͡z t͡ɕ d͡ʑ/, and makes only minimal modifications to these so that the voicing assimilation logic treats them properly. All other sibilants are taken directly from the master phoneme file, `phonemes`.

In a first step, to maximize the perceptual distance between the three series of sibilants, definitions were added for the following phonemes, using audio recorded for other languages supported by eSpeak:

- /ʂ/: using the recording `sh3.wav`
- /ʐ/: using the recording `sh3.wav` and modified vowel transition
- /t͡ʂ/: using the recording `tsh2.wav`

- /d͡ʐ/: using the recording `tsh2.wav` and modified vowel transition

The above mentioned .wav files are already part of the eSpeak distribution and are used for other languages, namely Croatian and Swahili. `sh3.wav` and `tsh2.wav` match the retroflex articulation used in Polish more closely than the sounds used by default.



Praat spectrograms, from left to right: eSpeak /ʂ/ existing, human /ʂ/[1], eSpeak /ʂ/ after change.

As can be seen in the spectrograms, `sh3.wav` is closer to the human-produced /ʂ/; most obviously, it shows a formant at 1500 Hz frequency, which `sh.wav` lacks.

In addition, for the voiced phonemes, the formant transformations were modified so that the boundaries with neighbouring vowels continue to sound acceptable.

- /t͡s/ and /d͡z/: using `ts2.wav`

This recording shows a shorter release than the default `ts.wav` used medially: 0.028 s compared to 0.073 s. The shorter release allows it to be distinguished more easily from the phonemically distinct stop-fricative sequences /ts dz/. Unlike the default phoneme definition of /t͡s/, the one used here does not distinguish final and non-final realizations; subjectively, they are not necessary.

6.2.1 Allophones before /i/

Taking the secondary palatalization, described in the previous chapter, into account, the above phonemes were modified with a `nextPh` condition so that they are realized with their original recordings when they precede /i/. In this way, a worsening of the output in these environments was avoided.

[1] Spoken by the author.

6.3 Tweaking of existing phonemes

As a simple way to further improve the naturalness of eSpeak's output, a handful of phonemes whose current treatment is not critical to comprehensibility were tweaked to use recordings closer to their usual Polish realizations.

- /t/: `t_dnt2.wav` is used, unless the /t/ is in an onset and the following consonant is one of /m n r/, in which cases the existing `t_dnt.wav` is used. The latter recording contains a slight aspiration after the release. /t/ may occur aspirated in any position, especially when the syllable containing it is contrastively stressed, but aspiration is nearly universal in the above mentioned environment and relatively rare outside it.
- /ɛ/: Unstressed /ɛ/ word-finally, as well as after palatalized consonants, is realized by eSpeak more centrally than is usual in Polish. This is corrected by removing the `nextPh(isPause)` condition in phoneme `E` as well as replacing the formants used by phoneme `E#` by those used in that same removed condition.
- /ɲ/: By default, this phoneme is realized with a duration of about 150 ms. Although this fits natural speech in Spanish, for which it was first defined in eSpeak, it is about twice as long as the Polish realization. This is corrected by importing the default definition and setting the length to 75 ms.

6.4 Tweaking of voice file

Currently, eSpeak NG's voice file for Polish is barren. All settings apart from the intonation modifier are taken from the defaults, which were originally set with English in mind. eSpeak's concept of "voice" includes parameters such as pitch range, segment length, effects of stress, effects of word boundaries, and intonation variants.

Polish differs from English in a few of these aspects. For instance, to generalize, Polish prosody is syllable-timed, while English is stress-timed. This discrepancy manifests (subjectively) as the system adding unnatural length to stressed syllables. To fix this, a `stressLength` line is added with smaller differences between unstressed, stressed and contrastively stressed syllables.

Moreover Polish exhibits voicing assimilation across word boundaries. The Polish `rules` implement this assimilation nearly perfectly, but the effect in the spoken output sounds unnatural due to the short pause that is added by default between words. This is fixed by adding `words 0 1`, which suppresses the pauses except when they lie between two vowels.

7 Evaluation

To evaluate the effectiveness of the changes made to eSpeak NG in increasing the comprehensibility of its output, an evaluation corpus was created and given to a few Polish speakers to annotate.

7.1 Evaluation design

7.1.1 Variables

The experiment was designed to measure differences between the original and modified versions of eSpeak in the following variables:

- Comprehensibility
- Naturalness

for the following kinds of utterances:

- No modified phonemes present (only prosody differs)
- With modified sibilants
- With modified alveolo-palatal nasal (including before front mid vowel)
- With modified front mid vowel

7.1.2 Corpus

The evaluation corpus consists of a mixture of prose and verse, from works that are out of copyright or available on permissive terms, as well as individual dictionary words. The corpus was processed by unmodified and modified versions of eSpeak NG to produce two corresponding WAV files of each sentence, line or word. The following sources were used to create the corpus:

- 17 sentences from out-of-copyright literary works [Mickiewicz, 1828; Sienkiewicz, 1896]
- 47 sentences from Wikipedia articles [Wikipedia]
- 120 content words from HunSpell dictionary [Miłkowski, 2008]

Each stimulus contained either no modified phonemes, or modified phonemes of only one of the three classes. For stimuli with modified phonemes consisting of more than a single word, care was taken to also include at least one unmodified phoneme from the same class as the modified phoneme(s).

The full corpus can be found in Appendices C and D.

7.1.3 Software

Various software was used to perform the evaluation, some written specifically for this project.

- Words and sentences were selected manually. They were shuffled and assigned to formats using the basic Unix utilities, `shuf`, `nl`, `sort` and `sed`.
- Stimuli were presented using a simple hand-written webpage that took a comma-separated list of stimuli as input, showed the appropriate sequence of forms and saved annotator input in JSON format.
- Words typed by annotators were scored using the Levenshtein Python module [Ohtamaa et al., 2022].
- Statistics were calculated using Gnumeric [de Icaza et al., 2022].

7.1.4 Experimental design

Recordings in the corpus were presented in one of three formats. They were randomized and assigned formats once and given in the same order, both among and within tasks, to all annotators. For ease of analysis, all formats were coded to produce the same range of values, $[0, 4]$, for both measures.

Single-word format

This format was used for all recordings of single words as well as lexical phrases of up to four words. Annotators were presented with a single recording from either original or modified version of eSpeak. They were asked to type the word spoken and to rate how natural the word sounds on a five-point scale. No feedback was given as to whether the input was correct.

For comprehensibility, correctly typed words were coded as 4; for incorrectly typed words[1], the result was $\max\left(0, 4 - \frac{\text{ed}}{\text{len}}\right)$, where ed is the character-based Levenshtein distance to the correct word and len is the number of characters in the correct word. The naturalness rating was coded on a scale from 0 to 4.

Rating format

This format was used for half of the sentence recordings. Annotators were presented with a single recording from either original or modified version of eSpeak. They were asked to rate the sentence separately with respect to comprehensibility and naturalness, each on five-point scales, coded from 0 to 4.

Comparison format

This format was used for the other half of the sentence recordings. Annotators were presented with two recordings of the same sentence, one from the original and one from the modified version of eSpeak, in either order. They were asked to select the more comprehensible recording as well as the more natural-sounding recording. The selected recording was coded as 4 and the non-selected as 0.

In total, three annotators were each presented with 336 stimuli (of which 32 contained two recordings). Each was presented once automatically; annotators could press a button to hear the stimulus one additional time. Annotators were told to judge the stimuli as quickly as possible. To avoid fatigue affecting the results, annotators were asked to take a break after every 85 completed tasks.

The forms shown to the annotators can be found in Appendix B.

[1] Input was first reviewed manually; obvious typing errors without any possible phonetic motivation (letters adjacent on the keyboard to the correct letter) were corrected before calculation. After manual review, to avoid noise from homophones, the words were normalized so that each phoneme is represented by only one grapheme in a given environment.

7.2 Results

As seen in the table below, the changes taken together brought increases in both comprehensibility and naturalness.

Subset of corpus	Comprehensibility	Naturalness
Whole corpus	+0.21 ($p < 0.007$)	+0.23 ($p < 0.004$)
No modified phonemes	No difference	No difference
Modified sibilants	No difference	+0.43 ($p < 0.008$)
Modified palatal nasal	+0.52 ($p < 0.0002$)	+0.54 ($p < 0.0001$)
Modified front mid vowel	No difference	No difference

Mean difference in comprehensibility and naturalness between original and modified eSpeak NG, on a scale of 0 to 4, grouped by phoneme class. p-values calculated using two-sided Student's t-test, significance threshold $p \leq 0.05$.

Raw annotations for each stimulus and annotator can be found in the file `exp.log`, and "cooked" annotations after post-processing of typed words in `absolute.csv`, both located in the archive `scripts.tar.gz`.

7.2.1 Effect of changes to prosody

In an effort to increase naturalness, prosody parameters (stress length modifiers and word pause length) were modified. Stimuli containing only these changes were included to provide a baseline for evaluating the other, possibly comprehensibility-affecting, changes. No significant effect of the changes to prosody on comprehensibility or naturalness can be found.

7.2.2 Effect of changes to sibilants

The main focus of this project was to improve the comprehensibility of synthesized speech containing the sibilants, and for this purpose, phoneme definitions were created or modified for /t͡s d͡z t͡ʂ d͡ʐ ʂ ʐ/.

Naturalness of stimuli containing these phonemes was improved (0.43 increase in score, $p < 0.008$), but contrary to expectations, while comprehensibility scores increased (+0.22), this increase was not significant ($p = 0.18$). It may be useful to examine this discrepancy further by increasing sample size or by specifically choosing members of minimal pairs as stimuli.

7.2.3 Effect of changes to palatal nasal

The change of the length of /ɲ/, being the simplest of all modifications applied, and done only as an afterthought meant to slightly improve naturalness, surprisingly showed the highest improvements in both comprehensibility (+0.52, $p < 0.0002$) and naturalness (+0.54, $p < 0.0001$) scores.

7.2.4 Effect of changes to mid front vowel

No significant effect of the changes to /ɛ/ alone were found. It bears mentioning that stimuli with /ɛ/ following /ɲ/ were lumped with those containing /ɲ/ only for this analysis. To disentangle the effects of these modifications, a further study could be performed using multiple versions of eSpeak containing only one modification each.

8 Conclusion

Progress in high-quality text-to-speech is occurring mainly in systems that replace human-defined rules with machine learning. At the same time, traditional rule-based text-to-speech systems continue to fill niches that more modern systems cannot, such as environments with limited storage or processing power, or where speech must be synthesized at high speeds. A notable example of such a niche is screen readers for visually impaired people, which run parallel to other processes on a personal computer and, due to the sheer amount of information available in user interfaces designed for visual use, are usually set to speak several times faster than natural speech.

This paper has examined eSpeak NG, a minimalistic, additive speech synthesis system, and shown based on eSpeak NG's support for the Polish language that the naturalness and comprehensibility of such systems can be improved with relatively minor changes that have negligible storage and runtime impact, especially when it comes to languages with relatively transparent orthographies.

Further directions regarding this topic could include an examination of other languages supported by eSpeak NG for similar low-hanging fruit, as well as additional studies to determine why some of the changes made in this project did not have the predicted effects.

References

Ł. Abramowicz. Using sociolinguistic data to illuminate a theoretical debate: The case of person/number marking in Polish. *University of Pennsylvania Working Papers in Linguistics*, 14(1), 2008. URL https://repository.upenn.edu/cgi/viewcontent.cgi?article=1017&context=pwpl.

M. de Icaza, J. Goldberg, H. Ashburner, S. Atkinson, M. Berkelaar, J. Brefort, and others. Gnumeric – the GNOME spreadsheet, 2022. URL https://gitlab.gnome.org/GNOME/gnumeric/.

W. Doroszewski. Objaśnienia wyrazów i zwrotów. *Poradnik językowy*, 10:472, 1960.

J. Duddington. eSpeak text to speech. 2010. URL http://espeak.sourceforge.net/.

R. H. Dunn, V. Vitolins, J. Hiltunen, A. H. Ngyuen, et al. eSpeak NG text-to-speech, 2022. URL https://github.com/espeak-ng/espeak-ng/blob/master/docs/index.md.

K. Długosz-Kurczabowa and S. Dubisz. *Gramatyka historyczna języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego, 2006.

W. Jassem. A phonologic and acoustic classification of Polish vowels. *STUF – Language Typology and Universals*, 11(1-4):299–319, 1958.

Y. Leviathan and Y. Matias. Google Duplex: an AI system for accomplishing real-world tasks over the phone. 2018.

A. Mickiewicz. *Konrad Wallenrod*. 1828. URL https://pl.wikisource.org/wiki/Konrad_Wallenrod.

M. Miłkowski. Polish dictionary for HunSpell, 2008. URL https://extensions.openoffice.org/en/download/1716.

Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19), 2019. URL https://www.mdpi.com/2076-3417/9/19/4050/pdf.

M. Ohtamaa, A. Haapala, and M. Bachmann. Levenshtein Python C extension module, 2022. URL https://github.com/maxbachmann/Levenshtein.

P. Pawelec. The sonority sequencing generalization and the structure of consonant clusters with trapped sonorants in Polish. *Anglica Wratislaviensia*, 50, 2012. URL https://wuwr.pl/awr/article/download/155/134.

M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il, and N. Mastorakis. An overview of text-to-speech synthesis techniques. *Latest trends on communications and information technology*, 2010. doi: 10.1.1.669.5477.

M. Rochon and B. Pompino-Marschall. The articulation of secondarily palatalized coronals in Polish. 1999.

H. Sienkiewicz. *Quo vadis: Powieść z czasów Nerona*. 1896. URL https://pl.wikisource.org/wiki/Quo_vadis.

G. F. Simons and C. D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, 2017. URL https://www.ethnologue.com.

O. E. Swan. *A grammar of contemporary Polish*. Slavica Publishers, 2002.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2000. URL http://hts.sp.nitech.ac.jp/?plugin=attach&refer=Publications&openfile=tokuda_icassp2000.pdf.

J. P. Van Santen. Phonetic knowledge in text-to-speech synthesis. In *The integration of phonetic knowledge in speech technology*, pages 149–166. Springer, 2005. doi: 10.1007/1-4020-2637-4_9.

Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous. Tacotron: A fully end-to-end text-to-speech synthesis model. *Computing Research Repository*, 2017. URL http://arxiv.org/abs/1703.10135.

R. J. Weiss, R. J. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma. Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis. *Computing Research Repository*, 2020. URL https://arxiv.org/abs/2011.03568.

J. C. Wells. Phonetic transcription and analysis. In *Encyclopedia of Language and Linguistics*, pages 386–396. Elsevier, 2006.

Wikipedia. Wikipedia, wolna encyklopedia. URL https://pl.wikipedia.org/. Individual sentences from various articles.