



university of  
 groningen

campus fryslân

# From Tolkien's Novel to Synthetic Speech: Developing TTS Systems for Quenya

Author:  
Wangyiyao Zhou

Supervisors:  
Ph.D. Candidate Phat Do

MASTER THESIS

A thesis submitted in fulfillment of the requirements  
for the degree of Master of Science  
in the

VOICE TECHNOLOGY  
CAMPUS FRYSLÂN

June 10, 2024

## Abstract

Text-to-speech (TTS) technology has been successfully implemented in various applications, serving as a means to preserve endangered languages and cultures. However, TTS systems for constructed languages have not been extensively studied. J.R.R. Tolkien created Quenya for the elves in his novels, inspired by the phonetic patterns and structures of Finnish. While enthusiasts have developed courses based on Tolkien's materials, attempts to synthesize Quenya speech remain limited. This study uses the articulatory features as inputs for speech synthesis and evaluates the outcomes of applying transfer learning from models based on more resourced languages. Using the IMS-Toucan system from the University of Stuttgart, based on the FastSpeech 2 architecture, this research developed a TTS system for Quenya by fine-tuning three models with a 34-minute Quenya dataset: one pre-trained on Finnish, one pre-trained on English, and a multilingual model. The results showed that the Finnish fine-tuned model produced better speech than the English model, while the multilingual model produced the most natural and accurate speech. This study provides insights for developing TTS systems for other constructed and ancient languages requiring phonetic reconstruction.

## Acknowledgements

I would like to express my deepest gratitude to my supervisor, Ph.D. Candidate Phat Do, for his invaluable guidance throughout this project. I am also grateful to Associate Professor Dr. Matt Coler for his support and insights. Additionally, I would like to thank all the teachers in the Voice Technology program for their advice and encouragement. I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

My thanks also go to the original creators of the IMS-Toucan system for providing such a convenient tool. Special thanks to the authors of *Atanquesta* and *Glaemscrafu*; this project would not have been possible without your open-source recordings. I also extend my gratitude to all the participants in the questionnaire survey and the Quenya enthusiasts who were willing to write detailed listening reports. Finally, I am deeply grateful to the volunteer speaker, Igor Marchenko. Despite your busy schedule leading up to graduation, you provided exceptionally high-quality recordings. You are the best classmate I have ever had.

# Contents

<b>Abstract</b>	<b>1</b>
<b>Acknowledgements</b>	<b>2</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Research Question and Hypothesis . . . . .	6
1.2 Research Contributions . . . . .	6
1.3 Thesis Outline . . . . .	7
<b>2 Background</b>	<b>8</b>
2.1 Phonetics and Phonology . . . . .	8
2.1.1 Phonemes and Allophones . . . . .	8
2.1.2 International Phonetic Alphabet (IPA) . . . . .	9
2.1.3 Phonology of Quenya . . . . .	10
2.2 Existing Quenya Language Corpora . . . . .	12
2.3 Non-Autoregressive TTS Technologies . . . . .	13
2.3.1 Evolution from Concatenative to Autoregressive Models . . . . .	14
2.3.2 Fastspeech . . . . .	15
2.3.3 Fastspeech 2 . . . . .	17
2.3.4 Vocoder . . . . .	18
2.3.5 IMS-Toucan . . . . .	19
<b>3 Related Study</b>	<b>22</b>
3.1 Constructed languages . . . . .	22
3.2 Low-Resource Language TTS . . . . .	24
<b>4 Methodology</b>	<b>26</b>
4.1 Quenya Dataset and Grapheme-to-Phoneme (G2P) Script . . . . .	26
4.1.1 Building Quenya Dataset . . . . .	26
4.1.2 Building Quenya G2P . . . . .	26
4.2 Training the Finnish and English TTS Models . . . . .	27
4.2.1 Finnish Dataset . . . . .	27
4.2.2 English Dataset . . . . .	27
4.2.3 Model Training and Evaluation Results . . . . .	27
4.3 Fine-Tuning Models with Quenya Data . . . . .	28
4.4 Evaluation . . . . .	28

<b>5</b>	<b>Result</b>	<b>29</b>
5.1	Mean Opinion Scores (MOS) . . . . .	29
5.2	Listening Reports . . . . .	30
<b>6</b>	<b>Discussion</b>	<b>32</b>
6.1	Challenges with the Finnish Model . . . . .	32
6.2	Limitations . . . . .	33
6.2.1	Prosody and Pronunciation Challenges . . . . .	33
6.2.2	Dataset Quality and Training Parameters . . . . .	33
6.3	Future Research . . . . .	33
<b>7</b>	<b>Conclusion</b>	<b>35</b>
	<b>References</b>	<b>36</b>
	<b>Appendices</b>	<b>42</b>

# 1 Introduction

Text-to-speech (TTS) technology has changed how we interact with digital devices and media, enabling machines to convert written text into spoken words. In the field of language learning, this technology provides immediate voice output, greatly enhancing interactivity and accessibility for learners, especially for those with visual impairments. Additionally, TTS technology is crucial for preserving endangered languages, offering digital support that helps safeguard cultural heritage. Although TTS technology has achieved significant success with widely used languages, matching human-like quality in some instances, the focus on low-resource languages has also been increasing in recent years. However, the development in constructed languages still shows insufficient progress.

A constructed language is deliberately crafted by an individual or group, rather than emerging from natural evolution over time (Adams, 2011; Okrent, 2009). These languages can be categorized as planned languages (Gobbo, 2017; Janton, 1993; Tonkin, 2015), fictional languages (Barnes & Van Heerden, 2006; Kazimierczak, 2010; Schreyer, 2021b), or artificial languages (Schreyer, 2021a). Constructed languages like Esperanto are designed as international auxiliary languages (auxlangs) to facilitate communication (Forster, 1982), while others like Klingon in "Star Trek", Dothraki in "Game of Thrones" or Quenya in Tolkien's writings serve artistic purposes. J.R.R. Tolkien was a pioneer in popularizing language creation, developing historical language families, a technique still used by many language creators. Tolkien focused on the art and aesthetics of his languages, deeply influenced by Finnish grammar, which made his languages "heavily Finnicized in phonetic pattern and structure" (Carpenter et al., 1981). Although Quenya distanced itself somewhat from Finnish over time, the influence never completely disappeared (Perälä, 2002). Tolkien reduced his borrowing of Finnish words, but the phonetic and structural influence of Finnish on Quenya deepened (Tikka, 2007).

In building TTS systems, constructed languages and low-resource languages face similar challenges. Constructed languages are often considered endangered because they are rarely learned as a first language at home, have few speakers, and lack official status and prestige (Schreyer, 2011). Both types of languages are usually learned voluntarily later in life (Christoph, 2012). Various techniques have been proposed to address the low-resource TTS problem, with transfer learning being one of the commonly used methods (Weiss, Khoshgoftaar, & Wang, 2016). This method involves pre-training the acoustic model in a different language with sufficient training data (the source language) and then fine-tuning the model with the limited data available for the target low-resource language (Tan, Qin, Soong, & Liu, 2021). However, cross-lingual transfer learning presents challenges, primarily due to mismatches between the

input embeddings of the source and target languages caused by differences in phoneme sets or orthographic characters. To address this, [Lux and Vu \(2022\)](#) fixed previous shortcomings by using a linguistically motivated representation of the inputs to such a system (articulatory and phonological features of phonemes) that enables cross-lingual knowledge sharing and applying the model-agnostic meta-learning framework to the field of low-resource TTS for the first time. [Do, Coler, Dijkstra, and Klabbers \(2023\)](#) investigated the effectiveness of phone labels versus articulatory features for cross-lingual transfer learning in TTS applications for low-resource languages. Currently, there has been limited research on TTS for constructed languages, with only a few examples available. For instance, Esperanto TTS systems include *Parol*<sup>1</sup> and *EsperantoTTS*<sup>2</sup>, and [Jokisch and Eichner \(2000\)](#) developed a TTS system for Klingon. However, it seems that no TTS system has been developed for Quenya. Therefore, this study aims to fill that gap by utilizing the ToucanTTS<sup>3</sup> system ([Lux et al., 2023](#)), which supports the use of phonological features extracted from the IPA transcription of audio data to enhance acoustic feature mapping. The study will fine-tune TTS models pre-trained on English, Finnish, and multilingual datasets using a Quenya dataset to determine which performs best.

## 1.1 Research Question and Hypothesis

**Research Question:** Will a Quenya TTS system benefit more from transfer learning using a Finnish language model compared to an English language model in terms of performance?

**Sub-Questions:** Is there an advantage in using multilingual models over monolingual models in enhancing the quality of synthetic speech for such a system?

**Hypothesis:** Based on the previous discussion ([Christoph, 2012](#); [Do et al., 2023](#); [Tan et al., 2021](#)), this study hypothesizes that a model trained through transfer learning from Finnish will perform better than one trained from English. Furthermore, a multilingual model is anticipated to outperform both the Finnish and English models, achieving the best overall performance in synthesizing Quenya speech.

## 1.2 Research Contributions

Studying contact and constructed languages is important for understanding the full range of human linguistic possibilities ([N. H. Lee, 2020](#); [Schreyer, 2021b](#)). This research aims to support educational purposes by providing richer auditory reference materials for enthusiasts learning

---

<sup>1</sup><https://parol.martinrue.com/>

<sup>2</sup><https://54696d21.github.io/esperantoTTS/>

<sup>3</sup><https://github.com/DigitalPhonetics/IMS-Toucan>

Quenya. Currently, learners rely on limited printed materials and recordings from other enthusiasts, which often do not meet their needs. Furthermore, developing speech synthesis technology for constructed languages is crucial for preserving linguistic diversity, especially for languages with extremely limited or non-existent resources. The experience gained from developing TTS for constructed languages can also be applied to extinct and ancient languages that require speech reconstruction (Ivnova, 2023). By using small datasets read by expert linguists, the speech of these languages can be recreated, aiding in the preservation and understanding of cultural heritage.

### 1.3 Thesis Outline

The structure of this thesis is organized as follows. Section 2 provides the necessary background, including basic knowledge of Quenya phonology and the theoretical foundations of the non-autoregressive model, FastSpeech 2-based ToucanTTS. Section 3 reviews related studies, covering important research on constructed languages and low-resource language TTS. Section 4 outlines the experimental methodology, describing the processes and techniques employed. Section 5 presents the results, providing insights into the performance of the Quenya TTS systems. Section 6 discusses these findings about the research questions and explores future research directions. Finally, Section 7 concludes the thesis, summarizing the experiments and their implications for TTS technologies in low-resource languages.



## 2 Background

This section provides essential background knowledge to facilitate a better understanding of the methodologies discussed later. Section 2.1 introduces fundamental concepts in phonetics, the International Phonetic Alphabet (IPA), and the phonology of Quenya. Section 2.2 details the Quenya language corpus utilized in this study. Finally, Section 2.3 explores mainstream non-autoregressive deep learning models for speech synthesis, including FastSpeech, FastSpeech 2, and ToucanTTS, which are employed in this research.

### 2.1 Phonetics and Phonology

Phonetics is the study of human sounds, covering how sounds are produced, transmitted, and understood, including the mechanisms of speech production, pronunciation features, and modes of expression. This research is crucial for fields such as speech synthesis and speech recognition. As a fundamental component of language structure, alongside grammar, vocabulary, and text, understanding the pronunciation methods or phonetics is essential for converting text into speech that humans can comprehend.

#### 2.1.1 Phonemes and Allophones

In linguistics, a phoneme is the smallest unit of sound that can distinguish meaning in language (Bett, 2002). This means that a change in a single phoneme can alter the meaning of a word, thereby distinguishing two different words (Barlow & Gierut, 2002). For instance, in English, the difference in the phonemes [p] and [b] differentiates the words *pat* and *bat*. A phoneme is not the actual sound produced but is a theoretical concept used to describe those elements of speech that can differentiate meanings. An important aspect of phonemes is their variant pronunciations known as allophones. Allophones are context-dependent variants of a phoneme that, despite slight differences in pronunciation, are perceived as identical in a specific linguistic context. For example, the phoneme [p] in English has slightly different articulations at the beginning of the word *pin* and within the word *spin*, but these variations do not change the meaning of the words, hence are considered allophones of the phoneme [p].

Phonemes can be categorized into two broad types: vowels and consonants. Vowels are sounds produced with a free airflow in the oral cavity, such as [aɪ] in *like*, whereas consonants are sounds where the airflow is obstructed or blocked at some point in the mouth, like [k] in *cat*. Classification of phonemes also involves their point of articulation in the mouth; vowels are typically categorized based on tongue position (such as high, mid, low) and lip rounding (rounded or unrounded), while consonants are classified by place of articulation (such as labial,

dental), manner of articulation (such as fricative, stop), and voicing (voiced or voiceless).

These classifications enable linguists to accurately describe and analyze phonetic phenomena across languages. For example, the letter *c* represents different phonemes in English and Spanish: in English, it might sound as [s] or [k], while in Spanish, it varies between [θ] (as in *Barcelona*) or [k] (as in *casa*). The correspondence between phonemes and the letters that represent them has developed independently in each language. Understanding these relationships helps us grasp the complexity and diversity of languages, and how subtle differences in pronunciation can distinguish and elucidate meanings.

### 2.1.2 International Phonetic Alphabet (IPA)

The International Phonetic Alphabet (IPA) is a widely used symbol system designed to provide a consistent and accurate means of transcribing the phonetic sounds of all languages globally. Established by the International Phonetic Association in 1886, its purpose is to offer a common standard for linguists, teachers, students, and speech therapists to unambiguously record the pronunciations of different languages. The IPA is based on the Latin alphabet, supplemented by some Greek letters and other special symbols to represent specific phonetic sounds not covered by the Latin script. These symbols are designed to be as intuitive as possible, illustrating the sound production features such as the direction of airflow, articulation sites, and voicing. In consonants, the IPA categorizes sounds based on physiological mechanisms of production, such as the place of articulation (labial, dental, apical, etc.) and the manner of articulation (stop, fricative, nasal, etc.). For instance, the English [t] sound is represented in IPA as [t̪], which is a voiceless apical stop. Vowel symbols are based on tongue position (high, mid, low) and the shape of the oral cavity (front, central, back), for example, the vowel in the English word *see* is noted in IPA as [iː], a high front vowel. Stress and tone symbols play a crucial role in the IPA; they not only indicate the stress position within syllables but also significantly affect the meanings of words in some languages. For example, in English, the word *record* [ˈrɛk.ɔːrd] can be pronounced with the primary stress on the first syllable when used as a noun meaning “a document or result that can be stored and accessed,” but as *record* [rɪˈkɔːrd] with the primary stress on the second syllable when used as a verb meaning “to capture information in written or other permanent form.”

Using the IPA in speech synthesis helps accurately mimic sounds from different languages. Speech synthesis systems use IPA transcriptions to connect phonological features like voicing, place of articulation, and vowel position with sound features. This method does not depend on any specific language and uses common features of how humans make sounds. For example, features like voicing or where the sound is made in the mouth are the same in any language

when taken from IPA. This makes it possible for systems trained with these features to work well with new languages, even those not included in the initial training.

### 2.1.3 Phonology of Quenya

Tolkien described how Elves, Men, and Hobbits pronounced Elvish languages in various sources. Quenya closely resembles a natural language (Destruel, 2016) and has evolved since Tolkien first created it. The earliest version, known as Qenya, dates back to at least 1915 when Tolkien wrote the *Qenya Lexicon* (Tolkien, 1992, p. 246–248). Initially, Tolkien represented the sound [k<sup>w</sup>] with a single *q*. Therefore, despite different spellings, Qenya and Quenya are pronounced the same [k<sup>w</sup>eñä]. While writing *The Lord of the Rings*, Tolkien changed the spelling from *k* to *c* and used *qu* instead of *q*. These changes were purely aesthetic and did not alter the pronunciation. These spelling changes show that the language has been gradually evolving. Tolkien continuously updated his ideas, but he never clearly marked a split between Qenya and Quenya. Therefore, the differences in pronunciation and vocabulary between them might not always be clear. Thus, this research has adopted the pronunciation guidelines from Tamás Ferencz’s tutorial *Atanquesta*<sup>4</sup>, using it as the standard for developing the pronunciation rules in the TTS system for Quenya. Additionally, the open-source audio files from this tutorial are a crucial part of creating the dataset, as detailed in Section 2.2.

**Vowels** Quenya has five vowels, distinguished by length: the short vowels and long vowels, as shown in the table below.

Short Vowels					Long Vowels				
a	e	i	o	u	á	é	í	ó	ú
[a]	[ɛ]	[i]	[o]	[u]	[aː]	[eː]	[iː]	[oː]	[uː]

Table 1: Quenya Vowels: Short and Long Forms

For [a, i, u, o], short and long forms have identical vowel quality. However, [e] differs; the short is [ɛ], and the long is a high-mid [eː].

**Diphthongs** In diphthongs, the two vowels are in close contact with each other and the first vowel (which receives more stress) effortlessly glides into the second one.

In Quenya, unlike modern English where the final *e* in words like *home* and *mole* is silent, every vowel and diphthong is pronounced regardless of its position in a word. For example, *ende* is pronounced [ende] rather than [end], and *mule* is [mule] instead of [mul].

<sup>4</sup><https://middangeard.org.uk/aglardh/atanquesta>

ai	oi	ui	au	eu	iu
[ai̯]	[oi̯]	[ui̯]	[au̯]	[eu̯]	[iu̯]

Table 2: Quenya Diphthongs

This also applies to vowel combinations; *ea* is always pronounced as [ea], not as in *mean* [mi:n]. Tolkien used a dieresis (two dots) over vowels like *e* and *o* in his publications to indicate that these vowels should be pronounced fully, as in *ë* and *ö*.

**Consonants** The consonants *f*, *h*, *k* (also represented as *c*), *l*, *m*, *n*, *p*, *s*, *t*, *v* in Quenya are articulated similarly to their counterparts in English, with the notable exception that *p*, *t*, and *k* are not aspirated. The phoneme *y*, outside of palatalized contexts, is pronounced as in the English words *boy* and *year*; for instance, in Quenya *yára*.

	Bilabial	Labiodental	Dental	Alveolar	Palatal	Velar	Pharyngeal	Glottal
<b>Plosive</b>	p b			t d	c	k g	q	
<b>Nasal</b>	m			n	ɲ			
<b>Vibrant</b>				r				
<b>Tap or Flap</b>				ɾ				
<b>Fricative</b>	ɸ	f v	θ	s	ç	x		h
<b>Approximant</b>	w				j			
<b>Lateral Approximant</b>				l	ʎ			

Table 3: Quenya Consonants

The consonants *k* and *c* are consistently represented by the sound [k], regardless of their written form, and never manifest as [s] or [ts]. Similarly, the letter *θ* (also written as *th*) invariably represents the sound [θ], like the English *thin* or *thick*. The letters *t* and *v* have practically merged in pronunciation, both now typically rendered as [v], although the orthographic form preserves their etymological origins.

For the consonants *hw*, *hr*, and *hl*, *hw* [ɸ] resembles the English *wh* as pronounced in conservative Received Pronunciation or in Scottish and Irish English, exemplified in words like *where* and *what*. The phoneme *hr* [ʀ̥], found in the Icelandic word *hrafn*, and *hl* [ɬ̥], pronounced as the Welsh *ll* in *Llandudno*. However, pronouncing these voiceless consonants can be challenging. If speakers find it difficult, they will not significantly err by pronouncing them as their usual voiced counterparts: *w*, *r*, or *l*. In this project, due to current limitations of the ToucanTTS toolkit in supporting specific articulatory features, the sounds [ʀ̥] and [ɬ̥] are unavailable. Therefore, I have substituted these with [r] and [l] respectively. Furthermore, the consonants *b*, *d*, and *g* appear exclusively within consonant clusters and are never encountered as standalone sounds in native Quenya words.

Quenya includes several labialized and palatalized consonants with counterparts in other languages. The sound *kw* [kw], similar to the English *quick* and *quantum*, is typically spelled as *qu* in Quenya. The consonant *ny* [ɲ] mirrors the Dutch *oranje*, while *ty* [tʃ] is like the Hungarian *tyúk* and Romanian *chin*. The *ly* [ʎ] sound is found in American English *million*, and *hy* [ç] resembles what is heard in the English *hue* and German *nicht*. Double consonants such as *ll*, *pp*, *nn*, *ss*, *tt*, *rr*, *mm* are frequent. The complete set of Quenya consonants is in Table 3.

**Stress** Quenya’s stress rules for *Atanquesta* words depend on the number of syllables and their lengths. A syllable is **long** if it contains a **long vowel** (*á, é, ó, ú, î*), or a **diphthong** (*ai, oi, ui, au, ou*), and/or the vowel is followed by a **double consonant/consonant cluster**. In this respect, *Atanquesta* treats the palatal and labial consonants *qu, ly, ny, ty, hy* as clusters. Examples, with the long syllable in bold: *malle, ampano, keante, kára, huine, nalye*.

A syllable is **short** if it contains a single short vowel and is followed by a single consonant, or a vowel in hiatus. Examples: *kare, tuluwa, toa, keante*.

If the word has one or two syllables, then the stress falls on the **first syllable** of the word (shown by capitalizing the stressed syllable): *MÁ; MÁra; KAre; LASse*.

If the word has at least three syllables:

- if the second-from-end (**penultimate**) syllable is long, then that receives the stress: *kaRINwa; FeanÁro; ambalOTse; kaNASta*
- if the second-from-end syllable is short, then the stress falls on the syllable before it, i.e., the third-from-end (**antepenultimate**) syllable: *KÁrima; NAHtana; LINDale; MÁlime, TELume*.

## 2.2 Existing Quenya Language Corpora

The extensive fan base for Tolkien and Quenya is supported by numerous dedicated communities. This section enumerates several key resources, including notable works and websites, that are relevant to the study.

*The Elvish Writing Systems of J.R.R. Tolkien*<sup>5</sup> by Matt Coombes, *Atanquesta*<sup>6</sup>, *The Elvish Linguistic Society*,<sup>7</sup> *Merin Essi ar Quenteli*<sup>8</sup>, *Ardalambion*,<sup>9</sup> *Glémscrafu*,<sup>10</sup> *Tolkien Gateway*,<sup>11</sup> *Sindanórie*,<sup>12</sup> and *Parma Tyelpelassíva*<sup>13</sup> are among the highlighted resources. For online dictionaries, *Eldamo*,<sup>14</sup> and *Parf Edhellen*<sup>15</sup> provide extensive lexical databases for Elvish languages.

However, recordings of Quenya are scarce within these communities. Only *Glémscrafu*, *Atanquesta*, and *Merin Essi ar Quenteli* provide actual audio recordings. *Glémscrafu* features Quenya poetry and selected sentences from novels, recited by Bertrand Bellet and Benjamin Babut. *Atanquesta* is a comprehensive Quenya language course aimed at the general public, with lessons read by Tamas Ferencz. *Merin Essi ar Quenteli* was not selected for inclusion as it only contains isolated Quenya words, lacking substantial audio content. In this study, a dataset was created utilizing open-source recordings from *Glémscrafu* and *Atanquesta*, resulting in a total duration of 28 minutes. Additionally, linguistic experts recited selected Quenya poetry from *Sindanórie* and *Parma Tyelpelassíva*, contributing an additional 6 minutes of recordings to create a custom dataset. For more details, refer to the section 4.

## 2.3 Non-Autoregressive TTS Technologies

To better understand non-autoregressive TTS, this section is structured into five subsections. Section 2.3.1 provides a concise overview of the progression from initial concatenative methods to advanced deep learning autoregressive models. Section 2.3.2 discusses FastSpeech, and Section 2.3.3 delves into FastSpeech 2, outlining its advancements and improvements over the original. Section 2.3.4 focuses on the role of vocoders in synthetic speech enhancement. Finally, Section 2.3.5 examines IMS-Toucan, a system based on FastSpeech 2.

---

<sup>5</sup><https://www.kickstarter.com/projects/614014046/the-elvish-writing-systems-of-jrr-tolkien>

<sup>6</sup><https://middangeard.org.uk/aglardh/atanquesta>

<sup>7</sup><http://www.elvish.org/>

<sup>8</sup><https://realelvish.net/>

<sup>9</sup><https://folk.uib.no/hnohf/>

<sup>10</sup><https://glaemscrafu.jrrvf.com/english/index.html>

<sup>11</sup>[https://tolkiengateway.net/wiki/Main\\_Page](https://tolkiengateway.net/wiki/Main_Page)

<sup>12</sup><http://sindanoorie.net/>

<sup>13</sup><http://www.science-and-fiction.org/elvish/index.html>

<sup>14</sup><https://eldamo.org/index.html>

<sup>15</sup><https://www.elfdict.com/>

### 2.3.1 Evolution from Concatenative to Autoregressive Models

In the early development of speech synthesis, concatenative synthesis was a common approach (Taylor, 2009). This method involves stitching together pre-recorded audio clips to generate speech. While it delivered clear audio, it was not very flexible, struggling to capture various voice styles and emotions effectively. Technologies have evolved to incorporate Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) (Mu, Yang, & Dong, 2021). The traditional Statistical parametric speech synthesis (SPSS) network is a complex pipeline containing many modules (Zen, Agiomyrgiannakis, Egberts, Henderson, & Szczepaniak, 2016), composed of the text-to-phoneme network, audio segmentation network, phoneme duration prediction network, fundamental frequency prediction network and vocoder. Building these modules will take a lot of time and effort, and errors in any component can complicate training. End-to-end TTS methods transform text into speech using a unified model that makes the process simpler (Wang et al., 2017). These models learn efficiently from large datasets, automatically mastering the best acoustic and pronunciation features.

End-to-end TTS models can be divided into autoregressive and non-autoregressive types based on their decoding processes during inference. Autoregressive models first convert input text into a sequence of fixed-length speech representation vectors, then generate the acoustic features over time, where each time step’s output depends on previous outputs and the current speech vector (Shen et al., 2018). Autoregressive acoustic models like Tacotron<sup>16</sup> and Tacotron 2<sup>17</sup>, and Transformer-TTS<sup>18</sup> are examples. Tacotron uses the Griffin-Lim (Griffin & Lim, 1984) vocoder to generate speech waveforms, which results in lower audio quality. Although Tacotron 2 uses the WaveNet (Van Den Oord et al., 2016) vocoder to improve synthesis quality, both models use an autoregressive uni-directional long short-term memory (LSTM)-based decoder with the soft attention mechanism (Bahdanau, Cho, & Bengio, 2014), which faces challenges with parallel computation. Compared to fully feed-forward architectures, this architecture leads to less efficient training and inference on modern parallel hardware (Elias et al., 2021). Autoregressive models generate sequences sequentially, which can lead to issues like repeated or skipped words, and difficulty in finely controlling speech pace and rhythm. Unlike autoregressive sequence generation, nonautoregressive models generate sequence in parallel, without explicitly depending on the previous elements, which can greatly speed up the inference process (Ren et al., 2019).

---

<sup>16</sup><https://google.github.io/tacotron/>

<sup>17</sup><https://github.com/NVIDIA/tacotron2?tab=readme-ov-file>

<sup>18</sup><https://github.com/as-ideas/TransformerTTS>

### 2.3.2 FastSpeech

The architecture of FastSpeech differs from the conventional sequence-based encoder-attention-decoder structure. It primarily consists of two main components: a Feed-forward Transformer (FFT) and a Length Regulator (LR). The Feed-forward Transformer (FFT) replaces the conventional attention mechanism, and the Length Regulator (LR), which includes a Duration Predictor, adjusts the output sequence length to match the duration of the phonemes. The Duration Predictor estimates the durations of the phonemes to guide the Length Regulator. The overall model architecture of FastSpeech is shown in Figure 1.

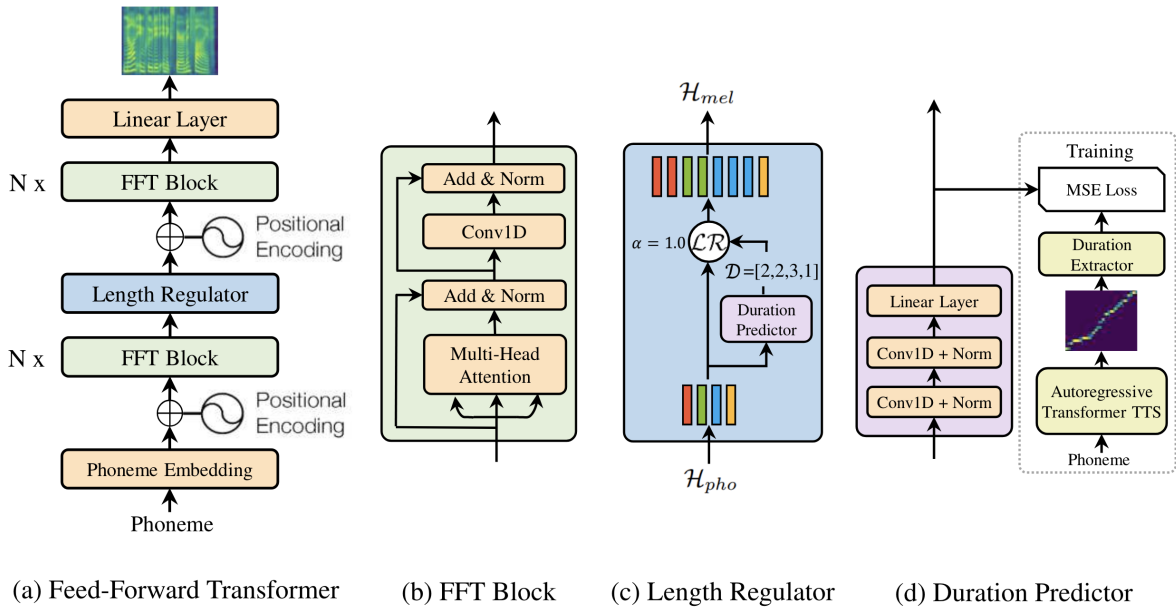


Figure 1: The overall architecture for FastSpeech. The duration predictor (Ren et al., 2019).

**Feed-forward Transformer** FFT is based on the Transformer (Vaswani et al., 2017) and one-dimensional convolutional feed-forward networks (Gehring, Auli, Grangier, Yarats, & Dauphin, 2017; Ping et al., 2018). As shown in Figure 1(a), it stacks multiple FFT modules to establish the mapping relationship between phoneme sequences and the Mel spectrogram. There are  $N$  FFT modules on both the phoneme and Mel spectrogram sides, with a Length Regulator in between to compensate for the length disparity between phonemes and the Mel spectrogram. Each FFT module features a multi-head attention mechanism and a two-layer one-dimensional convolutional network with ReLU activation. The positional information of phonemes is handled using positional encoding from the Transformer model. The multi-head attention in FFT is used to extract positional information, while the one-dimensional convolutional network ensures that the connections between two adjacent hidden states are as close



as the connections between phonemes and the Mel spectrogram.

**Length Regulator** The Length Regulator resolves mismatches between the phoneme sequence length and the Mel spectrogram sequence, and controls speech speed and certain prosodic aspects. Typically, the phoneme sequence is shorter than the Mel spectrogram sequence, with each phoneme corresponding to multiple Mel spectrograms. The Length Regulator expands the hidden states of the phoneme sequence to match the length of the Mel spectrograms. For instance, if the hidden states of the phoneme sequence are  $H_{\text{pho}} = [h_1, h_2, \dots, h_n]$  and the phoneme durations are  $D = [d_1, d_2, \dots, d_n]$ , the Length Regulator is represented by the equation  $H_{\text{mel}} = LR(H_{\text{pho}}, D, \alpha)$ , where  $\alpha$  is a hyperparameter that determines the length of the expanded sequence, thereby controlling the speech speed. Adjusting  $\alpha$  changes the speech speed, and modifying the duration of space characters can alter some aspects of the speech prosody.

**Duration Predictor** The Duration Predictor consists of a two-layer 1D convolutional network with ReLU activation, followed by layer normalization and a dropout layer, and an additional linear layer that outputs a scalar, which represents the predicted phoneme duration. This module is stacked on top of the FFT blocks on the phoneme side and is jointly trained with the FastSpeech model to predict the length of mel-spectrograms for each phoneme using mean square error (MSE) loss. Length prediction is performed in the logarithmic domain, which renders the data more Gaussian and easier to train. The Duration Predictor is trained using true phoneme durations extracted from an autoregressive teacher TTS model, as shown in Figure 1d. The training process involves first training an autoregressive encoder-attention-decoder-based Transformer TTS model following reference (Li, Liu, Liu, Zhao, & Liu, 2019). For each training sequence pair, decoder-to-encoder attention alignments are extracted from the trained teacher model. Due to the multihead self-attention (Vaswani et al., 2017), multiple attention alignments are available, and not all attention heads demonstrate the diagonal property where the phoneme and mel-spectrogram sequences are monotonically aligned. A focus rate  $F$  is introduced to measure how closely an attention head approximates a diagonal alignment:  $F = \frac{1}{S} \sum_{s=1}^S \max_{1 \leq t \leq T} a_{s,t}$ , where  $S$  and  $T$  are the lengths of the ground-truth spectrograms and phonemes, respectively, and  $a_{s,t}$  denotes the element in the  $s$ -th row and  $t$ -th column of the attention matrix. The focus rate for each head is calculated, and the head with the highest  $F$  is selected for the attention alignments. The phoneme duration sequence,  $D = [d_1, d_2, \dots, d_n]$ , is then extracted based on the duration extractor  $d_i = \sum_{s=1}^S [\arg \max_t a_{s,t} = i]$  meaning the duration of a phoneme is the number of mel-spectrograms attended to it according to the selected attention head. Adjustments to the duration of space characters in the sentence allow for control over parts of the speech prosody.

### 2.3.3 Fastspeech 2

FastSpeech 2 addresses the issues found in FastSpeech and enhances the solution to the one-to-many mapping problem in text-to-speech synthesis. It improves the model by directly training with ground-truth targets rather than relying on simplified outputs from a teacher model. Additionally, it introduces more variation information in speech, such as pitch, energy, and more accurate duration, which are used as conditional inputs to enrich the model’s performance and output quality. This variance adaptor includes several components: a duration predictor (also known as the length regulator from FastSpeech, which is detailed in paragraph 2.3.2.), a pitch predictor, and an energy predictor, as illustrated in Figure 2.

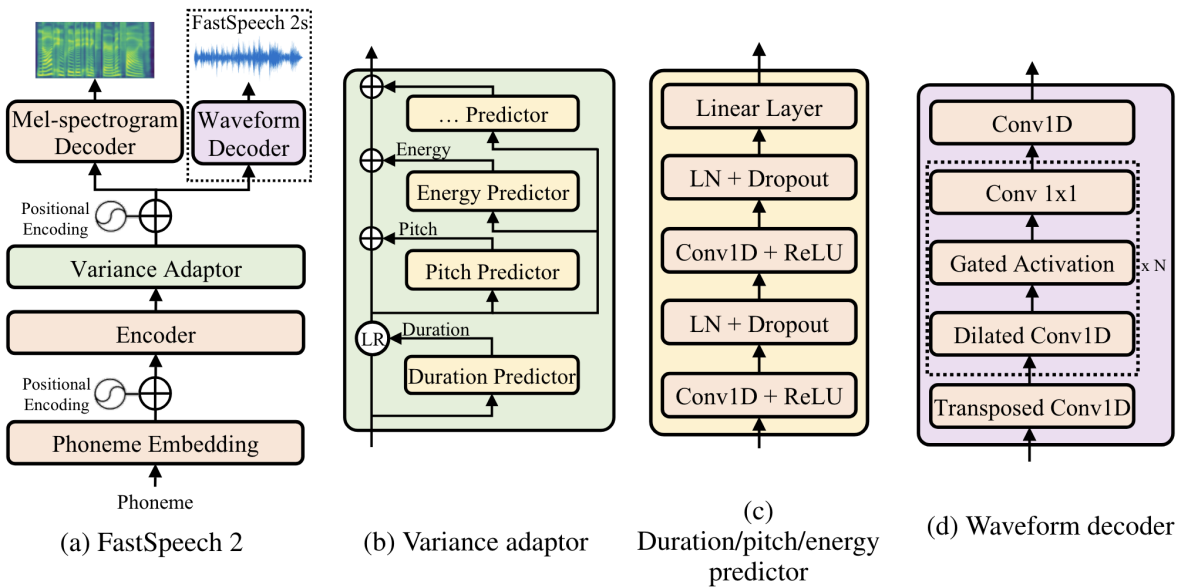


Figure 2: The overall architecture for FastSpeech 2 and 2s (Ren et al., 2020).

**Duration Predictor** Unlike FastSpeech, which relies on a pre-trained autoregressive TTS model for phoneme duration extraction, this method employs the Montreal Forced Alignment (MFA) (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) tool to enhance alignment accuracy and minimize the information gap between model input and output.

**Pitch Predictor** To more accurately predict pitch contour variations, the continuous wavelet transform (CWT) is used to break down the ongoing pitch series into a pitch spectrogram (Hirose & Tao, 2015; Suni, Aalto, Raitio, Alku, & Vainio, 2013). This spectrogram is then used as the training target for the pitch predictor, which is optimized using mean square error (MSE) loss. During inference, the pitch predictor outputs a pitch spectrogram that is converted back into a pitch contour with the inverse continuous wavelet transform (iCWT). For both training

and inference, the pitch contour  $f_0$  (fundamental frequency) for each frame is quantized into 256 possible values on a logarithmic scale, transformed into a pitch embedding vector  $p$ , and added to the expanded hidden sequence.

**Energy Predictor** The L2-norm of the amplitude of each short-time Fourier transform (STFT) frame is computed as the energy. The energy of each frame is then quantized to 256 possible uniform values, encoded into an energy embedding  $e$ , and added to the expanded hidden sequence similarly to pitch. An energy predictor is used to predict the original values of energy rather than the quantized values, and this predictor is optimized with MSE loss.

### 2.3.4 Vocoder

WaveGlow (Prenger, Valle, & Catanzaro, 2019) is a flow-based vocoder that combines the ideas from Glow, a generative flow network, and WaveNet (Van Den Oord et al., 2016), a powerful autoregressive model. Its main advantage is its ability to generate high-quality speech relatively quickly due to its non-autoregressive nature. Compared to traditional vocoders that may face challenges with speed or require significant computational resources, such as WaveNet, WaveGlow offers a more efficient alternative with minimal compromise on quality. For FastSpeech (Ren et al., 2019), using a pre-trained WaveGlow model as a vocoder means that the output Mel spectrograms can be processed into audio samples more quickly and efficiently. This is beneficial as FastSpeech generates these Mel spectrograms through a non-autoregressive process, inherently enhancing the overall speed of speech synthesis. However, compared to newer vocoders, WaveGlow still requires considerable computational resources, which may limit its use in resource-constrained environments. FastSpeech 2 (Ren et al., 2020) advances these capabilities by incorporating newer neural vocoders like Parallel WaveGAN (Yamamoto, Song, & Kim, 2020) and HiFi-GAN (Kong, Kim, & Bae, 2020), which further improve the efficiency and quality of speech synthesis introduced by WaveGlow.

**Parallel WaveGAN** A non-autoregressive model that uses a generative adversarial network (GAN) architecture, significantly speeding up the speech synthesis process by allowing faster and parallel computation of audio samples from Mel spectrograms. The audio quality produced by Parallel WaveGAN is commendable, often nearing that of autoregressive models like WaveNet, but at a fraction of the computational cost.

**HiFi-GAN** Known for producing high-fidelity audio, HiFi-GAN improves upon earlier GAN-based vocoders by optimizing the generator and discriminator architecture for better audio quality. It is especially renowned for producing clear, crisp, and natural-sounding audio at high speeds, making it an excellent match for FastSpeech 2. HiFi-GAN not only ensures rapid

waveform generation but also handles various nuances in audio, making the synthesized speech sound more natural and less processed.

### 2.3.5 IMS-Toucan

IMS Toucan (Lux et al., 2023) is a toolkit from the Institute for Natural Language Processing at the University of Stuttgart in Germany. It’s designed to help people learn and use the latest voice synthesis technology. Introduced during the 2021 Blizzard Challenge, this toolkit primarily uses a modified version of FastSpeech 2 (Ren et al., 2020). In IMS Toucan, FastSpeech 2 has two key changes. First, it averages the pitch and energy of speech sounds (phonemes) based on their duration, a technique from FastPitch (Łańcucki, 2021) used in ESPnet (Hayashi et al., 2020; Watanabe et al., 2018) to enhance voice control. This means users can adjust the pitch, energy, and length of speech sounds during voice generation for detailed customization. Second, IMS-Toucan uses the Conformer (Gulati et al., 2020), which merges convolutional networks and transformers. Initially designed for voice recognition, the Conformer is also highly effective in speech synthesis and is part of ESPnet (Guo et al., 2021). During the 2023 Blizzard Challenge, the submission was improved from the previous one in 2021. These improvements are the result of two years of development on the IMS Toucan toolkit, which now includes various designs to better manage multilingual capabilities, controllability, and scenarios with limited resources. The system that integrates all these features is named ToucanTTS. An overview of this architecture is presented in Figure 3.

**Text-to-Phoneme** To convert text into phonemes, an open-source phonemizer<sup>19</sup> utilizing `espeak-ng`<sup>20</sup> as its backend is employed. The process begins with basic text cleaning, followed by transforming the input into a sequence of phonemes using IPA notation. These phonemes are then converted into articulatory vectors through a lookup table (Lux & Vu, 2022), which provides more effective cross-lingual modeling than traditional phoneme labels (Do et al., 2023). Each vector represents a one-hot encoding of the human vocal tract’s configuration during sound production. Additionally, these vectors are enhanced with extra dimensions to represent nonsegmental markers such as lengthening, shortening, and lexical stress (Lux, Koch, & Vu, 2022b). The phonemizer also generates symbols that, while not forming phonemic units on their own, modify adjacent phonemes by altering the corresponding dimensions in the articulatory vector.

---

<sup>19</sup><https://github.com/bootphon/phonemizer>

<sup>20</sup><https://github.com/espeak-ng/espeak-ng>

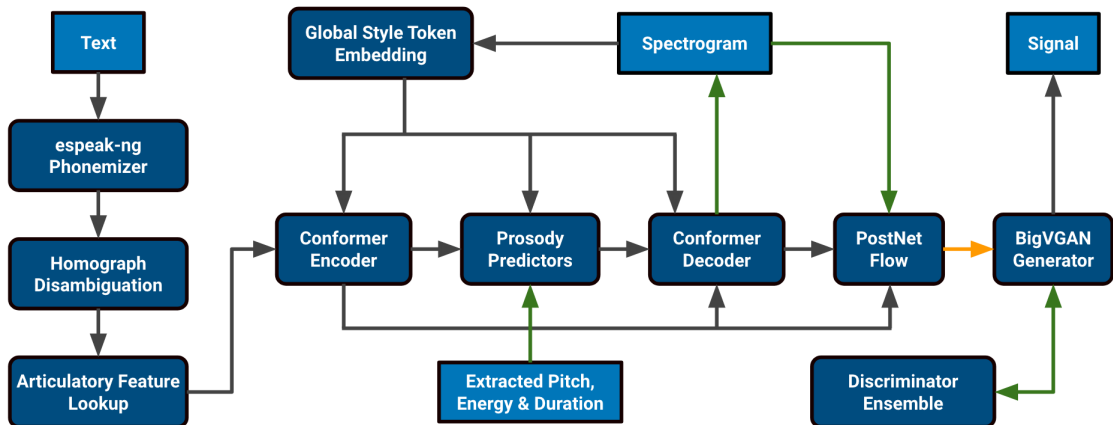


Figure 3: Overview of all the components in our system. The green arrows show the losses applied at training time. The orange arrow only exists during inference, the gradient is not passed through at training time (Lux et al., 2023).

**Spectrogram-to-Alignment and Embedding** The approach relies on precise alignments of phonemes to spectrogram frames, influenced by the model’s learned durations and the averaging of pitch and energy values over these durations. Precise alignments are achieved by training a simple speech recognition system with a CTC objective (Graves, Fernández, Gomez, & Schmidhuber, 2006), which models the likelihood of all phonemes over time. The resulting posteriogram is input into an auxiliary spectrogram reconstruction model, which aims to reconstruct the inputs to ensure sharper phoneme boundary definitions. For extracting alignments from posteriograms, the axis containing phoneme likelihoods is reordered according to the phoneme sequence in the transcription, and a monotonic alignment search (MAS) is conducted.

To disentangle and capture varying acoustic conditions and speaking styles, the Global Style Token embedding approach is employed, augmented by strategies from AdaSpeech 4 (Wu et al., 2022). These include a style token disentanglement loss and an increase in the number of style tokens to 2000. The embeddings are integrated after every encoder block, decoder block, and each layer in the prosody predictors, using concatenation followed by projection.

**Phoneme-to-Spectrogram** The spectrogram generation network utilizes the basic structure of FastSpeech 2 (Ren et al., 2020), augmented with phoneme-wise averaging of pitch and energy following the FastPitch (Łańcucki, 2021) methodology. This configuration provides a high level of fine-grained control over the generated speech. For enhanced efficiency, the Conformer architecture (Gulati et al., 2020) is used as both encoder and decoder, known for its effectiveness across various speech tasks. Additionally, the system incorporates a PostNet

using normalizing flows, inspired by analyses in PortaSpeech (Kim, Kim, Kong, & Yoon, 2020; Ren, Liu, & Zhao, 2021).

**Vocoder** As the neural vocoder for performing spectrogram inversion, ToucanTTS utilizes a generative adversarial network (GAN) setup that includes the BigVGAN generator (S.-g. Lee, Ping, Ginsburg, Catanzaro, & Yoon, 2022), which offers improvements over HiFi-GAN (Kong et al., 2020), along with discriminators from MelGAN (Kumar et al., 2019), HiFiGAN, and Avocodo (Bak et al., 2023).

## 3 Related Study

### 3.1 Constructed languages

A constructed language is deliberately crafted by an individual or group, rather than emerging from the natural evolution and changes of languages over time (Adams, 2011; Okrent, 2009). Within academia, individuals who develop constructed languages are variously referred to as language creators, language planners, language inventors, language engineers, or language architects (Adams, 2011; Gobbo, 2017; Peterson, 2015; Schreyer, 2021b). These languages themselves fall under categories such as planned languages (Gobbo, 2017; Janton, 1993; Tonkin, 2015), fictional languages (Barnes & Van Heerden, 2006; Kazimierczak, 2010; Schreyer, 2021b), and artificial languages (Schreyer, 2021a). Schubert (2011) also refer to the overall study of planned languages as “interlinguistics”. Constructed languages are often categorized based on their method of creation and their intended purpose. In terms of creation, languages are described as either a priori, which means they are made from scratch without influence from other languages, or a posteriori, indicating they are constructed with influences from one or more existing languages (Schreyer, 2021b). Regarding purpose, constructed languages are classified as auxlangs, which are intended as international auxiliary languages to facilitate communication (Forster, 1982), such as Esperanto. Alternatively, arlangs are used for artistic purposes in media or literature, like Klingon in the "Star Trek" series, Dothraki in "Game of Thrones," or Quenya and Sindarin in J.R.R. Tolkien’s Middle-earth writings.

It is widely believed that the oldest recorded constructed language is Lingua Ignota (Latin for "unknown language"), created by the twelfth-century nun Hildegard von Bingen (Higley, 2007). L.L. Zamenhof was an important figure in the history of constructed languages. He created Esperanto, aiming to improve global communication and understanding. Esperanto is now one of the most successful constructed languages, with a worldwide community of speakers that crosses national borders. Another constructed language with a significant community is Klingon. Created by linguist Marc Okrand for the 1984 film *Star Trek III: The Search for Spock*, Klingon now has up to 7,500 learners, about 120 fluent speakers (Windsor & Stewart, 2017). Initially, Okrand built on words from past *Star Trek* episodes created by actor James Doohan (Okrand, Adams, Hendriks-Hermans, & Kroon, 2011). Since its creation, Klingon has grown worldwide. Various linguistic studies have examined Klingon, including a typological analysis (Sutrave, 2017) and surveys of other constructed languages like Quenya, Dothraki, and Na’vi. These studies found that while Quenya, Dothraki, and Na’vi often follow natural language patterns, Klingon defies many of Greenberg’s Linguistic Universals, making it unique (Destruel, 2016). Learning Klingon is supported by resources such as a Klingon dictionary



(Okrand, 1992) and books like *The Klingon Way: A Warrior's Guide* (Okrand, 1996) and *Klingon for the Galactic Traveler* (Okrand, 1997). These materials cover dialects, specialized vocabulary, idioms, and slang, helping establish a robust speech community. The *Klingon Language Institute*<sup>21</sup>, founded in 1991, further supports this virtual community.

One of the pioneers in popularizing language creation was J.R.R. Tolkien. His method involved developing historical language families, a technique that many language creators still use today. Tolkien focused deeply on the art and aesthetics of his languages, personally investing in their sounds and beauty. The discovery of Finnish grammar had a profound impact on Tolkien. According to him, his invented languages "became heavily Finnicized in phonetic pattern and structure." (Carpenter et al., 1981). This led to the creation of Qenya, the language of the High Elves (Tikka, 2007), which was later renamed Quenya. The change was mainly orthographic, with a slight difference in pronunciation, as discussed in detail in Section 2.1.3. Finnish had a strong influence on the early forms of the language, especially in vocabulary, where many words were Finnish in style. Over time, Quenya distanced itself somewhat from Finnish, but the influence never completely disappeared (Perälä, 2002). Tolkien reduced his borrowing of Finnish words, but the phonetic and structural influence of Finnish on Quenya deepened (Tikka, 2007). While Finnish influence on Quenya is noticeable, Quenya remains a unique language with parallels to many different languages. Tolkien didn't directly borrow from languages but used them as inspiration, creating a language with both uniqueness and depth.

In building TTS systems, constructed languages and low-resource languages face similar challenges. Constructed languages are often considered endangered because they are rarely learned as a first language at home, have few speakers, and lack official status and prestige (Schreyer, 2011). Both types of speech communities are similar because their members usually learn the languages later in life and do so voluntarily (Christoph, 2012). Contact languages are tied to community identity, and this idea can also apply to constructed languages like Esperanto, which has been described as an Eastern European contact language (Lindstedt, 2009). Both contact and constructed languages can encode various types of local knowledge and provide insight into the creator's worldview (N. H. Lee, 2020; Schreyer, 2021b). When constructed languages like Esperanto become popular, people who share similar values may learn them to understand this worldview better (Schreyer, 2021b). Studying contact and constructed languages is important for understanding the full range of human linguistic possibilities (N. H. Lee, 2020; Schreyer, 2021b). Language creation is always influenced by existing languages because they are made by humans who use language (van Oostendorp, 2019). The relationship between

---

<sup>21</sup><https://www.kli.org/>



language planning and planned languages shows that language revival processes, such as those for Hebrew or Cornish (Romaine, 2011; Tonkin, 2015), involve similar steps of lexical expansion and standardization. Comparisons of lexical expansion projects for Te Reo Māori and Esperanto (Krägeloh & Neha, 2014) show similar processes in both communities. Speakers of endangered languages can learn from constructed language speakers about using media, information technology, and other language planning methods (Schreyer, 2011). Creating a strong, enthusiastic interest in endangered languages could inspire more people, especially community members, to learn and use these languages (Schreyer, 2015).

### 3.2 Low-Resource Language TTS

As detailed in Section 2.3, the advance of deep learning (Goodfellow et al., 2014; Vaswani et al., 2017) has led to significant improvements in the field of TTS. End-to-end models, such as Tacotron 2 (Elias et al., 2021; Shen et al., 2018), TransformerTTS (Li et al., 2019), FastSpeech 2 (Ren et al., 2020), FastPitch (Łańcucki, 2021), have achieved unprecedented quality and controllability in speech synthesis. These models typically rely on vocoders like WaveNet (Van Den Oord et al., 2016), MelGAN (Kumar et al., 2019), Parallel WaveGAN (Yamamoto et al., 2020), and HiFi-GAN (Kong et al., 2020) to convert parametric representations into waveforms. Models like EATS (Donahue, Dieleman, Bińkowski, Elsen, & Simonyan, 2020) and VITS (Kim, Kong, & Son, 2021) have been developed to generate waveforms directly from grapheme or phoneme input sequences. While these methods perform remarkably well with sufficient data, cross-lingual data usage remains a significant challenge in TTS. For instance, the Tacotron model requires more than 10 hours of training data to produce high-quality synthesized speech (Chung, Wang, Hsu, Zhang, & Skerry-Ryan, 2019). Collecting such large amounts of speech data is expensive and time-consuming, which poses substantial challenges for developing TTS systems for the many less widely spoken languages around the world.

Various techniques have been proposed to address the low-resource TTS problem, with transfer learning being one of the commonly used methods (Weiss et al., 2016). This method involves pre-training the acoustic model in a different language with sufficient training data (the source language) and then fine-tuning the model with the limited data available for the target low-resource language. This approach uses underlying similarities between languages, such as pronunciation patterns and semantic structures, to improve the mapping between input (text or phoneme sequence) and output (speech features) in the target language (Tan et al., 2021). However, cross-lingual transfer learning presents challenges, primarily due to mismatches between the input embeddings of the source and target languages caused by differences in phoneme sets or orthographic characters. To address this, researchers have proposed solutions like the Phonetic Transformation Network (Tu, Chen, Yeh, & Lee, 2019), which includes

an automatic speech recognition component to map input symbols across languages based on their sounds. Wells and Richmond (2021) have experimented with using phonemes and phonological features as inputs, utilizing linguistic expertise to enhance the mapping of embeddings between source and target languages. Gutkin (2017) also applied phonological features to low-resource TTS with considerable success. Lux and Vu (2022) fixed previous shortcomings by using a linguistically motivated representation of the inputs to such a system (articulatory and phonological features of phonemes) that enables cross-lingual knowledge sharing and applying the model-agnostic meta-learning (MAML) framework to the field of low-resource TTS for the first time. Do, Coler, Dijkstra, and Klabbers (2021) confirmed the improvement in output speech quality in multilingual models over their monolingual counterparts. Lux et al. (2022b) demonstrated that with a simple encoder design, a mechanism to encode word boundaries, and the language agnostic meta learning training procedure, a low-resource capable multilingual zero-shot multispeaker TTS can be achieved. Do et al. (2021) found that language family classification, despite its widespread use, was ineffective for selecting source languages. Instead, they proposed using Angular Similarity of Phoneme Frequencies (ASPF), which measures the similarity between the phoneme systems of two languages (Do, Coler, Dijkstra, & Klabbers, 2022). They also investigated the effectiveness of phone labels versus articulatory features for cross-lingual transfer learning in TTS applications for low-resource languages (Do et al., 2023).

There has been limited research on TTS for constructed languages, but some attempts have been made. For example, Esperanto TTS systems include *Parol*<sup>22</sup> and *EsperantoTTS*<sup>23</sup>, and Jokisch and Eichner (2000) developed a TTS system for Klingon. The experience from developing TTS for constructed languages can also be applied to extinct and ancient languages needing speech reconstruction. Using small datasets read by expert linguists, researchers can recreate the speech of these languages, helping to better understand and preserve cultural heritage.

---

<sup>22</sup><https://parol.martinrue.com/>

<sup>23</sup><https://54696d21.github.io/esperantoTTS/>

## 4 Methodology

This section outlines the development and evaluation processes for a Quenya TTS system. Section 4.1 focuses on the creation of the Quenya dataset and Grapheme-to-Phoneme Conversion (G2P). Section 4.2 describes the training of TTS models for both Finnish and English using ToucanTTS as the source language model. Section 4.3 details the fine-tuning of these models with a Quenya dataset to improve the quality of synthesized speech, also utilizing the multilingual pre trained points available on ToucanTTS. Section 4.4 outlines the specific methods for evaluating the Quenya TTS system.

### 4.1 Quenya Dataset and Grapheme-to-Phoneme (G2P) Script

#### 4.1.1 Building Quenya Dataset

The dataset for the Quenya text-to-speech system comes from two sources. Firstly, as mentioned in Section 2.2, one source is the public recordings from *Atanquesta* and *Glémscrafu*. As previously discussed in Section 2.1.3, Qenya and Quenya are considered the same language, so both the Qenya<sup>24</sup> and Quenya<sup>25</sup> sections from *Glémscrafu* are included in the dataset. The second source is a custom dataset created by a linguistic expert who recited selected Quenya poetry from *Sindanórie* and *Parma Tyelpelassíva*. The expert is a young male, a native Russian speaker with a minor in Finnish, familiar with the phonetics of most Indo-European and Uralic languages.

The structure of the dataset mimics the **LJSpeech**<sup>26</sup> format, respecting sentence boundaries and segmenting audio into audio files ranging from a maximum of about 13 seconds to a minimum of 2 seconds, with transcription texts manually annotated. The dataset was manually segmented in Audacity<sup>27</sup>. The recordings from *Glémscrafu* underwent noise reduction using voicefixer<sup>28</sup>. Finally, the dataset from the public open-source recordings totals **28** minutes, while the custom-built dataset adds an additional **6** minutes, making a combined total of **34** minutes.

#### 4.1.2 Building Quenya G2P

As mentioned in the 2.3.5 paragraph, ToucanTTS employs an open-source Phonemizer, relying on espeak-ng to convert text into phonemes (Lux et al., 2023). However, since espeak-ng does

---

<sup>24</sup><https://glaemscrafu.jrrvf.com/english/qenya.html>

<sup>25</sup><https://glaemscrafu.jrrvf.com/english/quenya.html>

<sup>26</sup><https://keithito.com/LJ-Speech-Dataset/>

<sup>27</sup><https://www.audacityteam.org/>

<sup>28</sup><https://github.com/haoheliu/voicefixer>

not support Quenya, a Quenya Grapheme-to-Phoneme (G2P) script was developed using the knowledge of Quenya phonology discussed in Section 2.1.3 and is openly available on GitHub; the link can be found in the appendices. This script maps words to phonemes and annotates stress. As an artificial language, Quenya has relatively fixed rules with few exceptions, enabling the script to correctly annotate most pronunciations. To better recognition and synthesis of early Quenya variants, the script was also informed by the Quenya dictionary *Eldamo*, including some obsolete letters such as  $\theta$  and  $\tilde{n}$ .

## 4.2 Training the Finnish and English TTS Models

### 4.2.1 Finnish Dataset

The dataset used for training Finnish is derived from **CSS10**<sup>29</sup>, a collection of single-speaker speech datasets for ten languages (Park & Mulc, 2019). The total duration of the Finnish dataset is 10 hours and 32 minutes. During its use, some transcription errors were found in the dataset, such as chapter numbers included in the transcribed texts that were not read aloud, and some text content not matching the spoken audio. To enhance the quality of synthesized speech and avoid potential errors, transcriptions were manually corrected. Normalized texts were also selected and reformatted to match the LJSpeech format for processing.

### 4.2.2 English Dataset

The dataset used for training the English TTS is **LJSpeech**, which includes 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books, totaling about 24 hours. To ensure a fair comparison, this study randomly extracted 10 hours and 31 minutes from the dataset for training, equivalent to the Finnish dataset used. No other modifications were made to the dataset.

### 4.2.3 Model Training and Evaluation Results

Both Finnish and English were trained from scratch using ToucanTTS, with identical training configurations: batch size of **12**, learning rate of **1e-3**, and a total of **80,000** training steps. Both languages used the built-in aligner from ToucanTTS, with no fine-tuning applied to the aligner. Each model then generated **10** sentences that were not present in the training set, and the word error rate (WER) was tested using OpenAI’s Whisper<sup>30</sup> automatic speech recognition, ignoring case sensitivity. The English model achieved of **3.7%**, while the Finnish model achieved of **3.4%**. For comparison, human speech achieved a WER of **0.9%** for English

---

<sup>29</sup><https://github.com/Kyubyong/css10>

<sup>30</sup><https://github.com/openai/whisper>

and **1.7%** for Finnish. These results indicate that both models have reached a reasonable level of intelligibility. Audio demonstrations generated by these models can be found in the links provided in the Appendix.

### 4.3 Fine-Tuning Models with Quenya Data

First, the Quenya G2P script mentioned earlier was integrated into the text front end of ToucanTTS. Fine-tuning was then performed using an example file provided by the system, with the training configuration including a learning rate of **1e-5**, batch size of **6**, and **6000** training steps. Additionally, the original aligner was also fine-tuned. After fine-tuning the Finnish and English models, further adjustments were made using the same configuration on a multilingual pre-trained checkpoint developed with data from 12 languages: English, German, Spanish, Greek, Finnish, French, Russian, Hungarian, Dutch, Polish, Portuguese, and Italian, totaling 389 hours and including all speech datasets from CSS10 and LJSpeech. After completing the training, the final three checkpoints were averaged and consolidated into a single optimized checkpoint, for use during inference. To evaluate the effectiveness, each model generated **10** sentences that were not present in the training data.

### 4.4 Evaluation

To better assess the quality of the speech, this study used traditional Mean Opinion Score (MOS) evaluations and detailed listening reports. Surveys were distributed to communities of Quenya enthusiasts and linguists, and two proficient Quenya speakers were invited to identify specific errors in sentences generated by the TTS and to provide detailed listening reports. In the MOS evaluation, participants were provided with a total of 9 sentences, each available in 4 different versions: real human voice recordings, sentences fine-tuned from a multilingual model, sentences fine-tuned from Finnish, and sentences fine-tuned from English. These sentences came with their transcribed and phonetic texts, but the participants were not informed which voices were produced by humans. Participants were required to rate each voice sample on a scale from 1 to 5, where 1 point indicates the voice is extremely unnatural and almost unrecognizable, and 5 points signify the voice is very natural, nearly indistinguishable from a real human voice. The survey was conducted using the Qualtrics<sup>31</sup> web platform and distributed online. As it was disseminated among targeted groups and communities, no personal information such as native language or gender was recorded from the participants. An example of the survey instrument is available in the Appendix.

---

<sup>31</sup><https://www.qualtrics.com/>

## 5 Result

### 5.1 Mean Opinion Scores (MOS)

After excluding incomplete and evidently randomly filled questionnaires, such as those where all responses were rated as 1 point, a total of **31** valid surveys were collected. The Mean Opinion Scores (MOS) for different speech models in this task are presented in Figure 4 and Table 4, while the statistical significance of the results is confirmed by the Wilcoxon Signed-Rank Test shown in Table 5.

Table 4 illustrates the average MOS for each speech model. The Real Voice model achieves the highest score of **4.59**, indicating the highest level of satisfaction among participants. Following closely, the Meta Model has an average score of **4.33**, demonstrating a high quality of synthetic speech, though slightly less preferred than the Real Voice. The Finnish Model, with a MOS of **3.84**, performs moderately well but still lags behind the Real Voice and Meta Model, suggesting that there is room for improvement in its speech synthesis quality. The English Model has the lowest MOS of **2.25**, reflecting the least satisfactory performance among the evaluated models.

Model	MOS
Real Voice	4.59
Meta Model	4.33
Finnish Model	3.84
English Model	2.25

Table 4: Average Mean Opinion Scores (MOS) for Different Speech Models

Comparison	p-value	Significance
Real Voice vs Meta Model	0.0113	Significant
Meta Model vs Finnish Model	< <b>0.001</b>	Highly Significant
English Model vs Finnish Model	< <b>0.001</b>	Highly Significant

Table 5: Wilcoxon Signed-Rank Test Results

Table 5 provides the p-values from the Wilcoxon Signed-Rank Test, offering insights into the statistical significance of the differences between the models. The comparison between the Real Voice and Meta Model yields a p-value of **0.0113**, which is less than **0.05**, indicating a significant difference. This result suggests that while both models perform well, participants significantly prefer the Real Voice over the Meta Model. The comparison between the Meta

Model and Finnish Model results in a p-value of  $< \mathbf{0.001}$ , indicating a highly significant difference. This result shows that the Meta Model is significantly preferred over the Finnish Model, reinforcing the superiority of the Meta Model's synthetic speech quality. The comparison between the English Model and Finnish Model yields a p-value of  $< \mathbf{0.001}$ , also indicating a highly significant difference. This finding underscores the substantial preference for the Finnish Model over the English Model, despite the Finnish Model itself needing improvements.

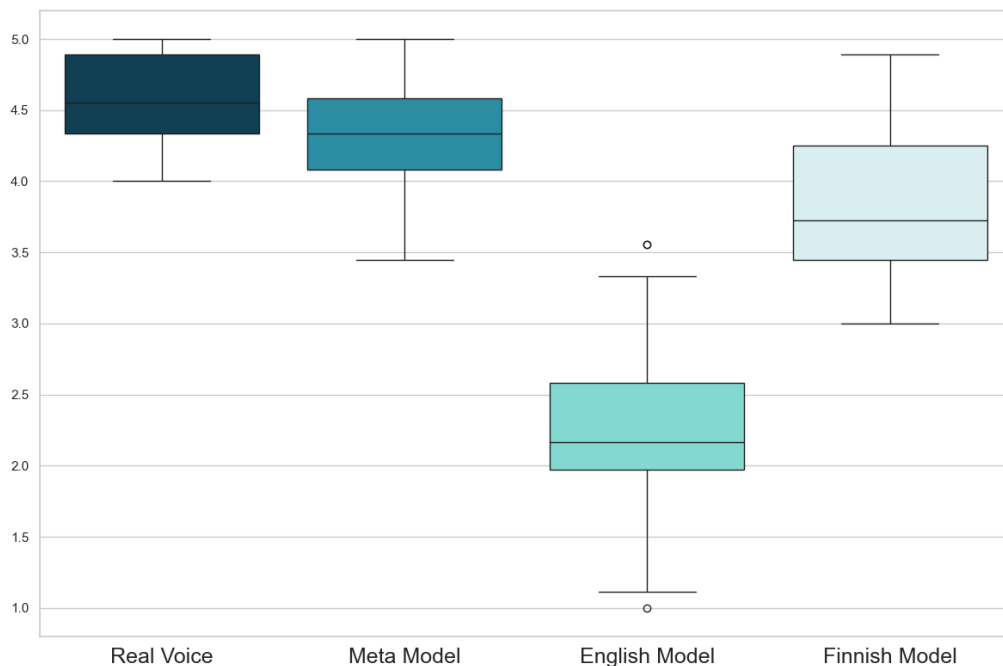


Figure 4: Mean Opinion Scores (MOS) for Different Speech Models

## 5.2 Listening Reports

In manually marking incorrect or missing words in sentences, both professional speakers indicated that due to the low intelligibility of the model tuned from English, there was no need to meticulously annotate its errors. Therefore, the main marking effort was concentrated on the models tuned from Finnish and the multilingual pre-trained point. Here, one particularly interesting and typical examples are selected for detailed analysis.

- Ai! laurië lantar lassi súrinen, yéni únótimë ve rámar aldaron!
- 'ai! l'aurië l'antar l'assi s'u:rinɛn, j'e:ni u:n'o:time ve r'a:mar 'aldaron!

This line of poetry comes from the most famous Quenya poem, **Namárië**, known as "Farewell," which translates to: **"Ah! like gold fall the leaves in the wind, long years numberless**

**as the wings of trees!"** The parts highlighted in **red** indicate errors made by the Finnish-tuned model, while those highlighted in **green** represent errors made by the meta-tuned model. From this example, it is clear that Finnish lacks the Quenya vowels **ɛ** and consonants **j**, with pronunciations in this language still sounding like the Finnish **e** and **y**. Both the Finnish-tuned and meta-tuned models are unable to produce the trilled **r**, and instead produce sounds more like to a tap **r**. This shows that source language models lacking target language phonetics can significantly impact the quality of synthesized speech. Further details and implications of this finding will be discussed in the following section 6.



## 6 Discussion

According to the results presented in Section 5, the study confirmed the hypothesis proposed in Section 1.1, demonstrating that the model transferred from Finnish indeed performed better than the one transferred from English. And the results generated by fine-tuning from a multilingual model exceeded those from both the Finnish and English models. Although the hypothesis was validated, the Finnish model’s performance was not as strong, and the best-performing multilingual model still exhibited noticeable differences from real voice and lacked naturalness, indicating room for improvement. This section will delve into a detailed discussion of the results, reflect on the limitations of this study, and propose directions for future research.

### 6.1 Challenges with the Finnish Model

The Finnish model’s limited performance can be attributed to Finnish not being the most suitable source language for Quenya. Although Finnish had been a major source of inspiration, Tolkien was also fluent in Latin and Old English, and familiar with Greek, Welsh, and other ancient Germanic languages during his development of Quenya. Quenya lacks the front vowels *ä*, *ö*, and *y*, which are characteristic of Finnish, and it follows Latin-based stress rules that are entirely alien to Finnish. In Finnish, the accent is always on the first syllable, and the front vowels *ä*, *ö*, and *y* cannot occur in the same word with their back vowel equivalents *a*, *o*, and *u* (a phenomenon known as vowel harmony). The fact that Tolkien did not incorporate these two noticeable aspects of Finnish phonology into Quenya indicates his method of creating languages: he aimed for originality while ensuring his languages felt archaic and rich by being rooted in reality.

As shown in the results, the Finnish-tuned model struggled with adopting new stress patterns and lacked some Quenya-specific phonemes, particularly certain consonants and vowels like *ç*, *j*, and *ɛ*, leading to mispronunciations. In contrast, models fine-tuned from the multilingual dataset successfully addressed these issues. Stress is crucial in many of the languages included in the multilingual model, such as Spanish and Russian, allowing it to adeptly learn and generalize various stress patterns. Additionally, this model encompasses all the IPA symbols required for Quenya, for example, *nicht* [nɪçt] in German and *year* [jɪr] in English. The success of the multilingual model in this study highlights the importance of a broad linguistic foundation when developing speech synthesis systems for low-resourced languages. By leveraging the strengths of multiple languages, the multilingual model can overcome the limitations of single-language models and provide a more natural and accurate representation of speech. This finding suggests that future efforts in TTS development for constructed lan-

guages should consider the benefits of multilingual training to achieve higher quality and more authentic-sounding results.

## 6.2 Limitations

### 6.2.1 Prosody and Pronunciation Challenges

Although the model fine-tuned from a multilingual dataset outperformed the models fine-tuned from Finnish and English datasets in MOS scores, it still showed significant differences compared to real voice, indicating room for improvement in its naturalness. Enhancing the model’s prosody is particularly crucial for constructed languages like Quenya, which place a high emphasis on aesthetic appeal. In this project, the TTS model can only produce sentences with a standard reading tone, failing to mimic the distinctive prosodic patterns of Quenya poetry, thus impacting the speech’s naturalness. The short pauses between commas are too brief, making the speech sound rushed and unnatural. Additionally, the lack of stylized pronunciations further detracts from the authenticity. Despite the speaker’s distinct and stylized trilled **r** in the Quenya dataset, none of the models managed to produce a clear trilled **r**, sounding more like a tap **r** instead. Both proficient speakers agreed that a stylized trilled **r** is essential in Quenya, particularly in poetry. While it is true that in everyday spoken language, people often simplify the trilled **r** to a tap **r** for efficiency, such a simplification in Quenya, a language not primarily used for communication, diminishes its naturalness and aesthetic value. Tolkien’s own Quenya recordings also feature a very prominent trilled **r** that is slightly longer than in everyday speech.

### 6.2.2 Dataset Quality and Training Parameters

During the training of the Finnish model, although many transcription errors were manually corrected, time constraints prevented a complete cleansing of the CSS10 Finnish dataset. The training logs still showed some erroneous files were detected and skipped, which may have introduced potential issues affecting the overall performance of the Finnish model. Moreover, during the fine-tuning of the model, only the number of training steps was modified while default parameters were used; optimizing these parameters could potentially enhance training outcomes and achieve better performance.

## 6.3 Future Research

For all low-resource languages, the fundamental issue is the scarcity of data. Although improvements can be made by training models on limited datasets, as previously discussed, the outcomes are often not as effective as those achieved with larger, high-quality datasets.

Therefore, research should focus on how to collect more high-quality datasets recorded by professional Quenya speakers and on optimizing G2P scripts to handle pronunciation exceptions. Additionally, ToucanTTS has already provided capabilities for cloning prosody across speakers and research on TTS for poetry (Koch et al., 2022; Lux, Koch, & Vu, 2022a), which can be leveraged to further enhance the naturalness of Quenya synthetic speech.

Furthermore, it may be worthwhile to enhance the articulatory features in ToucanTTS to support phonemes currently not supported, such as  $\text{ɬ}$ , making Quenya pronunciation more precise. Various methods can be explored to optimize the model’s parameters to generate better speech. For instance, fine-tuning hyperparameters such as learning rate, batch size, and the number of training epochs can lead to significant improvements. Implementing advanced techniques like transfer learning with larger pre-trained models, or using data augmentation to artificially increase the size and variability of the training dataset, can also be beneficial.

Moreover, exploring which multilingual training setups yield better synthesis results for Quenya could be insightful. Perhaps using a multilingual model trained exclusively on Germanic and Uralic language families might produce higher-quality Quenya speech. This is because such a model would not need to generalize across languages that do not contribute phonetic or grammatical features relevant to Quenya, thus focusing more effectively on the specific characteristics needed. Limiting the training languages to those more closely related or influential could reduce the noise from irrelevant linguistic features and improve the model’s performance.

## 7 Conclusion

This study successfully developed a text-to-speech (TTS) system for Quenya using the IMS-Toucan, fine-tuning models from English, Finnish, and multilingual datasets. The results confirm that the Finnish-tuned model performs better than the English-tuned model, while the multilingual model outperforms both, demonstrating the advantage of diverse linguistic features. Despite these successes, challenges remain, particularly with certain Quenya-specific phonemes and stress patterns. Evaluation using Mean Opinion Scores (MOS) and detailed listening reports from Quenya enthusiasts and linguistic experts provided a thorough assessment, highlighting areas for improvement. Future research should focus on higher-quality datasets, optimizing training parameters, and refining prosody to match Quenya’s unique aesthetic. This study not only provides a functional TTS system for Quenya but also offers valuable insights for developing TTS systems for other constructed, low-resource, and even linguistically reconstructed extinct languages, emphasizing the significance of cross-lingual transfer learning and multilingual models.

## References

- Adams, M. (2011). *From elvish to klingon: exploring invented languages*. Oxford University Press, USA.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Bak, T., Lee, J., Bae, H., Yang, J., Bae, J.-S., & Joo, Y.-S. (2023). Avocodo: Generative adversarial network for artifact-free vocoder. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 12562–12570).
- Barlow, J. A., & Gierut, J. A. (2002). Minimal pair approaches to phonological remediation. In *Seminars in speech and language* (Vol. 23, pp. 057–068).
- Barnes, L., & Van Heerden, C. (2006). Virtual languages in science fiction and fantasy literature. *Language Matters: Studies in the Languages of Southern Africa*, 37(1), 102–117.
- Bett, S. (2002). The number of phonemes in english. *Memory of Ken Ives (1917–2002)*, 30(1).
- Carpenter, H., et al. (1981). *The letters of jrr tolkien* (Vol. 140). Boston: Houghton Mifflin.
- Christoph, K. G. (2012). Esperanto and minority languages: A sociolinguistic comparison. *Language Problems and Language Planning*, 36(2), 167–181.
- Chung, Y.-A., Wang, Y., Hsu, W.-N., Zhang, Y., & Skerry-Ryan, R. (2019). Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6940–6944).
- Destruel, M. (2016). *Reality in fantasy: linguistic analysis of fictional languages* (Unpublished doctoral dissertation). Boston College.
- Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2021). A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages. *Interspeech 2021*, 16–20.
- Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2022). Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In *Proceedings of the 1st annual meeting of the elra/isca special interest group on under-resourced languages* (pp. 16–22).
- Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2023). The effects of input type and pronunciation dictionary usage in transfer learning for low-resource text-to-speech. *arXiv preprint arXiv:2306.00535*.
- Donahue, J., Dieleman, S., Bińkowski, M., Elsen, E., & Simonyan, K. (2020). End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*.

- Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Skerry-Ryan, R., & Wu, Y. (2021). Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*.
- Forster, P. G. (1982). *The esperanto movement* (No. 32). Walter de Gruyter.
- Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *International conference on machine learning* (pp. 1243–1252).
- Gobbo, F. (2017). Are planned languages less complex than natural languages? *Language Sciences*, 60, 36–52.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (pp. 369–376).
- Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2), 236–243.
- Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... others (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.
- Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., ... others (2021). Recent developments on espnet toolkit boosted by conformer. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5874–5878).
- Gutkin, A. (2017). Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages. In *Interspeech* (pp. 2183–2187).
- Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., ... Tan, X. (2020). Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7654–7658).
- Higley, S. (2007). *Hildegard of bingen’s unknown language: An edition, translation, and discussion*. Springer.
- Hirose, K., & Tao, J. (2015). *Speech prosody in speech synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer.
- Ivnova, V. (2023, August). *Synthesising proto-indo-european using phonological features for zero-shot synthesis*. Retrieved from <https://campus-fryslan.studenttheses.ub.rug.nl/371/>

- Janton, P. (1993). *Esperanto: Language, literature, and community*. Suny Press.
- Jokisch, O., & Eichner, M. (2000). Synthesizing and evaluating an artificial language: Klingon. *parameters*, 4, 5.
- Kazimierczak, K. (2010). Adapting shakespeare for "star trek" and "star trek" for shakespeare: "the klingon hamlet" and the spaces of translation. *Studies in Popular Culture*, 32(2), 35–55.
- Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33, 8067–8077.
- Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning* (pp. 5530–5540).
- Koch, J., Lux, F., Schauffler, N., Bernhart, T., Dieterle, F., Kuhn, J., . . . Thang Vu, N. (2022). PoeticTTS - Controllable Poetry Reading for Literary Studies. In *Proc. interspeech 2022* (pp. 1223–1227). doi: 10.21437/Interspeech.2022-10841
- Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33, 17022–17033.
- Krägeloh, C., & Neha, T. N. (2014). Lexical expansion and terminological planning in indigenous and planned languages: Comparisons between te reo māori and esperanto. *Language Problems and Language Planning*, 38(1), 59–86.
- Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., . . . Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.
- Łańcucki, A. (2021). Fastpitch: Parallel text-to-speech with pitch prediction. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6588–6592).
- Lee, N. H. (2020). The status of endangered contact languages of the world. *Annual Review of Linguistics*, 6, 301–318.
- Lee, S.-g., Ping, W., Ginsburg, B., Catanzaro, B., & Yoon, S. (2022). Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.
- Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019). Neural speech synthesis with transformer network. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 6706–6713).
- Lindstedt, J. (2009). *Esperanto—an east european contact language?* na.
- Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., . . . Vu, N. T. (2023). The ims toucan system for the blizzard challenge 2023. *arXiv preprint arXiv:2310.17499*.

- Lux, F., Koch, J., & Vu, N. T. (2022a). Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech. In *Proc. IEEE SLT*.
- Lux, F., Koch, J., & Vu, N. T. (2022b). Low-resource multilingual and zero-shot multispeaker tts. *arXiv preprint arXiv:2210.12223*.
- Lux, F., & Vu, N. T. (2022). Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. *arXiv preprint arXiv:2203.03191*.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kald. In *Interspeech* (Vol. 2017, pp. 498–502).
- Mu, Z., Yang, X., & Dong, Y. (2021). Review of end-to-end speech synthesis technology based on deep learning. *arXiv preprint arXiv:2104.09995*.
- Okrand, M. (1992). *The klingon dictionary: the official guide to klingon words and phrases*. Simon and Schuster.
- Okrand, M. (1996). *The klingon way: A warrior’s guide*. New York: Pocket Books.
- Okrand, M. (1997). *Klingon for the galactic traveler*. New York: Pocket Books.
- Okrand, M., Adams, M., Hendriks-Hermans, J., & Kroon, S. (2011). Wild and whirling words. *From Elvish to Klingon: exploring invented languages*, 111.
- Okrent, A. (2009). In the land of invented languages, new york, spiegel & grau. *En ligne: <http://inthelandofinventedlanguages.com>*.
- Park, K., & Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.
- Perälä, H. (2002). «are high elves finno-ugric? *Dostupné z*.
- Peterson, D. J. (2015). *The art of language invention: From horse-lords to dark elves to sand worms, the words behind world-building*. Penguin.
- Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., . . . Miller, J. (2018). Deep voice 3: 2000-speaker neural text-to-speech. *proc. ICLR*, 214–217.
- Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 3617–3621).
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.
- Ren, Y., Liu, J., & Zhao, Z. (2021). Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems*, 34, 13963–13974.
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.
- Romaine, S. (2011). Revitalized languages as invented languages. *From Elvish to Klingon:*



- Exploring Invented Languages*, 185–225.
- Schreyer, C. (2011). Media, information technology, and language planning: what can endangered language communities learn from created language communities? *Current Issues in Language Planning*, 12(3), 403–425.
- Schreyer, C. (2015). The digital fandom of na’vi speakers. *Transformative Works and Cultures*, 18.
- Schreyer, C. (2021a). Artificial languages. *See Stanlaw*.
- Schreyer, C. (2021b). Constructed languages. *Annual Review of Anthropology*, 50, 327–344.
- Schubert, K. (2011). *Interlinguistics: Aspects of the science of planned languages* (Vol. 42). Walter de Gruyter.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... others (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779–4783).
- Suni, A. S., Aalto, D., Raitio, T., Alku, P., & Vainio, M. (2013). Wavelets for intonation modeling in hmm speech synthesis. In *8th ISCA speech synthesis workshop* (pp. 285–290).
- Sutrave, N. (2017). Hol sarmey qed qulwi’ghitlh: A typological analysis of klingon.
- Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.
- Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.
- Tikka, P. (2007). The finnicization of queya. *URL: <https://www.oocities.org/petristikka/elvish/tikka.pdf> (accessed: 22.12. 2020)*.
- Tolkien, J. R. R. (1992). *The book of lost tales: Part two* (Vol. 2). Del Rey.
- Tonkin, H. (2015). Language planning and planned languages: How can planned languages inform language planning? *Interdisciplinary Description of Complex Systems: INDECS*, 13(2), 193–199.
- Tu, T., Chen, Y.-J., Yeh, C.-c., & Lee, H.-Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.
- Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... others (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.
- van Oostendorp, M. (2019). Language contact and constructed languages. *Handbook of language contact*, 124–135.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... others

- (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*, 164.
- Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ... others (2018). Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3, 1–40.
- Wells, D., & Richmond, K. (2021). Cross-lingual transfer of phonological features for low-resource speech synthesis. In *Proceedings of the 11th speech synthesis workshop, budapest, hungary* (pp. 160–165).
- Windsor, J. W., & Stewart, R. (2017). Can unnatural stress patterns be learned: new evidence from klingon. In *Actes du congrès annuel de l'association canadienne de linguistique 2017/proceedings of the 2017 annual conference of the canadian linguistic association*.
- Wu, Y., Tan, X., Li, B., He, L., Zhao, S., Song, R., ... Liu, T.-Y. (2022). Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *arXiv preprint arXiv:2204.00436*.
- Yamamoto, R., Song, E., & Kim, J.-M. (2020). Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Icassp 2020-2020 iee international conference on acoustics, speech and signal processing (icassp)* (pp. 6199–6203).
- Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., & Szczepaniak, P. (2016). Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. *arXiv preprint arXiv:1606.06061*.

# Appendices

## Audio Demonstrations and Source Code

Pre-generated audio samples of the Quenya TTS system are available at

<https://annie-zhou1997.github.io/QuenyaTTS.github.io/>

This site includes all the Quenya audio files and transcribed texts used for the MOS testing.

Interactive demo on Hugging Face:


<https://huggingface.co/spaces/AnnieZzz/Quenya-TTS>

For source code and detailed project documentation, visit the

<https://github.com/Annie-Zhou1997/Quenya-TTS>

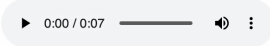
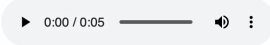
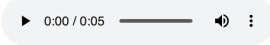
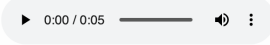
## Quenya MOS Survey Sample

Below is an example of the Mean Opinion Score (MOS) survey used in this study. The survey consists of 9 test sentences, each presented in the same format. Completing the survey typically takes about 15 minutes.

 university of  
 groningen

Et Eärello Endoreнна utúlien, Sinome maruvan ar Hildinyar, tenn'  
 Ambar-metta!

'et ear'ello endor'enna ut'u:lien, s'inome m'aruvan 'ar hild'ijar, tenn  
'ambar m'etta!

	Bad	Poor	Fair	Good	Excellent
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Next](#)

Figure 5: Quenya MOS Survey Sample