university of groningen / campus fryslân

Synthesising Proto-Indo-European using Phonological Features for Zero-Shot Synthesis

Victoria Ivanova

university of groningen / campus fryslân

University of Groningen

Synthesising Proto-Indo-European using Phonological Features for Zero-Shot Synthesis

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Assoc. Prof. Dr. Matt Coler (Voice Technology, University of Groningen)
and
Ph.D. Candidate Phat Do (Voice Technology, University of Groningen)

Victoria Ivanova (s5382726)

August 23, 2023

Contents

Acknowledgments

I am thankful to my supervisors Phat and Matt for the input and for the kind words and understanding during the process. Thank you to the teachers of the Voice Technology programme who taught me so much in one year. Thanks to the study advisor Hieke Hoekstra for the support.
Thank you, Ellemijn for being a great thesis-writing partner and a great friend!
Above all, thank you to my family whose love and support I always feel. And to Yannick for being the best and to my cat Ladefoged for being the cutest.

Abstract

Proto-Indo-European is a reconstructed language, from which the biggest language family, Indo-European, evolved. Linguists have reconstructed its phonology through the comparative method, analysing cognate words in its daughter languages. Some attempts at automation this process have been made, but fewer have attempted to take it a step further and synthesise its sound. This task could be seen as a zero-shot synthesis problem, meaning that it needs to be synthesised without any training data for a model to learn from.

We ask whether it is possible for this task to be achieved through the means of zero-shot synthesis using phonological features as input. Models utilizing this technique have been shown to produce successfully unseen languages in code-switching tasks and even synthesising unseen phonemes. We opt to use the IMS-Toucan toolkit, which is mostly build upon the FastSpeech2 architecture, with some additions, such as the use of the LAML optimizing framework. The toolkit is modular and we can modify its text-processing and phonemization modules to handle Proto-Indo-European input. Further, we fine-tune the multilingual model on Abkhaz, which has some similar features to Proto-Indo-European.

Our results find that our method improves significantly the naturalness of the synthesised speech in comparison to previous attempts at synthesising Proto-Indo-European, but the fine-tuning yields no significant improvement over the pre-trained model.

The user-friendly web app that we built is a useful tool for education or entertainment purposes. What is more, we believe that our system could be beneficial to language revitalization tasks and combined with other methods for automation of the reconstruction process, it could lead to better success in the efforts of keeping the languages of the world alive.

1 Introduction

Through pain-staking efforts, linguists have been able to trace back the sound changes that have occurred in the languages of the Indo-European language family and reconstruct the phonology of their ancestor – Proto-Indo-European. This is usually done by comparing cognate words across related languages, some of them also already extinct, and establishing the differences between them. On this basis, different sound change laws are formulated that allow us to back-track the changes in words through time and space. More recently, it has been shown that the manual reconstruction of words can be automated, using probabilistic models of sound change and alignment analysis (Bouchard-Côté et al., 2013; List et al., 2022).

These advancements pave the way for technology to enter the conservative and traditional world of language reconstruction to help with other tasks as well. Building upon the foundations of reconstruction efforts, this study will endeavor to develop a model that can generate speech in Proto-Indo-European from text input, thereby enabling us to hear what this language may have sounded like 6000 thousand years ago. Our research question is whether we can utilize a neural-network based approach to successfully synthesise the language from just text input in the traditional notation style.

Recent attempts at this task have already been made by Donnelly (2022), which used simple concatenation to synthesise Proto-Indo-European. The concatenated sounds were phonemes generated by the espeak-NG software. This is necessary due to the complete lack of audio data from this ancient language, as the people who spoke it lived approximately 6000 years ago and no recordings of their speech have survived to the present day. We intend to approach the task differently, namely as a zero-shot synthesis project. Such projects usually aim to synthesise contemporary languages with no resources, as there is no training data to train or fine-tune a traditional synthesis machine learning model. It could also be used for code-switching purposes, such as in Staib et al. (2020), where the authors synthesise code switching from English to German, which was previously unseen by the model. We hypothesise that using a zero-shot neural network technique will lead to considerable improvement upon the simpler concatenation methodology.

To realise our goal we intend to use a toolkit, presented first in Lux et al. (2021) that enables us to train the model to map phonological features, extracted from the IPA transcription of the audio data, to acoustic features. The phonological features of the phones, such as “vowel”, “openness”, etc., are language-independent, as they are mostly based on the way all humans use their articulators to form the sounds. This means that articulatory features data can be used across languages. Once the network has learned what the features “sound” like, it is able to synthesise any IPA-transcribed test, even if it is from a previously unseen language.

Further, even though the model we have chosen is capable of producing even unseen phonemes, we opt for fine-tuning a multilingual model, mostly trained on data from European languages, on data from the Caucasian language Abkhaz. We hypothesise that this would lead to better results, due to the exposure of the model to specific phonological features found both in Proto-Indo-European and Abkhaz.

A major challenge when working with an extinct language is that there are no native speakers

to evaluate the synthesised speech. We opt for a traditional MOS evaluation, which however focuses only on the naturalness of the speech, since the participants will not be able to judge the intelligibility.
We expect our work to be beneficial for education purposes, as students who study Proto-Indo-European do not have any auditory frame of reference for the language and rely on only learning it on paper. Furthermore, we aim for our tool to be user friendly and allow a wider public to learn about our history through discovering an extinct language, which is inevitably intertwined with ancient culture. What is perhaps even more valuable is that developing zero-shot synthesis techniques could be of great importance to language revitalization tasks, where there are extremely limited (or no) resources of a language, important to a community. Our approach can be combined with other technological advances in the field of language reconstruction and hopefully provide language conservationists with more tools when combating the rapidly disappearing languages of the world.

2 Thesis Outline

The next section of this thesis will discuss in length how the words we will synthesise were reconstructed and how our project fits in the newly expanding world of Proto-Indo-European reconstruction. Further, we will discuss how zero-shot synthesis has come to be and how it is realised and used in the state-of-the-art projects. Additionally, we present the TTS toolkit that we use. In the Methodology section we will discuss the modifications made to the toolkit that allow us to process Abkhaz and Proto-Indo-European, fine tune the model and build a web app that lets users easily execute inference on the models. In the Evaluation section we will outline how we set up the evaluation procedure and obtain the opinions of our evaluators. In Results we will discuss the findings of the evaluation procedure and we will analyse and put them in the context of our research question and hypothesis in the Discussion. Final conclusions are draws in the Conclusions section.

3 Background Literature

3.1 Proto-Indo European reconstruction

The beginnings of Indo-European comparative linguistics came with the discovery that Sanskrit is related to languages such as Greek, Latin and German when trade routes to India were opened around the beginning of the 16th century. The word forms found in the ancient Hindu scripts, the Vedas, that date to 1000 B.C. are less obscured by sound changes than those in Greek (Beekes, 2011), and it was famously noted by Sir William Johnson in 1786 that these three languages must stem from a possibly extinct common predecessor. Clackson (2007) notes that the real Indo-European parent language that was spoken is not the same as the language that we have reconstructed through comparative techniques. The words, syntax and grammar we are able to reconstruct may actually belong to different times or places where the Indo-European parent language was spoken, as it was always evolving and changing, just like any modern language. Thus, time and space are collapsed when we discuss Proto-Indo-European and we accept that some reconstructed forms might be much older than others, for example.

Nonetheless, the reconstructed Proto-Indo-European is able to reveal a lot about the ancient people who spoke it.

For example, by collecting all contemporary and old recorded words for wheel from the Indo-European languages, we find that they come from a common source – Proto-Indo-European. This shows us that they had a native word for wheel and therefore it was something present in their society. Exactly this method of comparing cognate, or etymologically related, words is in the heart of comparative linguistics.

Take for example the cognate set of words of English wolf – there is vílkas in Lithuanian, wulfs in the extinct but recorded Gothic, lúkos in Ancient Greek, lupus in Latin, vrka- in Sanskrit. Besides the differing position of the vowel and the liquid l or r (a change called metathesis), the words are phonologically similar and the meaning in all those languages is the same. These two conditions qualify them as cognate words. The analysis of this set has to take into account many details. Based on most of the forms and on what rules we know already, we can puzzle out that the Proto-Indo-European form could have been *wl̥k^{w}os[1]. However, this means that the word evolved unexpectedly in Latin and Greek as a syllabic l in PIE turned into al and ol instead. It has been suggested that the unexpected changes might be due to altering the word in fear of uttering the actual name of this scary animal, or to avoid it sounding similar to the words in those languages for fox.

Such is the process of reconstruction, a giant puzzle game with many rules that all need to be observed and yet, many exceptions that need to be accounted for. It is no surprise that people have endeavored to automatize the process of reconstruction, not only for PIE but for other proto-languages as well.

Often times automatization of the process starts with efforts to detect cognate words. Some algorithms which tackle this task rely on calculating pair-wise string alignment, a measure that tells us how similar strings are. Wieling et al. (2009) evaluates different algorithms that have been used for the purpose, such as the Levenshtein algorithm as well as the Pair Hidden Markov Model. List et al. (2022) further employs the pair-wise string alignment computation to automatically detect cognates in word lists, using language-specific scoring schemes.

Other tackle the process by first working on establishing the phylogenetic relation between languages, such as Bouckaert et al. (2012), which uses Bayesian phylogeographic approaches to model the expansion of the Indo-European family. Bouchard-Côté et al. (2013) develop a probabilistic model of sound change and a Monte Carlo inference algorithm, which they use to reconstruct word forms in Proto-Austranesian. They note that if their method is extended to jointly infer phylogenetic relations and cognate sets it could reduce the circularity of reasoning occurring in traditional reconstruction – relations between languages are based on the existence of cognate words between languages but the decision whether words are cognate or not is often motivated by considerations about relations between languages.

[1]The PIE diacritics notation differs from IPA in many ways, for example here the subscript ring means syllabic instead of devoiced. The asterisk before the word signifies that it is reconstructed.

3.2 PIE Phonology

In order to use articulatory features to synthesise Proto-Indo-European, we need to have a clear view of its phonology. Even though there is more or less a consensus on what sounds must have been present in Proto-Indo-European in order to observe all the sounds they evolved in in the daughter languages, there are still some areas of dispute. What is more, under the comparative linguistics framework even if there is a high level of accurateness in the reconstruction of the phonological form of the Proto-Indo-European words, the phonetic form can still be only approximated (Beekes, 2011). As Bičovský (2021) details, the abstract reconstructed phonological system of Proto-Indo-European is likely to be influenced by expectations of symmetry, an example of which is early reconstructions Brugmann and Delbrück (1967) suggesting that the sibilant series would follow the example of the plosives and have the members *s, *s^h, *z, *z^h with no support from data. Due to factors such as varying air pressure, muscle tension and execution of articulation gestures, the phonetic realisations of the phonological abstractions is never as symmetrical and precise. Figure 1 showcases the Proto-Indo-European phonology as represented by the currently used notation (Beekes, 2011).

	Traditional reconstruction			
CONSONANTS	occlusives/stops			
	labials	p	b	b^h
	dentals	t	d	d^h
	palatals	$\acute{k}$	$\acute{g}$	$\acute{g}^h$
	velars?	k	g	g^h
	labiovelars	k^w	g^w	g^{wh}
	fricatives	s		
	laryngeals	h_1	h_2	h_3
	sonants			
	liquids	r	l	
	nasals	m	n	
	semivowels	i	u	
VOWELS		e	o	
		$\bar{e}$	$\bar{o}$	

Figure 1: The phonology of PIE in the current notation as per (Beekes, 2011)

3.2.1 “Laryngeals"

For the purpose of this project it is necessary to assume phonetic realisations of the reconstructed phonological forms. Multiple theories have been brought forward through the decades about the phonetic realisations of the different series of reconstructed phonemes. Perhaps the most puzzling remains the question of the so-called “laryngeals". They are marked as h_1, h_2 and h_3. At first it was proposed by De Saussure (1879) that there were three vocalic elements, represented by *E, *A and *O in his notation, based on surprising behaviour of some Greek

verb paradigms, that could not be explained with the then-known sound change laws. Later on Cuny (1912) proclaimed them as “laryngeal", which remained a misnomer, despite modern theories not agreeing. The fact that when they appeared in interconsonantal position, they evolved into vowels in the daughter languages fed the idea that they were vowel-like. Reynolds et al. (Reynolds et al., 2000) analyse the matter through the perspective of generative phonology and mora representations and concludes that they could have been weak metric vowels.

In parallel, with the discovery that the ancient Anatolian language Hittite is part of the Indo-European family, it was also found that in the Hittite cognate words where the “laryngeals" were expected, we could find fricatives (Jasanoff, 2017). This became the basis of the fricative hypothesis, which is widely supported. According to it, the three laryngeals were the following sounds:

(1) *ʔ - glottal stop
*ʕ - pharyngeal fricative
*ʕʷ - labialized pharyngeal fricative

More recently, Hartmann (2021) demonstrates a less traditional approach to the topic by investigating the use of coarticulatory and statistical constraint effects governing the “laryngeals". This is done by setting up a deep neural network for each phonological feature that the sounds may exhibit. Each network was trained to detect the particular feature and to predict its presence or absence for unseen sound-environment data. The results of this study suggest the following values for the “laryngeals" based on what phonotactics the networks learned from the training data:

(2) *xʷ → ɦʷ – labialized velar fricative, later turned into labialized glottal fricative
*ʕ - pharyngeal fricative
*ɣʷ → ʁʷ – labialized voiced velar fricative, later turned into labialised voiced uvular fricative

However, despite this and other recent attempts to figure out the identity of the “laryngeals", most scholars still prefer the values of Example 1. We choose to stick with the more widely accepted theories and save the option of adjusting the phonology with which you synthesise on the fly for future expansions of this study.

3.2.2 Stops

The stop series of Proto-Indo-European also pose some questions in terms of their phonetic values, in particular the velar series. As seen in Figure 1, there are three different velar series, which is typologically and etymologically unlikely. So instead it was proposed by Gamkrelidze and Ivanov (1984) that the voiced stops were in fact glottalized, which is common in Caucasian languages. This theory suffered a lot of critique, to which the authors responded in Gamkrelidze (1989), and is still not widely accepted, despite Beekes (2011) stating that it is the more modern interpretation of the stops, as aforementioned, we choose to focus only on the traditional phonology of Proto-Indo-European and thus we will not interpret them as the Glottalic theory suggests.

3.2.3 Vowels

The vowel system of Proto-Indo-European is rather limited with only /e/ and /o/ as true vowels. There are also the semi-vowels /i/ and /u/, which are sometimes interpreted as full vowels, especially if they are in a nucleus position. Proto-Indo-European also exhibits vowel lengthening, thus there are also long variants of /e/ and /o/, but Beekes (2011) argues that this was probably only due to phonotactics.
It is quite remarkable that /a/ is lacking from the phonemic inventory and it is also a polarizing question to some. In general, it is assumed that whenever an /a/ appeared in a daughter language, it was mostly due to *h_2"coloring" an /e/ vowel. The main evidence for this is how Proto-Indo-European *o lengthens to lengthened 'a' in open syllables in Sanskrit, except from when the subsequent syllable contains a vowel that corresponds to /a/ in other daughter languages. It is suggested that this is possible because to the expected 'a' vowel is in reality a CV syllable *h_2e, which alters the syllabic structure and renders the previous syllable closed.
Overall, as stated we choose to follow a neutral path and stick to the traditional understanding of the PIE sounds. The following sections further delve into how having such extensive knowledge of possible reconstructions has inspired people to try and give this language a voice and how we intent to approach it.

3.3 Previous PIE synthesis attempts

Linguists have been long interested in bringing the sound of Proto-Indo-European to life. There are some short stories that are comprised of reconstructed words and follow reconstructed morphological and semantic rules that can be read aloud to showcase the sound of Proto-Indo-European depending on the different reconstruction theories. Such two stories are the "The Sheep and the Horses" ("H_2owis $h_1$ek̀wōs-k^{w}e") and "The King and the god" ("H_3rēk̀s deywos-k^{w}e") (Mallory & Adams, 1997).

More recently, Kloekhorst (2020) developed a mobile application named "Vanished Voices" which allows the user to listen to many utterance read by an expert in Proto-Indo-European and multiple other ancient languages, often in multiple reconstructed chronological stages. The app also provides information about the etymological evolution of the phrases.

There have been few efforts in bringing Proto-Indo-European and extinct languages in general to life through speech synthesis. That is not surprising, as there are multiple challenges that render standard TTS approaches unsuccessful. The over-arching challenge is the lack of audio data and the special nature of the reconstructed text data. Donnelly (2022) uses a concatenative synthesis approach to the challenge.

Firstly, in order to generate the segments which their model concatenates, they use the espeak-NG tool [2], which supports formant synthesis of 127 languages. The authors limit the scope to a 100 languages by mostly retaining the Indo-European languages of the selection and additionally some non-related languages to broaden the phoneme inventory they create. For each of the languages the authors comprise a Swadesh list – a list with a 100 words per language which are most often etymologically native to the language, such as words for nature elements, simple verbs and common animals. This is done with the intention to capture the phonemes native

[2] https://github.com/rhdunn/espeak

to the language and thus accumulate a good representation of the phonological inventory of the selected languages. Consequently, these words are synthesised for each language using the Praat software(Boersma, 2011) which works in combination with espeak-ng to produce .wav files.

At the time of inference, the user is able to input Proto-Indo-European in traditional notation and is able to choose what phonetic value will be assigned to each of the phonological symbols. The system then looks for matching tri-grams, bi-grams or, in the worst case, uni-grams in its inventory of synthesised phones. They are then concatenated and silence is added at word and sentence boundaries.

This approach results in a relatively flexible system but the naturalness is not very high due to the substantial size of the concatenated phone units employed and the simplistic manner in which they are combined, retaining traces from the phonetic environments they were taken from. We believe that taking a neural network based approach will in any case improve on the naturalness of the synthesised speech. However, neural network models typically need to be trained on data, which we have a complete lack of, which makes the task more intricate. Synthesising an unseen language with no available training data is referred to as zero-shot synthesis. The next section delves deeper into zero-shot methods.

3.4 Zero-shot synthesis

The question of managing code-switching is often entangled with that of zero-shot synthesis, as it calls for a system that is able to handle a different unseen phonology unexpectedly, and many zero-shot solutions are geared specifically towards this problem. Chiu et al. (2022) demonstrate the capabilities of a multi-language model to perform code-switching, however, not to a previously unseen language. This means that there is still a high requirement for training data, the lack of which is often the need for zero-shot capable systems. Transfer learning is a popular method when it comes to low-resource language solutions, which comes with challenges, such as the different sets of phonemes required to produce languages (Lux & Vu, 2022). Different solutions that aim to help map the knowledge of the source language to the target language input have been presented, such as the Phonetic Transformation Network, which utilizes a language recognition system to map orthographic symbols of different languages to their sounds (Tu et al., 2019). Do et al. (2022) proposes the joint use of Angular Similarity of Phoneme Frequencies to find the best possible source languages, and the use of phoneme mapping to improve on typical transfer learning. They suggest to bypass the issues of unseen phonemes by mapping them to the closest phoneme based on their phonological features.

A possible solution to avoid mapping issues is to execute joint learning and to include rich-resource languages and the low-resource languages in the training set, which is not an easy undertaking (Azizah et al., 2020; Xu et al., 2020). This however, comes with complex training procedures (Lux & Vu, 2022). Another approach to the input mapping problem is to simply alter the type of input that the model takes. Li et al. (2019) to use Unicode characters as input to Tacotron 2, but this does not tackle the issued of unseen phonemes and still requires alterations for new languages.

The heart of this problem is that each and every language of the world has a unique phono-

logical system and most have a unique orthography. In order to move towards a direction that leads to a more language-agnostic model, we need to convert the input to language-agnostic information. Using phonological or articulatory features information about the sounds and having the network learn the connections between these features and acoustic realisations is a possible approach. Staib et al. (2020) use this approach in combination with Tacotron 2 and the Griffin-Lim vocoder and are able to synthesis German code-switching with moderate success.

Lux and Vu (2022) takes this idea further by expanding it with the use of MAML (Model-Agnostic Meta Learning) and creating a highly modular, flexible toolkit intended for even less experienced users to employ it. We find that this toolkit suits our task very well as it requires only information about the phonology of the language and an adequate way of handling it and it is able to synthesise speech swiftly and successfully. In the next section will go into more detail about how the toolkit is built and how the system works.

3.4.1 IMS-Toucan Toolkit

This toolkit was created for the purpose of the 2021 Blizzard challenge which contained testing data with code-switching. It is largely based on the ESPNet toolkit (Watanabe et al., 2018). It is highly modular as their overall approach is meant to be compatible with different software. In Lux et al. (2021) they test different candidates for the system and eventually choose to go forward with an implementation of FastSpeech2 at the heart of the pipeline.

The main goal of FastSpeech (Ren et al., 2019) and FastSpeech2 (Ren et al., 2020) is to speed up the inference process by employing parallel mel-spectrogram generation, which differs from auoregressive models that generate each spectrogram, conditioned on the previous. FastSpeech utilizes a sperately trained teeacher model to obtain the correct durations for the phonemes. This is necessary due to the parallel synthesis model and the general one-to-many mapping problem, which relates to the many possible variations of durations of the phonemes.

However, the teacher model prooves to somewhat lead to inaccurate durations and sometimes leads to repetitions or skipping phonemes. FastSpeech2 handles this problem by training the model with actual ground truth durations, extracted from speech instead of the predictions from the teacher model. Further, FastSpeech2 includes more controlability of pitch and intonation.

The IMS-Toucan toolkit uses an Aligner (Rapp, 1995) to force-align the predicted phoneme sequences and then map the phonemic inventory of the Aligner to that of the phonemizer to get the necessary durations. They extract pitch and ton information and average this information over all the specific phoneme's spectrograms.

In the initial version of the toolkit, presented in Lux et al. (2021), the authors still rely on phoneme mapping and tackle the code-switching problem by reducing it to mapping unknown phones to similar ones. The later version of the toolkit that we utilize, uses phonological/articulatory features input, as aforementioned, as presented in Lux and Vu (2022).
The current IMS-Toucan uses two different ways to vectorize the phones using their features – PanPhon and Papercup (Mortensen et al., 2016; Staib et al., 2020). PanPhon's feature set

is comprised of 22 features, such as syllabic, voice, delayed release and more. Amongst the 10 multi-level features of the Papercup set we can find vowel roundness, voicing, consonant place and others. The role of these description systems is to encode the phones into numeric vectors, which are then fed into a fully connected layer, which is able to learn complex relationships due to its density and is expected to be able to learn the relations between the features autonomously. The aforementioned encoding function takes care of projecting the feature vectors into a 512 dimensional space.

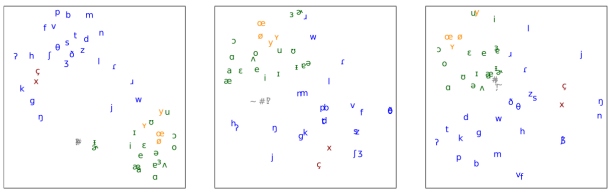
Another update on the toolkit is the use of the LAML – Language-Agnostic Meta Learning, which is a variation of MAML (Finn et al., 2017). This is a framework that allows the speeding up of the learning of new tasks when little data is available and it centers around optimizing the initialization weights of the model. The authors find that the MAML is not applicable to the tasks of learning a new language and a new speaker as it becomes very unstable. They adjust it by using it to calculate the loss per language and then updating the meta model directly using the Adam optimizer (Kingma & Ba, 2014). This simplifies the framework and stabilizes it.
After the model has been trained and is ready for inference, the input text is treated with the phonemizer and spectrograms are produced. The GitHub repositoryGitHub repository[3] where the code for this toolkit is freely available, indicates that the toolkit is currently built to work with the new Avocodo vocoder.
Overall, this toolkit provides an easy and adequate framework for us to build upon in order to reach out goal of synthesising Proto-Indo-European. In the next subsection, we describe why we think that part of our methodology should be to fine-tune a multilingual model on Abkhaz data.

3.5 Fine-tuning with Abkhaz

As shown in Lux and Vu (2022), the multilingual ToucanTTS model can generalize well into unseen phylogenetic branches and even synthesise unseen phonemes (see Figure 2), which implies that it is able to handle different phonologies well. However, in order to improve the chances of the synthesised Proto-Indo-European speech sounding intelligible and natural, we choose to fine-tune the model mostly trained on European languages with a small amount of data from a language that shares a lot of phonological traits with the reconstruction of Proto-Indo-European that we will follow for this project. This language is Abkhaz, spoken in the partially recognized Republic of Abkhazia, situated on the Caucasian Black Sea Coast. It is remarkable for its very limited vowel inventory – only two phonemic vowels. In contrast, there are 59 (Standard Abkhaz) and up to a 100 consonants, depending on the dialect and if we take into account geminatated consonants (Wier, 2005). Its three way contrast in its stops series between voiceless aspirated, voiced and ejective (glottalized) is reminiscent of what the Glottalic theory (as mentioned in Subsection 3.2.2) suggests for the stops series of Proto-Indo-European, namely a three-way contrast of voiceless (long), voiceless ejective (glottalized) and normal voiceless. In fact, Beekes (2011) specifically mentions that the Glottalic theory is more typologically likely than the standard reconstruction as a similar three-way contrast is seen in come Caucasian languages.

Further, the secondary phonological features of Abkhaz also overlap with those of Proto-Indo-European to some extent, namely labialization and palatalization.

[3]GitHub repository



(a) 66 dimensional Featurespace (24 PanPhon and 42 Papercup features)

(b) 512 dimensional embeddingspace learned during Tacotron 2 training

(c) 384 dimensional embeddingspace learned during FastSpeech 2 training

Figure 2: Generalization over unseen phones in (Lux & Vu, 2022)

Overall, we expect that it would be beneficial for the model to be exposed to the Abkhaz data that contains a rich variety of phonemes that are similar to some more topologically rare Proto-Indo-European reconstructions.

4 Methods

This section details the steps that we took to modify the IMS-Toucan toolkit to be compatible with the Abkhaz fine-tuning data and the Proto-Indo-European text input at inference time. This includes data restructuring, changes to the text pre-processing and phonemization modules of the pipeline, creating an appropriate fine-tuning pipeline and adjustments to the inference process. Additionally, we created a user-friendly web-based app for easy inference using the pre-trained and fine-tuned mosdules.

4.1 Abkhaz data pre-processing

The first focus of our methodology is the fine-tuning of the provided multilingual pre-trained model and that begins with ensuring our data is compatible with the system. The public GitHub repository, which contains all relevant code to this project, also contains a Python notebook (*abkhaz_data_processing*) which executes all of the below-described steps, including the phonemization.

Firstly, the Common Voice data needs to be prepared as these datasets are not typically meant to be used for TTS tasks. The format is thus not readily compatible with IMS-Toucan and requires restructuring. We convert the format to this of the popular dataset IJSpeech, which is defined by a folder, filled with .wav files, and a folder, filled by the corresponding .txt files containing the annotations. The Abkhaz data consists of a folder with audio files, which are the sentences recorded by the users, and a spreadsheet that contains the meta-information for each sentence and its annotation. Before the restructuring, we filter the spreadsheet so that we are left with information only about the utterances spoken by the most prolific speaker. This speaker, a woman in her fifties, recorded 1039 sentences. This constitutes about 5% of the training set of the data, or approximately 88 minutes, as illustrated in Figure 3.

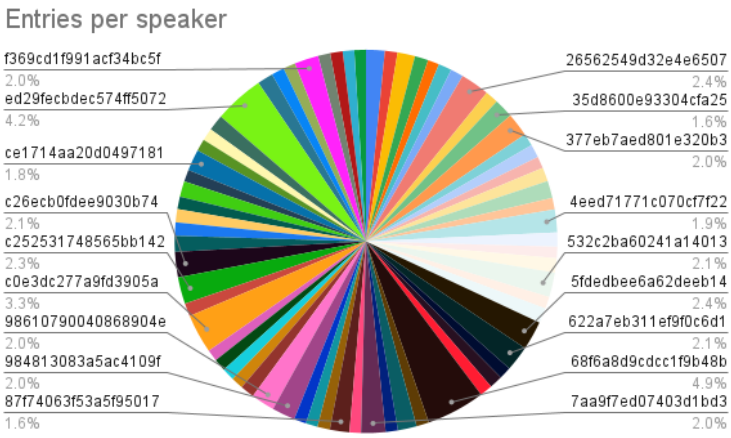


Figure 3: Entries per speaker in the Abkhaz dataset

The phonemization of the Abkhaz data, or simply the conversion of the orthographic annotations to IPA, proved to be somewhat challenging as it is a relatively low-resource language, for which it is difficult to find a ready-to-use phonemizer. We opted to use the web tool provided by Baltoslav, an organization which provides keyboards, language games and other tools for multiple languages such as Upper Sorbian, Chechen and others. The most straight-forward way to incorporate this phonemizing tool was to pre-phonemize the text annotations of the data and thus skip the phonemization step of the text handling of the system during training time. This reorganization of the process is a streamlined task due to the high modularity of the system, as illustrated by the summary of the modified pipeline outline (Figure 4).

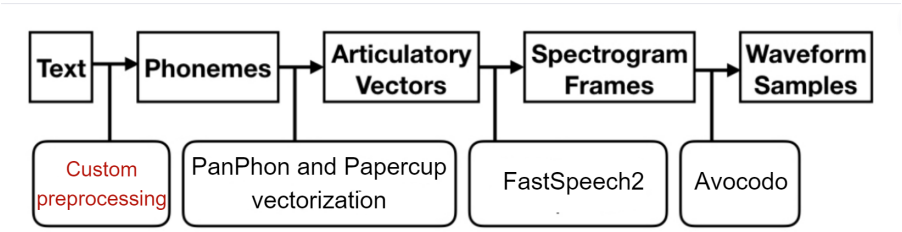


Figure 4: Modified IMS Toucan pipeline

4.2 Fine-tuning pipeline

After the data has been prepared, we can proceed with the fine-tuning of the pre-trained model. The IMS Toucan system is geared towards beginners, which means that there are clear instructions provided for most possible actions, including fine-tuning from a pretrained model checkpoint. There are checkpoints available in the repository for multiple trainable parts of the system – a massively multilingual model, a self-contained aligner, an embedding function, a vocoder and an embedding GAN. As vocoders are generally user independent as also noted in the documentation of the toolkit, we do not need to train a vocoder and can limit ourselves to only training the multilingual model.

The model was trained on data from 12 languages (English, German, Spanish, Greek, Finish, French, Russian, Hungarian, Dutch, Polish, Portuguese and Italian). The total amount of hours

equals 389, with English data constituting the most hours (89) and Greek the least (4). As noted in Lux and Vu (2022), the authors use an embedding function that first applies a non-linear function on the articulatory features data from both the PanPhon and Papercup systems and then projects onto a 512 dimensional space. Further, they employ the LAML fucntion which allows the model to optimize its initialization parameters and is proven to benefit the training, specifically in low-resource fine-tuning scenarios.

The toolkit offers an example file which aids the user in preparing their fine-tuning pipeline. Our adjustments included simply assuring that the system is led correctly to the corpus and that it process the Abkhaz IPA annotations as expected, namely just taking them for input and skipping the phonemization step. The model is fine-tuned for 5000 steps. We performed the fine-tuning on HàbròK, the computer cluster of the University of Groningen, using one GPU unit.

After the fine-tuning, the latest checkpoint is compressed and it can now be easily used for inference. Thus, we can pass to the next step of our methdology, namely synthesising Proto-Indo-European with the pre-trained and fine-tuned models.

4.3 PIE handling

The IMS-Toucan toolkit mostly employs espeak-NG as a backend phonemizer, which does not support Abkhaz, which caused the need for an alternative handling of the fine-tuning data. Of course, it is also unable to handle Proto-Indo-European.

The modification of the toolkit pre-processing module, which allows us to handle Proto-Indo-European input happens at the stage when the phones are turned into vectors that encode the information about their features, where usually the espeak-NG phonemizer is plugged in. A traditional phonemizer aims to do more than simply map each grapheme to a phoneme. It also captures contextual variation and phonetic changes, such as assimilation and coarticulation. Since we are unable to recover Proto-Indo-European speech in such detail and we rely purely on the phonological information we can reconstruct, we can utilize a simple script that maps each notation grapheme to an IPA symbol with no additional transformations. Our code mostly consists of a dictionary, where the graphemes are matched to a phoneme. This is done in accordance with the phonology theory that we choose to follow as described in Section 3.2.

Some of the more peculiar features of Proto-Indo-European require specific solutions. Such were the cases of palatalized phonemes, labialized phonemes and syllabic phonemes. These features are not yet supported by the model. A user of the GitHub community raised the question of including palatalisation as a feature of consonants in the context of synthesising Polish. As explained by the author in a comment, this is relatively difficult as the aligner, which works with the identity of the phones as opposed to the feature vectors, would require modifications as well. The use proposes instead decomposed palatalization, which means reading in the diacritic /ʲ/ as simply the phone [j]. This is reported to lead to audible improvement of the inference.

Since Proto-Indo-European exhibits labialized phones, we include decomposed labialization in the code similarly as to how the user implemented the decomposed palatalization, which would

mean read in in the diacritic ʷ simply as the labial approximant [w]. In the case of the syllabic phonemes, specifically the liquids, we chose to direct the system to read them as the liquid sound, followed by a schwa to signify the expected vocalization.

These adjustment allowed us to be able to input Proto-Indo-European text in the inference code and have the model synthesise speech successfully.

4.4 Building the web app

The toolkit is provided with a streamlined inference procedure that can be easily customized. We built upon this feature by creating a web app using the Streamlit open-source Python library, as per the example of Do et al. (2022) who also built a TTS we app with the library. Our web app allows the user to easily input text in standard Proto-Indo-European notation and synthesise it with the press of a button. The user can choose between using simply the pre-trained multi-language model or its fine-tuned version. The app is directly connected to the GitHub repository of the code and reflects changes, made to the repository in real time. A screenshot of the web app is attached in appendix 7.

Having completed the task of successfully synthesising speech in Proto-Indo-European, we can continue on the evaluation procedure in order to determine whether the zero-shot with phonological features approach is suitable for this task. The results will allow us to establish whether our method is an improvement upon the recent attempt at the same task by Donnelly (2022).

5 Evaluation

The evaluation of the synthesised speech in this project proves to be a major challenge as traditional synthesis evaluation methods usually involve native speakers. Recently, there has been interest surrounding objective measures based on neural networks, such as MOSNet (Lo et al., 2019). They are trained on a lot of data, comprising of sythesised speech samples and their MOS scores given by humans. However, Cooper et al. (2022) showcase that they perform only “moderately well" in zero-shot scenarios, such as ours.

Alternatively, given the possible applications of our project in education, we can also choose to ask experts in the field of comparative Proto-Indo-European linguistics to give their opinion on the synthesised speech. However, by choosing to involve non-experts we aim to gather feedback from a broader audience. A secondary goal of this projects and others alike is often to bring Proto-Indo-European and other extinct languages to life with the intent of spreading the interest in this rich source of history and culture.

Our evaluation procedure focused on the naturalness of the speech, since evaluating intelligibility was not possible. We opted for a traditional mean opinion score (MOS) evaluation task. We lack natural human speech that we can use for reference, which also excludes the option of more elaborate subjective evaluation methods that employ a hidden reference anchor such as MUSHRA.

5.1 Stimuli

The closest that we have to natural speech is the sentences that can be heard in the Vanished Voices app (Kloekhorst, 2020). The eleven reconstructed sentences are read out by the author. However, it is important to note that the phonology that they chose to assign to the notation follows the Glotallic Theory discussed in Section 3.2. Thus, stops are often produces in an ejective manner, which is not in accordance with the phonological features we mapped to the notation graphemes. This invalidates the sentences even further from being considered the natural speech baseline when it comes to evaluating intelligibility. Still, they remain a good base to compare to overall in terms of naturalness as they are spoken by a human.

We include recordings of seven of these sentences in our evaluation and use their content to synthesise sentences with our system and the system by Do et al. (2022). The number of sentences was narrowed down with the consideration of avoiding a tedious evaluation with more than 40 recordings. The sentences in traditional notation and IPA, according to the phonology we adopt, are presented in Table 1.

PIE notation	IPA	Translation
kude tuh$_1$ h$_1$esi?	kude tuʔ ʔesi	Where are you?
k$^{\text{w}}$is h$_1$éḱwoh$_1$i sisde?	k$^{\text{w}}$is ʔek$^{\text{j}}$woʔi sisde	Who is getting on this horse?
d$^{\text{h}}$úgh$_2$tḗr tojdō̆m g$^{\text{w}}$eg$^{\text{w}}$ome.	d$^{\text{h}}$ugʕteːr tojdoːm g$^{\text{w}}$eg$^{\text{w}}$ome	Your daughter has come home.
péh$_2$ur h$_2$eh$_1$séh$_2$i dedh$_1$oje.	peʕur ʕeʔseʕi dedʔoje	The fire is burning in the fireplace.
g$^{\text{w}}$énh$_2$ wéh$_1$r pibh$_3$eti mélidk$^{\text{w}}$e h$_1$edsti.	g$^{\text{w}}$enʕ weʔr pibʕ$^{\text{w}}$ eti melidk$^{\text{w}}$e ʔedsti	The woman drinks water and eats honey.
swésōr h$_1$eso h$_1$ésu westoj.	swesoːr ʔeso ʔesu westoj	His sister is dressed well.
ne h$_1$ésu h$_1$esesm̥, ḱuō̆n k$^{\text{w}}$spénti ph$_2$énti h$_1$eleh$_2$d$^{\text{h}}$.	ne ʔesu ʔesesm̥, k$^{\text{j}}$uoːn k$^{\text{w}}$spenti pʕenti ʔeleʕd$^{\text{h}}$	I did not sleep well because the dog was barking all night.

Table 1: Evaluation sentences

We use the web app provided by Donnelly (2022) to synthesise the sentences. The app allows for a lot of modifications regarding to what phoneme the notation graphemes map and we choose the phonology that we adhere to, namely the widely accepted traditional phonology described in 3.2.
Finally, we also synthesise the sentences using both the multi-language pre-trained model and our fine-tuned version. All the sentences used in the evaluation can be listened to by going to this SoundCloud page.

5.2 Procedure

Thus, we are left with 28 sentences, seven from each of the four categories. We build the evaluation survey using the Qualtrics web platform and was distributed online. The recordings were presented in a randomized order. Upon agreeing to voluntary participation, the participants were instructed to focus on the naturalness of the speech and to use their intuition. They

graded the recordings from 1 to 5 on naturalness using sliders. We gathered responses from 32 participants from different ages, genders and nationality. Since this information is not relevant to the study, we did not gather this data.

6 Results

The gathered results were first preprocessed in Excel to improve readability and formatting. Further, they were statistically analysed using RStudio.

The following chart (Figure 5) reflects the average grading for each of the four categories of stimuli. It is observable that the samples, generated with the Donnelly (2022) model score the lowest, specifically 1.06 on average. The highest rated samples were the ones from the Vanished Voices app, as expected. The average for this group is 4.059. The two models used in this project score almost identically, with 3.29 and 3.25. The fine-tuned model scores slightly lower.

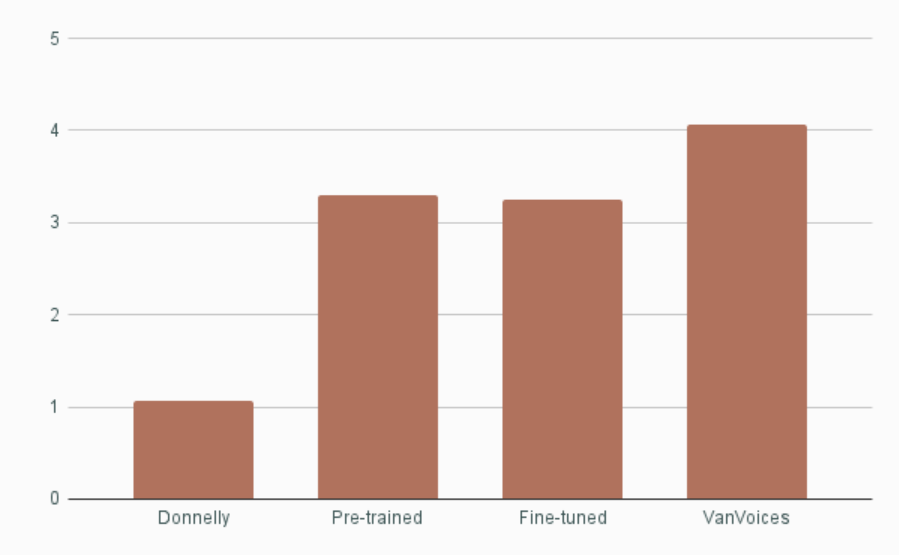


Figure 5: Barplot showcasing the avarage scores of each group

The next graph (Figure 6) details the results per sentence, which helps us scout more fine-grain differences in scoring. Unsurprisingly, the shortest sentence (Where are you?) is the highest rated, with all three models besides Donnelly's scoring close to 4. The other shorter sentences (Who is getting on that horse?, Your daughter has come home.) following this trend. Surprisingly, the longest sentence I did not sleep well because the dog barked all night. scores rather high as well, around 3.4.

The statistical analysis of the ratings per group confirms the intuition one might have upon observing the charts – the differences between the group ratings are significant, with the exception of the comparison between the two models presented in this project, which score with negligible difference. A one-way ANOVA test with dependent variable set as the rating and the independent variable being the group of the stimulus, resulted in a p-value of <2e-16, indicating

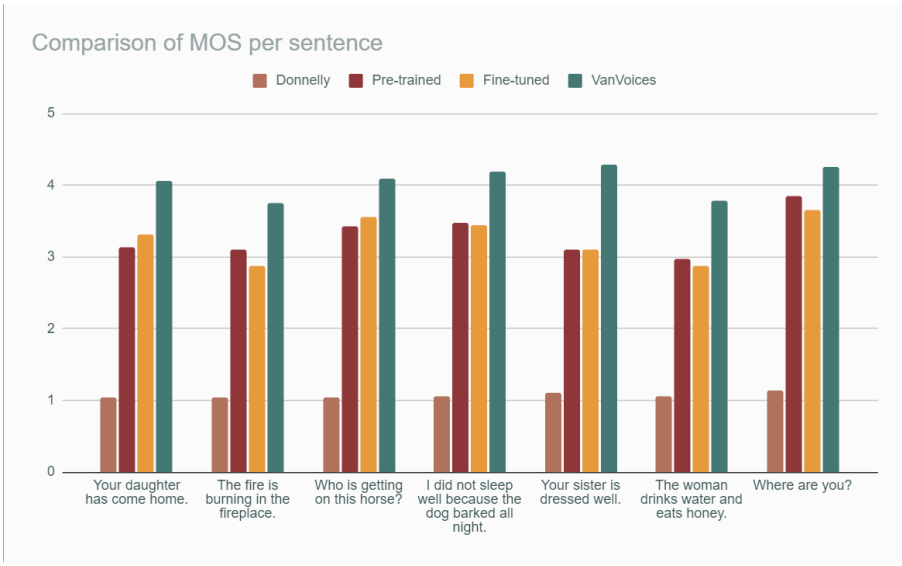


Figure 6: Barplot showcasing the avarage scores of each sentence

strong significance.

A post-hoc Tukey HSD test reveals the size of effect in between groups and the significance of the effect, which is summarised in Table 2. All of the comparisons exhbit a strong significant difference in rating, but the pair between the pre-trained and the fine-tuned model.

	Difference	Lower	Upper	p-value adjusted
finetuned-Donnelly	2.19226190	1.8149144	2.5696094	0.0000000
pretrained-Donnelly	2.20491071	1.8275632	2.5822582	0.0000000
VanVoices-Donnelly	2.99479167	2.6174442	3.3721392	0.0000000
pretrained-finetuned	0.01264881	-0.3646987	0.3899963	0.9997603
VanVoices-finetuned	0.80252976	0.4251823	1.1798772	0.0000010
VanVoices-pretrained	0.78988095	0.4125335	1.1672284	0.0000015

Table 2: Summary of results from the Tuket HSD test

These results allow us to confirm that both of the models used in this study produce more natural Proto-Indo-European speech than the previous attempt at this task by Donnelly (2022). We cannot confirm that the fine-tuned model produces more natural speech than the pre-trained model. Additionally, all of the models are rated significantly different than the human speech, suggesting that the synthesised speech is not yet natural-enough sounding. However, it is worth noting that the difference between the human voice and the two project models is lower than the other differences, albeit still significant.

7 Discussion

This section will first reflect on the results and what insights they give over our hypothesis and research question. Further, we will explore what could be improved upon or expanded in light of the outcomes of our study. Work on interdisciplinary topics like ours opens a bridge between the traditional field of comparative linguistics and modern machine learning techniques, which in turn opens the door to many novel ideas. Thus, we discuss multiple directions for future research on the topic.

The aim of this project was to explore the possibility of employing phonology features zero-shot synthesis to Proto-Indo-European. We hypothesised that using this advanced synthesis technique will be successful and an improvement upon previous attempts at Proto-Indo-European synthesis, specifically the concatenative TTS system by Donnelly (2022). Further, we hypothesised that fine-tuning a multilingual pre-trained model on the Caucasian language Abkhaz will help the model better generalise over the specific and sometimes unseen by the multilingual model phonological features of the Proto-Indo-European phones.

Our results confirm the hypothesis that our method is a significant improvement upon the naturalness of the Donnelly (2022) technique. Our participants judged the speech synthesised with the concatenative technology to be mostly very unnatural. This extremely low level of naturalness impedes the impressive features of the system, such as the flexibility in terms of the phonology chosen by the user.

The statistically significant difference in scores between the human speech and the models we used in this project suggests that our models do not yet produce a human-like speech that sounds fully natural. Nonetheless, the drastic improvement upon the Donnelly (2022) allows us to confirm our hypothesis partially.

The lack of significant (positive) difference in the ratings between the pre-trained model and our fine-tuned on Abkhaz version means that we cannot confirm, nor deny our hypothesis that fine-tuning will improve upon the pre-trained model. This could be caused by a multitude of reasons. Firstly, it is possible that Abkhaz was not an optimal choice as a fine-tuning candidate. We chose it due to its richness of consonants realisations, including extensive labialization and palatalization, common features to Proto-Indo-European. However, these are not entirely rare features, and are exhibited in more modest proportions by some of the European languages that constituted the training data for the multilingual model. Further, the system currently handles labialization and palatalisation in a decomposed manner, as elaborated in Section 4. Thus, it does not learn “labialized" as a feature of the preceding consonant but as a bilabial approximant /w/ following it. Therefore, the effect of presenting the model with labialised consonants perhaps does not lead to optimal learning results. Another useful feature that Abkhaz exhibits is ejective or pre-glottalized consonants. Possible future improvement include the possibility to expand our synthesis to cover multiple possible theorized phonological inventories of Proto-Indo-European, such as the Glottalic theory, which hypothesises the presence of ejectives in the language. In such case, training the model on data that exhibits ejectivized phones would be very beneficial. However, further improvements on the vectorization system would be required to recognize and be able to learn this feature. This leads to the conclusion that the Abkhaz data might be more useful is the system’s handling of some key features is improved.

Upon auditory inspection of the synthesised speech, produced by our modules, we could conclude that there is often an issue surrounding the three “laryngeal" sounds, the identities of which are actually a glottal stop, a pharyngeal fricative and a labialized pharyngeal fricative. Perhaps due to the under-exposure to pharyngeal sounds, the model seems to over-play the rest of the features that compose these sounds. This is in line with what was reported in Lux and Vu (2022), namely that the model can generalise over unseen phonemes and maps them to their own place in the articulatory space. However, the pronunciation is not guaranteed to be precise and can be understood mostly in the context of a longer utterance. Thus, often the glottal stop sounds like a vaguely front-of-the-mouth stop and the pharyngeal fricatives like the much more frequent alveolar or post alveolar fricatives. We observe that in the case of the glottal stop the fine-tuning seems to have somewhat improved the sound of the phone, possibly due to the frequent presence of glottal stops in the Abkhaz data. In fact, two of the three cases where the fine-tuned model scored higher in the MOS task, the sentences included only h_1 – the glottal stop “laryngeal".

Perhaps the biggest challenge of a zero-shot extinct language or near-extinct language is the evaluation process. Many traditional techniques require native speakers as evaluators or native natural speech for reference. For this reason often papers that detail systems suited for zero-shot synthesis employ a faux zero-shot scenario where the target language is not really low-resource (Lux & Vu, 2022; Staib et al., 2020). This allows them to still be able to evaluate the synthesised speech effectively.

Our approach to evaluation focused only on the naturalness of the speech as intelligibility in the context of matching a clear, native-like pronunciation is perhaps not a possible goal as it is difficult to establish this with no reference audio. It is possible that the very similar rating of the pre-trained model and the fine-tuned model is the result of out evaluation methodology. The IMS-Toucan toolkit harnesses the power of powerful modules from multiple TTS systems and is likely to produce human sounding speech in most cases after an appropriate amount of training. The goal of fine-tuning is often to be able to produce the phones of a target language more native-like, but not necessarily more human-like. Therefore, it is possible that our fine-tuned module achieved more correct pronunciations of certain phones, but that did not change the already relatively high naturalness and was thus not reflected in the evaluation. Nonetheless, with our evaluation we showed that this method is a viable tool in the exploration of Proto-Indo-European synthesis. Further research on the topic of evaluation in such case would likely greatly benefit the field.

A truly language agnostic model that operates using the universal notions of the articulatory characteristics of our speech could open the door to the revitalization of any extinct or near-extinct languages and lead to its popularization and development of different educational tools. Whalen et al. (2016) discusses the task of language revitalization in the context of the “extinct" Miami Native American tribe language. As mentioned in this article, 90% of the current 6000 human languages are predicted to disappear in the next century in a perhaps more brutal manner than Proto-Indo-European did, by being suffocated by globalization. Whalen et al. (2016) explores how a community losing its language can affect the physical and mental health of its members. The task of revitalization is comprised of much more than simply restoring the language, namely community and even political engagement, but it starts with reconstructing

the language and bringing new life to it, which can be done through zero-shot synthesis.

On the question of reconstruction, in this case we based ourselves on the manual reconstruction of Proto-Indo-European, as there is a multi century tradition and many resources. However, the manual comparative method suffers from some drawbacks, such as the inevitable circular reasoning that is caused by the connection between defining phylogenetic relations and cognate words (as discussed in section 3.1). Further, for some smaller language groups, there might not be enough tradition or human resources to manually reconstruct a language. This is why machine learning models that are able to derive a reconstructed word from a cognate set of words from daughter languages, such as Bouchard-Côté et al. (2013), would perfectly combine with a zero-shot synthesis approach to efficiently reconstruct and give voice to a language in a streamlined manner.

In the specific case of Proto-Indo-European, but also in other extinct language cases where we have a reconstructed vocabulary and even grammar, we can combine a synthesis model with a machine translation model. This would constitute not only a powerful learning tool but also a resource which is more engaging and friendly to the wide public.

8 Conclusion

In conclusion, our study was able to confirm that zero-shot synthesis using articulatory features is a viable route to explore when looking to synthesise Proto-Indo-European or any other extinct language that has a known or reconstructed phonology. We showed that a deep neural network based model is likely a better choice than other methods such as concatenative synthesise due to its higher level of naturalness of the synthesised speech, in confirmation of our hypothesis. Our study also sheds light on the importance of choosing the right language to fine-tune with, depending on your goals and the capabilities of the system to learn different phonological features.

Further, our study provides an insight on true zero-shot synthesis evaluation and its challenges. Future research on possible evaluation techniques would be beneficial.

Even though our focus is Proto-Indo-European, our conclusions can be generalised over other extinct languages. Possible future research directions include the combining of a synthesis model such as IMS-Toucan with a machine learning reconstruction method, which might limit the need for tedious or unavailable manual reconstruction.

References

Azizah, K., Adriani, M., & Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages. IEEE Access, 8, 179798–179812.

Beekes, R. S. (2011). Comparative indo-european linguistics. Comparative Indo-European Linguistics, 1–439.

Bičovský, J. (2021). The phonetics of PIE *d i: Typological considerations. Linguistica Brunensia, 5–21. https://doi.org/10.5817/LB2021-2-1

Boersma, P. (2011). Praat: Doing phonetics by computer [computer program]. http://www.praat. org/.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. Proceedings of the National Academy of Sciences, 110(11), 4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. Science, 337(6097), 957–960.

Brugmann, K., & Delbrück, B. (1967). Grundriss der vergleichenden grammatik der indogermanischen sprachen.

Chiu, C.-C., Qin, J., Zhang, Y., Yu, J., & Wu, Y. (2022). Self-supervised learning with random-projection quantizer for speech recognition. International Conference on Machine Learning, 3915–3924.

Clackson, J. (2007). Indo-european linguistics: An introduction. Cambridge University Press.

Cooper, E., Huang, W.-C., Toda, T., & Yamagishi, J. (2022). Generalization ability of MOS prediction networks. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8442–8446.

Cuny. (1912). Notes de phonétique historique. indo-européen et sémitique. Revue de phonetique, 2, 101–132.

De Saussure, F. (1879). Mémoire sur le système primitif des voyelles dans les langues indo-européennes. BG Teubner.

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. Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, 16–22.

Donnelly, P. (2022). Concatenative phonetic synthesis for the proto-indo-european language. Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), 193–201.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. International conference on machine learning, 1126–1135.

Gamkrelidze, T. V., & Ivanov, V. V. (1984). Индоевропейский язык и индоевропейцы.

Gamkrelidze, T. V. (1989). The indo-european glottalic theory in the light of recent critique. Folia Linguistica Historica, 22, 3–12.

Hartmann, F. (2021). The phonetic value of the proto-indo-european laryngeals: A computational study using deep neural networks. Indo-European Linguistics, 9(1), 26–84.

Jasanoff, J. (2017). The prehistory of the balto-slavic accent (Vol. 17). Brill.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kloekhorst, A. (2020). Vanished voices. Leiden.

Li, B., Zhang, Y., Sainath, T., Wu, Y., & Chan, W. (2019). Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5621–5625.

List, J.-M., Forkel, R., & Hill, N. W. (2022). A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. arXiv preprint arXiv:2204.04619.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019). Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352.

Lux, F., Koch, J., Schweitzer, A., & Vu, N. T. (2021). The IMS toucan system for the blizzard challenge 2021. Proc. Blizzard Challenge Workshop, 2021.

Lux, F., & Vu, N. T. (2022). Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. arXiv preprint arXiv:2203.03191.

Mallory, J. P., & Adams, D. Q. (1997). Encyclopedia of indo-european culture. Taylor & Francis.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3475–3484.

Rapp, S. (1995). Automatic phonemic transcription and linguistic annotation from known text with hidden markov models/an aligner for german. Workshop "Integration of Language and Speech in Academia and Industry.

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., 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.

Reynolds, E., West, P., & Coleman, J. S. (2000). Proto-indo-european 'laryngeals' were vocalic. Diachronica, 17(2), 351–387.

Staib, M., Teh, T. H., Torresquintero, A., Mohan, D. S. R., Foglianti, L., Lenain, R., & Gao, J. (2020). Phonological features for 0-shot multilingual speech synthesis. arXiv preprint arXiv:2008.04107.

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.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., & Chen, N. (2018). Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015.

Whalen, D. H., Moss, M., & Baldwin, D. (2016, May 9). Healing through language: Positive physical health effects of indigenous language use (5:852) [Type: article]. F1000Research. https://doi.org/10.12688/f1000research.8656.1

Wieling, M., Prokić, J., & Nerbonne, J. (2009). Evaluating the pairwise string alignment of pronunciations. Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELT&R 2009), 26–34.

Wier, T. R. (2005). Abkhaz. Language, 81(2), 516–517.

Xu, J., Tan, X., Ren, Y., Qin, T., Li, J., Zhao, S., & Liu, T.-Y. (2020). Lrspeech: Extremely low-resource speech synthesis and recognition. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2802–2812.

Appendices

A Web App

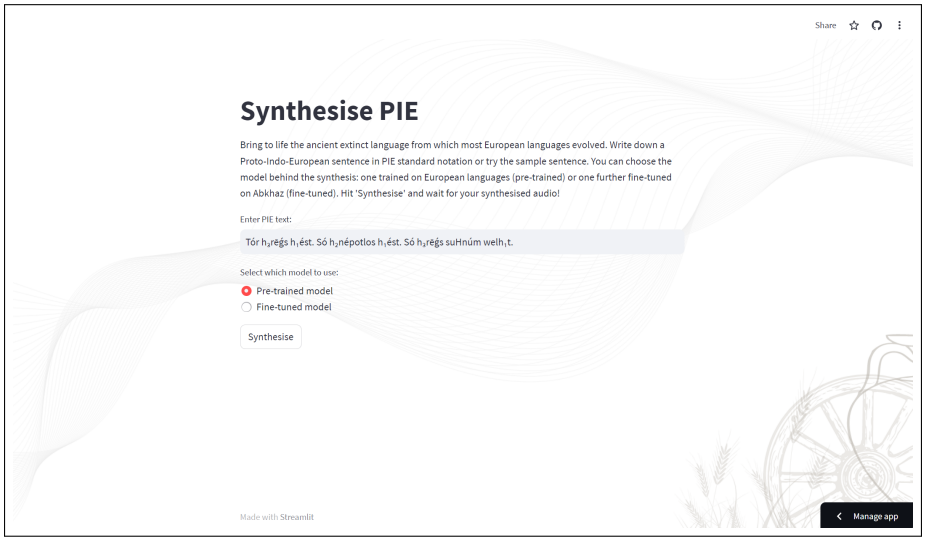


Figure 7: Web App