Concatenative Phonetic Synthesis for the Proto-Indo-European Language

Patrick J. Donnelly
Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, USA
patrick.donnelly@oregonstate.edu

Abstract

We propose a flexible concatenative text-to-speech system to synthesize hypothesized pronunciations of the reconstructed Proto-Indo-European language. To accomplish this, we synthesize speech examples in 100 extant languages and extract individual phones. Using this large database of phonetic sounds, we concatenate individual phonemes together to estimate pronunciations of Proto-Indo-European. Where available, we prioritize consecutive phones from the same source to help increase naturalness and intelligibility of the synthesized speech. Since the language's precise pronunciation is debated, we provide an interface to select the specific phonetic symbol(s) used for each of the language's phonemes and diphthongs. We provide this novel interactive tool to enable researchers and students to aurally explore the different and competing phonological hypotheses debated in the literature.

1 Introduction

Proto-Indo-European (PIE) is the reconstructed ancestor of all Indo-European languages. PIE is hypothesized to have been spoken as a single language sometime during the late Neolithic through the Early Bronze Age (between 4500 to 2500 BCE). According to the Kurgan hypothesis (Gimbutas, 1956), the language likely originated in the Pontic-Caspian steepe of eastern Europe. Over the following centuries, waves of Indo-European (IE) peoples migrated across much of the Eurasian continent. As they dispersed, their language split and underwent shifts in pronunciation, changes in morphology, and acquisitions of new vocabulary. This process continued for centuries, resulting in 448 extant daughter languages across eight subfamilies.[1]

[1] https://www.ethnologue.com/subgroups/indo-european

There is no historical record of PIE. Like other proto languages, the language was meticulously reconstructed using the comparative method (Hoenigswald, 1963). Although Indo-European (IE) linguists have largely converged on the phonemic inventory of PIE, there remains ongoing debate about the interpretation of these phonemes. Unlike living languages, there does not exist an agreed upon mapping of phonemes in PIE to specific phonetic symbols in the International Phonetic Alphabet (IPA). For this reason, there have not been previous attempts to synthesize PIE speech.

In this work we present a text-to-speech system that attempts to estimate PIE speech given a specific mapping of phonemes to IPA pronunciations. We sample phonetic sounds from 100 modern languages to build a concatenative speech synthesizer which is able to pronounce text in the reconstructed Proto-Indo-European language. We provide a flexible approach that allows listeners to tune the phonology to enable aural realizations of different hypothetical pronunciations. To our knowledge, this is the first attempt at speech synthesis for a prehistoric reconstructed language.

2 Proto-Indo-European Language

The phonology of PIE has been reconstructed based on the phonology of extant IE languages. This scholarship initially relied upon modern and well attested historical languages, such as Latin, Ancient Greek, and Vedic Sanskrit. However, the surprise discoveries of the Hittite and Tocharian languages in the early 20th Century provided new scholarly evidence that led to new understandings and sparked new academic debates (Jasanoff, 2017).

Hundreds of words have been reconstructed in PIE and scholars have largely converged on the morphology, although areas of debate still remain. In 1868, the linguist August Schleicher com-

		Labial	Coronal	Dorsal			Laryngeal
				palatal	plain	labial	
Nasals		*m	*n				
	voiceless	*p	*t	*ḱ	*k	*k^w	
Stops	voiced	(*b)	*d	*ǵ	*g	*g^w	
	aspirated	*b^h	*d^h	*$ǵ^h$	*g^h	*g^{wh}	
Fricatives			*s				*h_1, *h_2, *h_3
Liquids			*r, *l				
Semivowels				*y		*w	

Table 1: Common notation used for Proto-Indo-European phonology (Kapović, 2017). The preceding asterisk (*) denotes the phoneme is reconstructed rather than attested. The symbol *b is disputed and shown in parenthesis. The superscripts ^{h} and ^{w} stands for aspiration and labialization, respectively. The symbols *h_1, *h_2, *h_3 serve as phonemes for the three unknown laryngeal sounds. The phoneme *y represents the palatial semivowel (IPA /j/).

posed the short story "*The Sheep and the Horses*" ("H_2ówis h_1éḱwōs-k^we") in his version of reconstructed PIE (Adams, 1997). Over the decades, linguistics have published revisions to this story, accounting for new consensuses in the field or to advance their own linguistic hypotheses. In the absence of any text in PIE, this story has come to serve as the standard mechanism to demonstrate and compare different reconstructions. More recently several Indo-Europeanists linguists collaborated to each reconstruct their versions of another short story entitled "*The King and the god*" ("H_3réḱs deywós-k^we") (Adams, 1997).

2.1 PIE Phonology

Although linguistics have generally converged on the phonetic inventory of PIE, there remains significant debate regarding the pronunciation of these phonemes (Kapović, 2017). The pronunciations of certain sounds in PIE are not known, and may never been known. The majority of phonetic controversy concerns two issues. The first debate pertains to the pronunciation of the series of plosive stops. The second debate pertains to the belief that PIE had a set of phonemes that are not attested to in any extant daughter language.

2.1.1 Glottalic Theory

Glottalic theory proposes that PIE had ejective stops (*p', *t', *k') instead of the traditionally reconstructed plain voiced stops (*b, *d, *g). Once popular, this theory is no longer widely accepted by historical linguists (Barrack, 2002). The reconstruction of these phonemes is made more difficult by the *centum-satem* language split that divides northern and southern IE languages. This divide is named for the pronunciation of the word "hundred" in early PIE languages (Greek vs. Sanskrit). In centum languages, the plain and palatovelears merged together, while in the satem languages, the plain and labiovelars merged together.

2.1.2 Laryngeal Theory

The other controversy surrounding PIE phonology pertains to the number and pronunciation of vowels in PIE. Laryngeal theory proposes that there existed at least three sounds which do not survive in any extant daughter languages. Once reconstructed with five vowels, PIE is now commonly reconstructed with only two vowels, [e] and [o], that were colored by three hypothesized laryngeal sounds: *h_1, *h_2, and *h_3. Today most linguists accept the existence of the laryngeals but continue to dispute their exact phonetic realization (Keiler, 2015). As such, there exist numerous competing interpretations and revisions in the scholarly literature.

2.2 Open Questions

Given these significant open questions regarding its pronunciation, there have not been any previous attempts at automatic speech synthesis of PIE. Most modern synthesis approaches that seek to produce naturalistic speech require extensive knowledge about the pronunciation rules of the language or large datasets of spoken examples. Therefore, speech synthesis using cutting edge technologies is not readily possible for reconstructed languages.

3 Speech Synthesis

Speech synthesis is the task of converting written text into an audio waveform that represents a machine generated realization of the spoken text. These systems are also known as Text-to-Speech (TTS) tools. The majority of approaches for speech

synthesis utilize large corpora of text and audio examples. Accordingly, the majority of research in speech synthesis has focused on widely-spoken languages, particularly those with global influence.

Despite calls for more speech recognition tools for under-resourced languages (Besacier et al., 2014), there has been relatively little work in TTS for most of the world's languages. Nevertheless, speech synthesizers have been built for historical languages, such as Latin and Greek; constructed languages, such as Esperanto and Klingon (Jokisch and Eichner, 2000); and some endangered European languages like Basque or Irish (Chasaide et al., 2017). In an encouraging direction, the authors of a recent study collected and analyzed corpora, documented phonology, and built TTS systems for 12 different African languages (Ogayo et al., 2022).

Researchers have been attempting speech synthesis since the 1950's. Here we briefly review the principle categories of speech synthesis models.

3.1 Articulation Synthesis

Articulation Synthesis is a physical model which emulates various aspects of human pronunciation to synthesize speech. They provide tuneable parameters for the various aspects of human pronunciations (e.g., tongue, pharynx, vocal chords, etc.). While articulation synthesizers are able to produce high-quality speech, the large number of parameters make these systems computationally expensive. These constraints limit their practicality in real-time deployment. Neil Thapen's interactive Pink Trombone[2] is a recent interactive example of articulation synthesis.

3.2 Formant Synthesis

Formant Synthesis generates speech by sending a source signal through a series of filters modeling different formants of the desired speech sound. Speech created with formant synthesis often sounds less naturalistic than other approaches, but it is fast to produce and is generally intelligible (Lukose and Upadhya, 2017). In this work, we make use of the popular formant synthesizer `eSpeak-ng`[3] to generate various phonetic sounds across different languages spoken by a single artificial speaker.

[2] https://imaginary.github.io/pink-trombone/

[3] https://github.com/espeak-ng/espeak-ng

3.3 Concatenative Speech Synthesis

Concatenative Speech Synthesis is another traditional approach that relies upon large databases of sounds vocalized by the same speaker. These systems concatenate different prerecorded waveforms together in order to vocalize the desired text. Depending on the system, the individual constituent waveforms may be phones or phonemes, syllables, or entire words. Speech produced with concatenative synthesis is often intelligible but can potentially sound rather unnatural. This occurs because of the limited syntactic and semantic context as different sounds are spliced together at a very low structural level (Khan and Chitode, 2016).

3.4 Machine Learning Approaches

In recent years, these aforementioned signal processing methods have largely been superseded by new approaches using machine learning. Because these new approaches require large corpora of spoken audio examples, they are not suited for our attempt to estimate pronunciation of a non-attested reconstructed language. These approaches are beyond the scope of this work, but we briefly review the important recent developments here.

Techniques using statistical parametric estimation, such as hidden Markov models, consistently achieved robust and intelligible speech synthesis (Yamagishi et al., 2009). However, the recent abundance of big data and advances in deep learning algorithms have led to new speech synthesis techniques that dominant use in modern day TTS applications and research directions (see review (Ning et al., 2019)).

One such example is Wavenet, an autoregressive generative model for end-to-end speech synthesis (Oord et al., 2016) which models the waveform directly, without requiring a hybrid model or other processes to assemble the synthesized speech. Another such production-quality model, Deep Voice, synthesizes speech entirely from deep neural networks (Arık et al., 2017). Ongoing research in speech synthesis continues in many areas, such as emotional, dialogic, and spontaneous speech production (Delić et al., 2019).

3.5 Our Approach

Our approach requires the use of sets of phonetic sounds that do not occur together in any single language. And we require these sounds to originate from the same speaker. Unfortunately, there has

been very little development into multi-language speech synthesis (Malcangi and Grew, 2010). None of the aforementioned modeling techniques are particularly well suited for this task alone. For similar reasons, concatenative synthesis has previously been used to study some under-resourced languages (Van Niekerk and Barnard, 2009).

In this work we use formant synthesis to produce a library of spoken sounds in various languages. We then use concatenative synthesis to build speech from this corpus of sampled sounds.

4 Database of Phones

We present an approach that attempts to estimate the speech in PIE by concatenating different phonetic sounds extracted from various languages. To accomplish this task, we require phonemes across many different languages spoken by the same speaker. This is necessary in order to maintain a continuity of acoustic characteristics across the concatenation points. To build this dataset, we generate word lists in multiple languages, synthesize speech utterances for each word, splice the audio by individual phoneme, and save these phonetic sounds to a database.

4.1 Sampling Languages

To generate multi-language speech using the same speaker, we use the open-source `eSpeak-ng` tool, a popular TTS engine. `eSpeak-ng` currently supports 127 languages and accents.

For each language available in `eSpeak-ng`, users have carefully crafted lists of phonemes, pronunciation rules, and example words. These specific pronunciation rules allow the synthesizer to generate more realistic speech by considering the pronunciation context of phonemes, syllables, and words. These rules are developed with feedback from native speakers, and subsequently the languages available in `eSpeak-ng` reflect only those languages with a very large number of speakers.

Of the 127 available languages, we exclude constructed languages such as Esperanto or Klingon but we include the IE languages of Ancient Greek and Latin. For English, we retained only standard American and British pronunciation, and we exclude other less common dialects. For Spanish and Portuguese, we include both European and Latin American pronunciation. Although the majority of remaining languages are IE languages (56%), we also include non-IE languages in order to increase our phonetic inventory. Among those, we included three dialects of Chinese: Cantonese, Hakka and Mandarin. Altogether, we select 100 unique languages and five additional dialects for a total of 105 speakers from whom we synthesize speech.

4.2 Swadesh Lists

Swadesh lists were devised by linguist Morris Swadesh as a tool when measuring relationships between languages using glottochronology. A Swadesh list contains 207 words in a particular language (Swadesh, 1952). Given their long use in comparative linguistics, there exist complete or partial lists for many of the world's languages. We collect these lists for our 100 selected languages from Wiktionary.[4]

Some of these lists contain multiple synonyms or variants for each word. We exclude stand-alone suffices, but otherwise accept all complete words. Our goal is not to compare phonology but to generate a sampling of many possible phonemic pronunciations in the language. Therefore the number of words synthesized varies between languages.

4.3 Generating Phonemes

Next we synthesized each of these words using Praat.[5] Praat is popular tool for speech analysis in phonetics that also provides speech synthesis using `eSpeak-ng`. When generating speech, Praat labels and segments each of the phonemes used to create the synthesized utterance, using the Kirshenbaum phonetic encoding notation for IPA. This provides very specific IPA notation for each individual phoneme. These include various articulation diacritics, co-articulations, and diphthongs.

Praat supports a scripting language, which we use to automate the following tasks. For each language we generate a speech synthesizer. We use the default speaker voice "Female 1" and a speech rate of 150 words per minute. For each word in the language word list, we automatically generate the synthesized utterance. We iterate across the waveform to split and save each excerpt representing a single phoneme. We save these phonetic sounds as a single channel 16k Hz `wav` file. We snip all audio at zero-crossings in the waveform to prevent sonic artifacts when concatenating sounds together. We henceforth refer to these excerpts as phones, indicating that they have been extracted from the

[4] https://en.wiktionary.org/wiki/Appendix:Swadesh_lists

[5] https://www.praat.org

phonemic context of their source languages. Later, we will use sets of these phones to approximate phonemes in PIE. In a database, we log the source for each phone, and we make note of its context by logging the phones that precede and follow it.

4.4 Phonetic Inventory

In total, we collected 124,252 audio samples covering 339 uniquely labeled phonetic sounds. This set includes many short and long vowels, diphthongs, and consonantal articulations. Because the generated phones are shaped by the phonemic context within the source word and the specific pronunciation rules of their language, the duration of the excerpts differ even among sounds with the same precise phonetic IPA annotation.

Some of the sounds hypothesized to exist in PIE no longer occur in any IE daughter languages. While they do survive in some of the world's extant languages, they are quite rare. One such example is the voiced uvular plosive /ɢ/ (Prescott, 2018) which does not occur in the languages available.

Among our phonetic sounds extracted from these 100 widely spoken languages, we find examples of aspiration (e.g., /g^h/) and palatalization (e.g., /g^j/) but not examples of labialization (e.g., /g^w/). Instead we will need to approximate a few hypothesized sounds (e.g., /k^w/, /g^w/, and /g^{wh}/) using pairs of phones. For example, we substitute the sequence of two consecutive phones /gw/ for /g^w/. In languages such as Welsh, this sound occurs frequently, descended from this origin in PIE.

5 Concatenative Synthesizer

In this section we outline our concatenative speech synthesizer and describe our user interface.

5.1 Text Processing

In order to prepare the user-provided text for phonetic matching against our database of phones, we must perform a number of text manipulation steps. First, we normalize whitespace and convert the text to lowercase. Next, we split the text into individual sentences and phrases. We then split these phrases further into individual words. Lastly, we remap alternate orthography to a common notation. For example, the graphemes *ĝ and *ĝ are both remapped to *ǵ. Finally, we tokenize each word into phonemes in PIE, greedily grouping characters together to match the phonological notation given in Table 1.

5.2 Mapping Phonemes

In this step, we map the different PIE phonemes to specific pronunciations in IPA. Because there is no established pronunciation for PIE, our goal is to provide a flexible tool to realize different hypothetical pronunciations. To accomplish this, we read a JSON file containing a mapping of PIE phonemes to phonetic sounds. This mapping can be specified by the user at run-time to guide the desired pronunciation of the synthesized speech. The user is able to assign an IPA symbol to each consonant, vowel, and vowel-semivowel diphthong (e.g., *ew, *oy).

We also provide the user control over the pronunciation rules for the vowels following each of the unknown laryngeals *h_1, *h_2, *h_3. The potential "coloring" of the vowels following these sounds is an important part of hypotheses subscribing to Laryngeal Theory. Evidenced by pronunciations across its descendants, the presence of *h_2 is thought to color the vowel *e to *a while *h_3 colors *e to *o. Additionally, when the laryngeal occurs after the vowel, it likely lengthens the vowel.

5.3 Matching Phones

For each PIE phoneme in a word, we search our database for an example that represents the target sound. Of those found, we then examine the phone that precedes and that follows our target phone. We first attempt to retrieve a tri-gram of consecutive phones taken from a single source word. When unavailable, we next prioritize retrieving a bi-gram of phones. Finally, when such a pair is unavailable, we resort to a single phone. Our approach prioritizes finding consecutive phones from a single source word in order to attempt a more natural pronunciation. This is especially beneficial in the cases of short syllables or voiceless consonants whose pronunciation is shaped by adjacent vowels.

In this task, we consider silence as a possible target sound, which allows us to explicitly find sounds that start or end a syllable when present in the data. We prioritize matching these phones at the beginning and end of syllables to help shape a more natural pronunciation. In particular, this process helps reflect the natural attack and release that occurs at the beginning and end of words.

5.4 Audio Manipulation

Next, we manipulate the audio snippets that correspond to accented vowels. We provide three possibilities for handling the accent. In the first, we

ignore the accent altogether and use the unaltered phone. In the second option, we provide a stress accent. To simulate a stressed pronunciation, we increase the amplitude of the waveform containing the vowel by +3 dB. In the third option, we apply a pitch accent. To simulate a pitch accent, we raise the frequencies of the waveform by one-tenth of an octave. These default threshold values were selected to sound reasonable but can readily be tuned.

Once matching audio files have been identified for each PIE phoneme in the input, we concatenate these phones together. We add pauses of silence between syllables, words, and sentences. To help make the speech sound slightly more natural, we randomize the amount of silence added, scaled by the type of the pause. As a final step, we export a single channel 16 kHz audio file that can be saved to disk or displayed on an interactive web-page.

5.5 User Interface

We provide a graphical interface to demonstrate our concatenative speech synthesis tool, built in Python using the data science interface Streamlit.io.[6] This interface, shown in Figure 1, allows a user to select a particular mapping from each PIE phoneme to a specific IPA pronunciation. For each phoneme, we provide various possibilities hypothesized in the literature (Swiggers, 1989; Beekes, 2011; Meier-Brügger, 2013; Kapović, 2017; Byrd, 2018). However, some of these options are limited by what is available in our dataset of phonetic sounds.

After selecting the phonology, the user can enter a PIE word or phrase in the textbox. After the user clicks the "Speak" button, the app syntheses the text to speech using the process outlined in Section 5. This process is quick but may take a few seconds for longer texts. The interface then adds an audio player widget, allowing the user to listen to the speech or save the file to disk. Below the audio player, the interface prints the IPA transcription that was used to produce the speech. The user also is given an option to display a spectrogram generated from the waveform of the speech utterance.

Each time the user clicks the the "Speak" button, the app will generate a new utterance. Because we randomly select from the multiple audio examples available for each IPA symbol, each generated utterance will have a slightly different pronunciation, even when resynthesizing the same word.

[6] https://streamlit.io/

6 Discussion

Over the last two centuries of studies, linguists have meticulously reconstructed the Proto-Indo-European language. Since we may never know exactly what PIE sounded like, we can only estimate its pronunciation. Confounding this issue, there are multiple and competing hypotheses in the literature debating the language's pronunciation.

The quality of synthesized speech is often evaluated using human subjects and Mean Opinion Score tests (Streijl et al., 2016). Such an evaluation approach is not feasible for a reconstructed language which lacks consensus in its phonological interpretation. Nor can we use objective tests to compare the synthesized speech to spoken examples, since no such recordings exist. For these reasons, we do not attempt empirical evaluation of the quality of the naturalness of our speech. Instead we present this flexible and interactive tool as a way to estimate hypothetical pronunciations of PIE.

6.1 Limitations

Our system is a concatenative speech synthesizer using speech sounds generated by formant synthesis. Given the limitations of these older technical approaches, our synthesized speech will not sound naturalistic. Although our generated speech may sound mechanical and emotionless, it is highly intelligible. In this work, our goal is not naturalistic speech but to provide a means to realize hypothetical speech using custom pronunciations specified by the user. Our phonetic concatenative approach yields this flexibility whereas other models, whether they be articulation synthesizers or cutting-edge deep-learning approaches, cannot.

In our process, we naïvely decontextualize phonemes from their use in their source languages and reappropriate them as individual phonetic sounds towards our goal of approximating speech in PIE. Because we are divorcing individual phonetic sounds from their pronunciation context, we are ignoring the subtle differences in the pronunciation of the same IPA symbol in different languages. For this reason, the timing or transition of some generated sequences of phones occasionally do not flow naturally together, such as an undue pause between sounds. If a pronunciation is not satisfactory, the user can readily generate another rendering.

We do not make attempts to control for other high-level aspects of the pronunciation, such as emotion or prosody. Any attempt to define rules

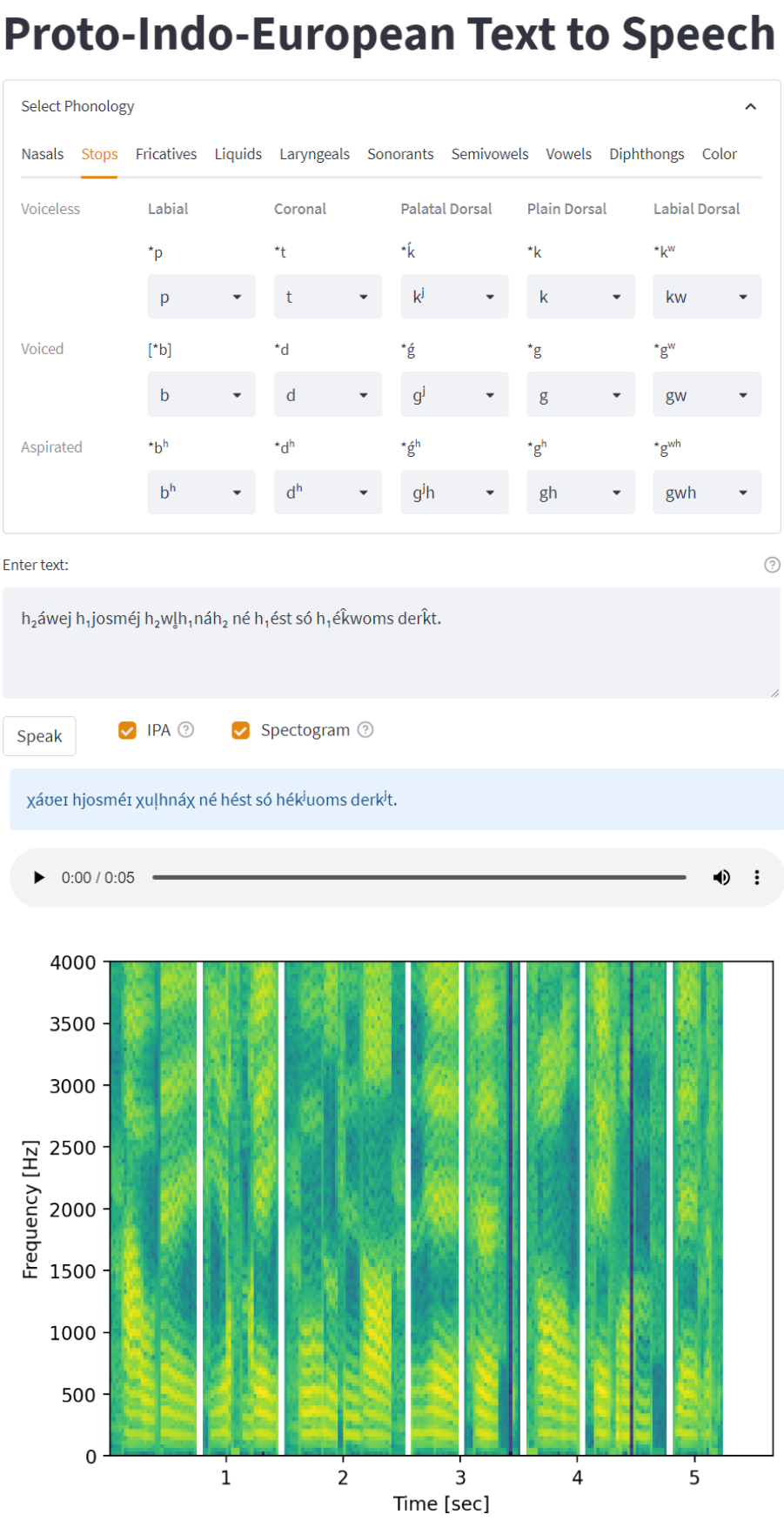

Figure 1: Screenshot of the user interface.

to control prosody in PIE would largely be based on conjuncture. We caution that attempts to make more our synthesizer sound more naturalistic would introduce even more bias, such as favoring the phonology of one language over another.

Although we sample from a large set of the world's popularly spoken languages, we are still missing several of phonetic sounds favored in some hypotheses of PIE phonology. One such sound, the voiced uvular plosive consonant represented in IPA as /ɢ/, is quite rare and occurs in only 2% of the world's languages.[7] Furthermore, articulation variants, such as /ɢʷ/, /ɢʰ/, and /ɢʷʰ/ are even more rare. We also lack examples of the labialized voiceless velar plosive /kʷ/. To compensate for these sounds we instead needed to substitute consecutive phonetic pairs from the same source utterance.

6.2 Future Work

To increase our phonetic inventory, we will consider larger words lists. This will allow us to encounter more naturally occurring sequences of two and three consecutive phones. However, increasing our inventory will also increase the disk space required to store the individual audio excerpts. Therefore, we will strategically sample from this data set to reduce excessive duplicates. Specifically, we will consider the phones that precede and follow as well as the length of a phone when deciding whether to retain or prune a phone from our dataset. In this way, we can retain a diversity of pronunciations and contexts, while eliminating those that are essentially duplicates.

As another strategy to improve our system, we will explore di-phone concatenative synthesis. A di-phone-based approach prioritizes concatenating sounds that cover the transition between individual phones. This typically consists of units sounding from the middle of a phone to the middle of the next phone. For those transitions not available in our dataset, we will substitute a single phone.

To improve our system's ability to realize all possible hypotheses of PIE phonology, we need to create examples of those few sounds we lack in our current approach. To do so, we intend to design and compile a `espeak-ng` speaker that uses the phonology we currently lack, such as the sound /ɢ/.

Although our current approach provides a few naïve options for pronunciation of PIE's accent, more work remains to improve this feature. We will consult with Indo-European linguistics to determine realistic parameters to more faithfully replicate the pitch and stress accents. We intend to add support for a rising accent, another hypothetical possibility for PIE's treatment of accent.

6.3 Contributions and Novelty

We describe a novel approach for a concatenative TTS synthesizer for the Proto-Indo-European language. We combine formant and concatenative synthesis to simulate a phonology of sounds that are not present together in any single language. When generating speech, we search among the multiple stored examples for each phone and consider the immediate phonetic context.

We present this unique and novel tool with the hopes of educating users about the ancestor language of 46% of the world's speakers. The approach presented here can be adapted by researchers to explore different hypothetical pronunciations of other reconstructed or extinct languages. For example, we envision extensions to this work to provide learners a way to aurally explore the effects of different sound laws, such as Grimm's law (Germanic), Burgmann's law (Indo-Iranian), or Winter's law (Balto-Slavic).

7 Conclusion

We propose a novel text-to-speech system to realize various hypothetical pronunciations of the Proto-Indo-European language. We synthesize speech waveforms from word lists using 100 different languages to extract a large library of individual phonetic sounds. Using this inventory, we build a concatenative speech synthesizer that combines phones in an attempt to recreate spoken speech in PIE. Since the precise pronunciation of many phonemes in PIE is uncertain, our system provides a user interface that permits users to select a specific IPA realization for each PIE phoneme and diphthong. Where available in the data, we prioritize extracting consecutive phones from the same original source utterance in order to provide continuity of the waveform and the pronunciation. We randomly select from the matching phones to provide a slightly different pronunciation each time a text is synthesized. We add small randomized pauses between syllables, words, and sentences to better emulate naturalistic speaking rates. We provide an interactive demonstration of our tool at `https://soundbendor.org/projects/PIE`.

[7] `https://phoible.org/`

References

Douglas Q Adams. 1997. *Encyclopedia of Indo-European Culture*. Taylor & Francis.

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. 2017. Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 195–204.

Charles M Barrack. 2002. The glottalic theory revisited: a negative appraisal. *Indogermanische Forschungen*, 107(1):76–95.

Robert SP Beekes. 2011. *Comparative Indo-European linguistics: an introduction*. John Benjamins Publishing.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Andrew Miles Byrd. 2018. 121. the phonology of proto-indo-european. In *Volume 3 Handbook of Comparative and Historical Indo-European Linguistics*, pages 2056–2079. De Gruyter Mouton.

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl. 2017. The abair initiative: Bringing spoken irish into the digital space. In *Proc. of Interspeech*, pages 2113–2117.

Vlado Delić, Zoran Perić, Milan Sečujski, Nikša Jakovljević, Jelena Nikolić, Dragiša Mišković, Nikola Simić, Siniša Suzić, and Tijana Delić. 2019. Speech technology progress based on new machine learning paradigm. *Computational intelligence and neuroscience*.

Marija Gimbutas. 1956. *The prehistory of eastern Europe*. 20. Harvard University.

Henry M Hoenigswald. 1963. On the history of the comparative method. *Anthropological linguistics*, pages 1–11.

Jay Jasanoff. 2017. The impact of hittite and tocharian: Rethinking indo-european in the 20th century and beyond. *Handbook of Comparative and Historical Indo-European Linguistics*, pages 220–238.

Oliver Jokisch and Matthias Eichner. 2000. Synthesizing and evaluating an artificial language: Klingon. In *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 1, pages 729–732.

Mate Kapović. 2017. Proto-indo-european phonology. *The Indo-European Languages*, pages 13–60.

Allan R Keiler. 2015. A phonological study of the indo-european laryngeals. In *A Phonological Study of the Indo-European Laryngeals*. De Gruyter Mouton.

Rubeena A Khan and JS Chitode. 2016. Concatenative speech synthesis: A review. *International Journal of Computer Applications*, 136(3):1–6.

Sneha Lukose and Savitha S Upadhya. 2017. Text to speech synthesizer-formant synthesis. In *2017 International Conference on Nascent Technologies in Engineering (ICNTE)*, pages 1–4. IEEE.

Mario Malcangi and Philip Grew. 2010. Toward language-independent text-to-speech synthesis. *WSEAS Transactions on Information Science and Applications*, 7(3):411–421.

Michael Meier-Brügger. 2013. *Indo-European Linguistics*. de Gruyter.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building african voices. *arXiv preprint arXiv:2207.00688*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Charles E. Prescott. 2018. Pharyngealization and the three dorsal stop series of proto-indo-european. In *The Meaning of Language*, page 220. Cambridge Scholars Publishing.

Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

P Swiggers. 1989. Towards a characterization of the proto-indo-european sound system. In Theo Vennemann, editor, *The New Sound of Indo-European: Essays in Phonological Reconstruction*, pages 177–208. De Gruyter Mouton.

Daniel R Van Niekerk and Etienne Barnard. 2009. Phonetic alignment for speech synthesis in under-resourced languages. In *Proc. of Interspeech*, pages 880–883.

Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. 2009. Robust speaker-adaptive hmm-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1208–1230.