

Oregon State University

Concatenative Phonetic Synthesis

for the Proto-Indo-European Language

Patrick Donnelly

December 17, 2022

soundbendor lab

Proto-Indo-European

Indo-European (IE) Family

- 448 extant languages
- 46% of the world's speakers
- e.g., English, Spanish, Germany, Greek, Russian, Hindi, ...

Proto-Indo-European

Indo-European (IE) Family

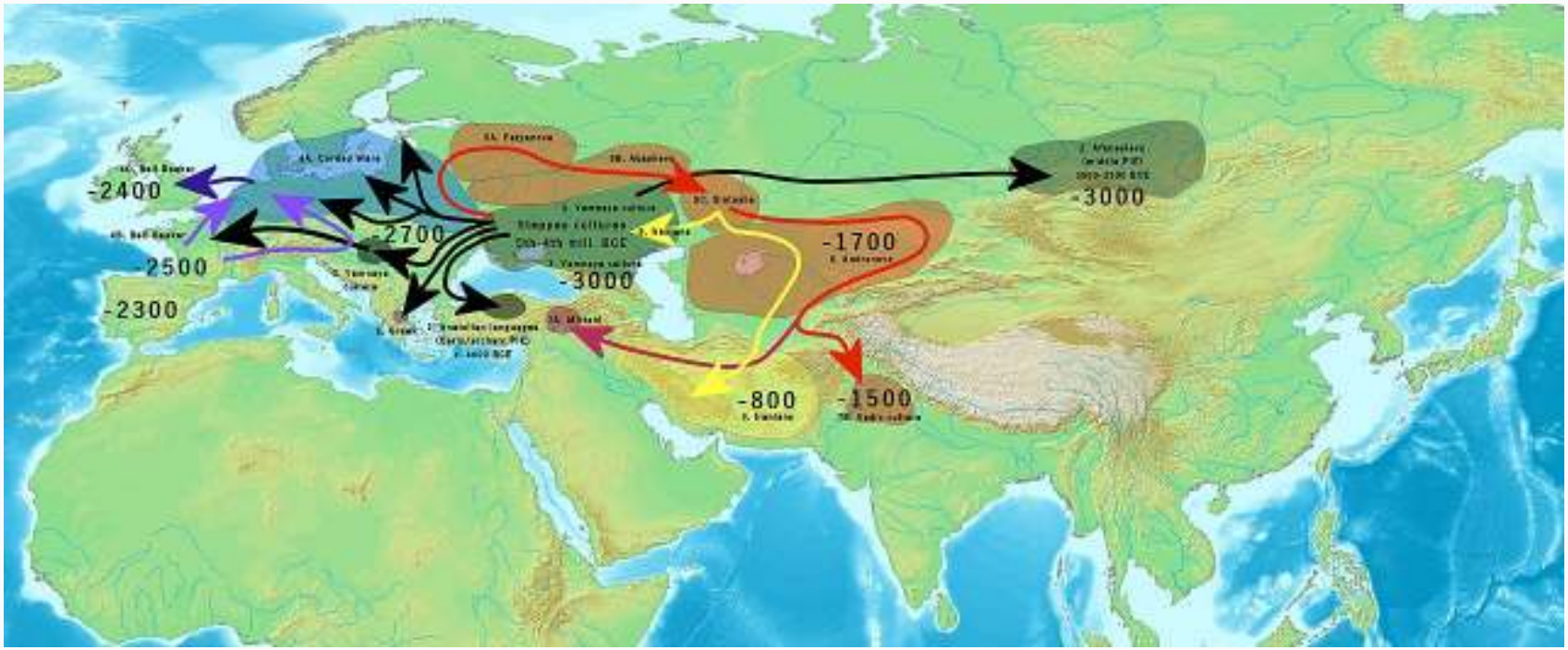
- 448 extant languages
- 46% of the world's speakers
- e.g., English, Spanish, Germany, Greek, Russian, Hindi, ...

Proto-Indo-European (PIE)

- reconstructed ancestor of IE languages
- spoken 4500 to 2500 BCE
- origins in the Pontic-Caspian steepe (Kurgan hypothesis [1])
- pronunciation is disputed [2]

Indo-European Migrations

(CC BY-SA 3.0)



Map of Indo-European migrations

Phonology

		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	

Common notation used for Proto-Indo-European phonology [3].

Speech Synthesis

- **Articulation Synthesis**: *physical model of human articulation*
  - computationally expensive
  - e.g., Pink Trombone
- **Formant Synthesis**: *source shaped by a series of formant filters*
  - fast and generally intelligible [4]
  - e.g., `eSpeak-ng`
- **Concatenative Synthesis**: *concatenated prerecorded waveforms*
  - often intelligible but robotic [5]

Speech Synthesis

- **Articulation Synthesis**: *physical model of human articulation*
  - computationally expensive
  - e.g., Pink Trombone
- **Formant Synthesis**: *source shaped by a series of formant filters*
  - fast and generally intelligible [4]
  - e.g., `eSpeak-ng`
- **Concatenative Synthesis**: *concatenated prerecorded waveforms*
  - often intelligible but robotic [5]
- **Hidden Markov Models** [6]
- **Deep Learning** (see review [7])
  - **recent models**: Wavenet [8], Deep Voice [9]
  - **ongoing research**: emotional, dialogic, and spontaneous speech [10]

Database of Phones

1 Sampling Languages

- TTS engine eSpeak-ng
- 105 languages/accents

Language Family	count
Indo-European	58
Balto-Slavic	15
Indo-Iranian	14
Italic	12
Germanic	10
Celtic	3
Hellenic	2
Albanian	1
Armenian	1
Turkic	11
Dravidian	4
Semitic	4
Sino-Tibetian	4
Uralic	4
Malayo-Polynesian	3
Bantu	2
Thai	2
Other	13
	105

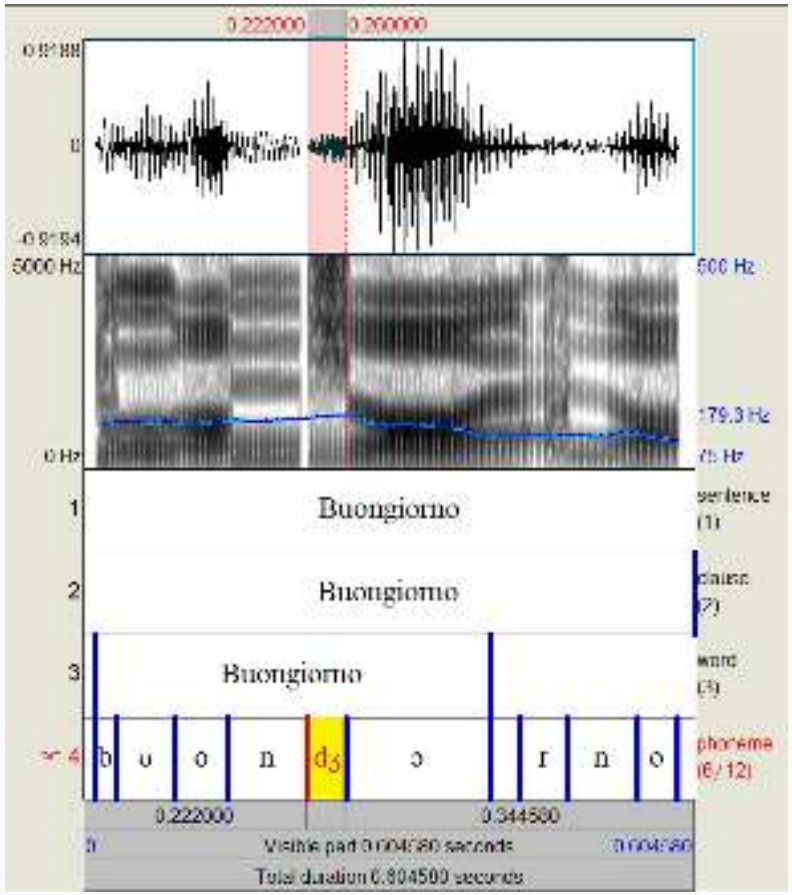
Database of Phones

1. **Sampling Languages**
   - TTS engine eSpeak-ng
   - 105 languages/accents
2. **Generating Speech**
   - Swadesh Lists
   - 207 common words
   - include synonyms

№	English	Italiano
...		
42	mother	madre
43	father	padre
44	animal	animale
45	fish	pesce
46	bird	uccello
47	dog	cane
48	louse	pidocchio
49	snake	serpente
50	worm	verme
51	tree	albero, pianta
52	forest	foresta, bosco
53	stick	bastone
54	fruit	frutta
55	seed	seme
56	leaf	foglia
57	root	radice
58	bark	corteccia
59	flower	fiore
60	grass	erba
...		

Database of Phones

1. **Sampling Languages**
   - TTS engine eSpeak-ng
   - 105 languages/accents
2. **Generating Speech**
   - Swadesh Lists
   - 207 common words
   - include synonyms
3. **Segmenting Phonemes**
   - Pratt scripting language
   - 124k audio samples
   - 339 unique IPA symbols



Concatenative Synthesis

1. **Text Processing**
   - normalize whitespace
   - set lowercase
   - split into individual words
   - remap notation
     - e.g., *ĝ and *g̃ → *ǵ

Concatenative Synthesis

1. **Text Processing**
   - normalize whitespace
   - set lowercase
   - split into individual words
   - remap notation
     - e.g., *ĝ and *ǧ → *ǵ
2. **Mapping Phonemes**
   - PIE phoneme → IPA
   - flexible mapping
     - from UI or JSON
   - easy to change

Concatenative Synthesis

1 Text Processing

- normalize whitespace
- set lowercase
- split into individual words
- remap notation
  - e.g., *ĝ and *g̑ → *ǵ

2 Mapping Phonemes

- PIE phoneme → IPA
- flexible mapping
  - from UI or JSON
- easy to change

3 Matching Phonemes

- longest source sequence
  - tri-, bi-, or uni-gram
- prioritize:
  - start of word
  - end of word

Concatenative Synthesis

1 Text Processing

- normalize whitespace
- set lowercase
- split into individual words
- remap notation
  - e.g., *ĝ and *g̊ → *ǵ

2 Mapping Phonemes

- PIE phoneme → IPA
- flexible mapping
  - from UI or JSON
- easy to change

3 Matching Phonemes

- longest source sequence
  - tri-, bi-, or uni-gram
- prioritize:
  - start of word
  - end of word

4 Audio Manipulation

- accented vowels
  - stress → +3 dB
  - pitch → $+\frac{1}{10}$ octave
- add randomized pauses
  - word, sentence, phrase

Limitations

- decontextualize phonemes
  - speech is mechanical
  - no emotion or prosody
- missing some sounds
  - e.g., /ɢ/, /ɢ^{wh}/, /k^{w}/

Limitations & Future Work

Limitations

- decontextualize phonemes
  - speech is mechanical
  - no emotion or prosody
- missing some sounds
  - e.g., /ɢ/, /ɢ^{wh}/, /k^{w}/

Future Work

- sample more words
  - prune near duplicates
- di-phone synthesis
- create `espeak-ng` speaker
- improve accent
  - add rising accent

Interactive Demonstration

https://soundbendor.org/projects/PIE

Conclusion

Summary

We present a **novel** text-to-speech system to realize hypothetical pronunciations of the Proto-Indo-European language.

- concatentative speech synthesizer
- flexible phonology
- stress and pitch accent

Conclusion

Summary

We present a **novel** text-to-speech system to realize hypothetical pronunciations of the Proto-Indo-European language.

- concatentative speech synthesizer
- flexible phonology
- stress and pitch accent

Applications

- aid in the study of PIE
- readily extensible
  - reconstructed, historical, or under-resourced languages

References

[1] M. Gimbutas, *The prehistory of eastern Europe*. Harvard University, 1956.

[2] A. R. Keiler, "A phonological study of the Indo-European laryngeals," in *A Phonological Study of the Indo-European Laryngeals*, De Gruyter Mouton, 2015. DOI: 10.1515/9783110812480.

[3] M. Kapović, "Proto-Indo-European phonology," *The Indo-European Languages*, pp. 13–60, 2017. DOI: 10.1163/22125892-00701002.

[4] S. Lukose and S. S. Upadhya, "Text to speech synthesizer-formant synthesis," in *2017 International Conference on Nascent Technologies in Engineering (ICNTE)*, IEEE, 2017, pp. 1–4. DOI: 10.1109/ICNTE.2017.7947945.

[5] R. A. Khan and J. Chitode, "Concatenative speech synthesis: A review," *International Journal of Computer Applications*, vol. 136, no. 3, pp. 1–6, 2016. DOI: 10.5120/ijca2016907992.

[6] J. Yamagishi, T. Nose, H. Zen, *et al.*, "Robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009. DOI: 10.1109/TASL.2009.2016394.

[7] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019. DOI: 10.3390/app9194050.

[8] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016. DOI: 10.48550/arXiv.1609.03499.

[9] S. Ö. Arık, M. Chrzanowski, A. Coates, *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 195–204. DOI: https://arxiv.org/abs/1702.07825.

[10] V. Delić, Z. Perić, M. Sečujski, *et al.*, "Speech technology progress based on new machine learning paradigm.," *Computational intelligence and neuroscience*, 2019. DOI: 10.1155/2019/4368036.