# Tironiculum—Latin Speech Recognition via Latin Text-to-Speech

**Lee Butterman**
Poeta ex Machina Labs
leebutterman@gmail.com

## Abstract

All this paper is divided into three parts. We introduce a text corpus of Latin prose, and we introduce a parallel text-audio corpus of synthetic Latin speech for both single words and lines of dactylic hexameter, to introduce the first Latin speech recognition system, Tironiculum, using wav2vec2. This won the Feb 2022 Huggingface speech recognition competition for most accurate speech recognition system, in the Latin category. Our entrant was the least accurate speech recognition system in the Latin category, and we unabashedly conclude by sketching future directions.

## 1 Motivation—Self-Supervised Speech Recognition

Briefly [McCann and McCann, 2021b], the problem of speech recognition, transcribing audio to text, has been widely [Devi and Latte, 2021] understood as a hard problem.

Latin is particularly useful for speech recognition space, even as English is the hegemonic default [Bender, 2019] in Natural Language Processing (and defaults are difficult to subvert [Hurtubise et al., 2021]), because 94% of the world's people do not have English as a first language [Leffert and Reed, 2021], and Latin was a *lingua franca* before the lingua Franca.

Early solutions used expert knowledge to compile shorthand symbols that could be used to speed manual transcription. Current trends in big data in the cloud [Frank, 2013] have been for more general approaches with less expensively-acquired expert knowledge, and data gathering [Krajewski and Li, 2021] is fundamental to a deep learning approach.

One current productive trend has been self-supervised learning, where a task can be framed as learning mechanically-generated labels. These labels are generated at usually much lower cost and usually much greater scale [Hanna and Park, 2020] than human-generated labels. Self-supervised learning often amounts to learning the inverse of a mechanical process: image recoloring for black-and-white photographs is learned as the inverse of stripping images of their color; super-resolution [Vincent, 2020] is learned as the inverse of downsampling images; language modeling is learned as the inverse of deleting a word in a sequence (at the end is called 'causal language modeling', in the middle is called 'masked language modeling'). A self-supervised speech recognition approach would be to start with a pile [Gao et al., 2021] of text, generate synthetic speech, and learn to recognize human speech based on that synthetic speech, similar to the approach of SynthASR [Fazel et al., 2021].

Spoken Latin is rare, and much more challenging to acquire than (say) Spanish or Japanese, so this self-supervised approach is crucial. (The careful observer will ask, why does one need speech recognition at all, if spoken Latin is very rare. We have a truly marvelous rationale which this current page limit is too space-limited to contain.) Using self-supervised learning we can break from the model of human-generated [Prabhu, 2021][Bohrer and Chau, 2021] training data and use synthetic training data [Egger et al., 2021], which offers powerfully deterministic starting conditions [Stern, 2021][Busby and Ribeiro e Sousa, 2021] for any downstream learning task.

Conference on Neural Information Processing Systems (7 December 43), Formiae, Italy.

## 2 Contribution: dataset of Latin text

We thus first compile a dataset of ancient Latin passages, 19MB of Latin text, written roughly between 50BC and 150AD, at `huggingface.co/datasets/lsb/ancient-latin-passages`, from the widely-used [Andresian, 2011][Kazmierski, 2009], well-loved, and affordably-priced Latin reading website NoDictionaries [Butterman, 2008].

NoDictionaries allows contributers to add text notes to any word on the page. Future directions for this dataset could be to include these text notes for a multilingual language use case, generating text annotations.

Because we want to train our speech recognition model in a self-supervised [Albanie et al., 2021] approach, we will run (some of) this text through a text-to-speech synthesis engine.

## 3 Contribution: dataset of Latin synthetic speech

Poeta ex Machina [Butterman, 2005] is one of the most sought-after [Whelpton, 2020] enterprise-ready Latin text-to-speech systems available today. Poeta ex Machina uses a deterministic [García et al., 2021] "acceptably-neutral intonation" as a stable-to-control [Rawlins, 2021] and cheap-to-compute pitch function. Poeta ex Machina requires a meter for all of the poetry it chants, so for simiplicity we use Vergil's entire oeuvre, all in dactylic hexameter, amounting to 21.4 hours of audio. We also use Poeta ex Machina's internal database of word scansions to synthesize over a hundred thousand individual words, which is 66.9 additional hours of audio. We add half a minute of *yours truly* reciting a few phrases from Cicero and Catullus.

The collection is available at `github.com/lsb/poetaexmachina-mp3-recitations`. Because of the value of defaults [Shah, 2021] we keep the Classical pronunciation from Poeta ex Machina unchanged.

Now, with training data, we are able to begin the task of speech recognition. (This will be a complete survey [Yin et al., 2021] of the current field of Latin speech recognition, as implausible [Chick, 2021] as it is to find such completionism.)

## 4 Contribution: Italian wav2vec2, fine-tuned on Latin

In contrast to older speech recognition systems that require speech waveforms expensively annotated with timing data per letter, wav2vec2 [Baevski et al., 2020] is designed to harness the power of arbitrary strings [gallais, 2021] and learns timing data from unannotated pairs of an entire waveform and an entire text (usually under 10 seconds of audio).

The community and infrastructure around wav2vec2 means that there are many wav2vec2 models trained on various modern languages. We can take a large pre-trained model whose training data is close to the target data distribution, and use it as a foundational [Bommasani et al., 2021] starting point [Li et al., 2017], instead of starting training from scratch. Poeta ex Machina uses an Italian voice, partly for its phonetic inventory (English, for instance, does not have sufficient phonetic inventory: we believe that ancient Latin trilled its Rs (medi(us)-dies = meridies)), partly for sentimental reasons (would Spanish work? Russian? Xhosa?). For similarly phonetic and sentimental reasons, and availability, we use a wav2vec2 model trained on the Italian dataset of Vox Populi, and fine-tune from there. Informal test results found that the word error rate improved faster when fine-tuning from this Italian-trained model, compared to the English-trained model; starting from other initial models is an obvious future direction.

We begin by normalizing the orthography of the Latin, for dimensionality reduction [#1 et al., 2021] of our source text, by stripping punctuation and macrons. We further normalize letters invented after 500AD like 'j' and 'v' into their original 'i' and 'u', taking advantage of the backwards compatibility[Copley, 2021] of the orthography. Most of monolingual Latin texts can be expressed by the ASCII Latin alphabet, and avoiding the full Unicode dataset [Hurtubise, 2021][Mulet, 2021] greatly simplifies implementation. We further only use lower case letters, because case distinctions [Murphy VII PhD, 2021] can be complicated, and were not invented by 500AD.

Wav2vec2 uses Connectionist Temporal Classification [Graves et al., 2006] to infer its transcription: at the bottom level [Wang, 2021] at each 20ms timestep we predict a letter or a break, and by analyzing the sequence afterwards [Madaan and Yao, 2021][Chuang and Wu, 2021a] we merge identical letters with no break in between to determine letter boundaries and word boundaries [Thorrez, 2021], terminating in a finite number of steps [Simmons, 2021].

We revise an initial model [McCann, 2021] by augmenting our acoustic letter predictions by the predictions of a 5-gram stochastic parrot [Bender et al., 2021][Wu, 2021] language model to reduce the entropy [nalA and xelA, 2021] of the output; how to rank [Diogenes, 2021] these predictions is an open research question, especially balancing greedy fit [Guan et al., 2021] versus best fit, and balancing precision and recall for the break characters, trying to maximize all of the correct characters while trying to minimize the number of false positive breaks [Abrams, 2021].

We follow the trend in using specialized hardware [McCraith, 2021] gaining power exponentially as per Moore's law [Efrati, 2021]: we train on GPU, in 16-bit and 32-bit floating point [Curry et al., 2021] precision, ensuring that all of our weights' and biases' normal base-2 [Jalaboi and Hansen, 2021] mantissas comport with the Strong Newcomb-Benford Law [Chuang and Wu, 2021b].

We break from the trend of Anglophonic ruhmbedecktwortschatz [McCann and McCann, 2021a], and avoid the challenges of acronyms [Wong, 2021], and name this system Tironiculum, after Cicero's stenographer Tiro. Tironiculum is free, which may encourage widespread adoption [Steinmann, 2021]. Not only is the code online (`github.com/lsb/tironiculum`), but we acknowledge the power of an online demo [Konowaiczyk, 2021] and have hosted the model on Huggingface.

Superhuman performance has long been of interest [Ashley et al., 2021] in the broader research community, and our results come astronomically close.

## 5 Results and future directions

The word error rate at the end of training was 0.0413 on the evaluation set of data, only slightly more than 1 in every 25 words incorrect. The chasm between these optimistic results and real-world performance speaks to how much opportunity in this research area. Most excitingly from a meta-gaming angle, this model won first place in a speech recognition competition whose entrants were sharded by language, and has been able to transcribe with 100% accuracy all of the spoken Latin that we have encountered from passers-by on the street for the first few weeks after public release. This initial model was the 0.25x size 'base' model, chosen to aid in prototyping speed. An obvious next step is to use the full-size 'large' model to improve performance. Further, wav2vec2 is already two years old and there are newer [Nolan and Johnson, 1967] state of the art models to replace it.

## 6 Impact statement, ethical concerns, and funding statement

Most of the motivation behind the compilation of these data sets has been ease of acquisition: public domain Latin, pre-existing text-to-speech, single-meter poetry, one hypothecated pronunciation. Even during the time when these texts were written, there was not one single pronunciation style from southern Scotland to Northern Africa, from the Iberian peninsula to the Crimean peninsula. Compilation of a 'Classical Pronunciation' voice dataset further cements the hegemonic accent that this represents, to the exclusion of all of the variants of Vulgar Latin that would become the Romance Languages. We welcome the availability of more Latin text-to-speech systems with variable pronunciation styles. For the 0.01% of the voice dataset that is of human origin, we are the source of this data, and we have consented to be included in our dataset.

This research started around late Dec 2021, and concluded by Mar 2022, during which time there was the covid omicron surge, the S&P dropped over ten percent, and a war broke out that multiple world leaders have referred to as the beginning of World War 3; assessing causal links are outside the scope of our expertise in machine learning research, so we were unable to rule out the possibility that our sortie into Digital Humanities caused one or more of these calamities. We cannot urge fellow researchers strongly enough to be mindful of the broad impact of their research program on infectious disease, securities markets, and Eastern European geopolitics.

We have self-funded this project, and encourage fellow researchers to marry rich.

## References


Prophet #1, Prophet #2, and Prophet #3. Universal Insights with Multi-layered Embeddings. *Sigbovik*, 2021.

Josh Abrams. On Sigbovik Paper Maximization. *Sigbovik*, 2021.

Samuel Albanie, Erika Lu, and João F Henriques. On the origin of species of self-supervised learning. *Sigbovik*, 2021.

Anna Andresian. Techno-teaching: practical, manageable online resources. 2011. URL https://www.magistrula.com/app/assets/docs/technoteaching.pdf.

Dylan R Ashley, Anssi Kanervisto, and Brendan Bennett. Back to Square One: Superhuman Performance in Chutes and Ladders. *Sigbovik*, 2021.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR*, abs/2006.11477, 2020. URL https://arxiv.org/abs/2006.11477.

Emily M Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, September 2019.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Rose Bohrer and Connie Chau. Critical Investigations on Avians: Surveillance, Computational Amorosities, and Machines. *Sigbovik*, 2021.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Philihp Busby and Daniel Ribeiro e Sousa. Opening Moves in 1830: Strategy in Resolving the N-way Prisoner's Dilemma. *Sigbovik*, 2021.

Lee Butterman. Poeta ex Machina. 2005. URL https://poetaexmachina.net.

Lee Butterman. NoDictionaries. 2008. URL https://nodictionaries.com.

Thomas Chick. "The SIGBOVIK paper to end all SIGBOVIK papers" will not be appearing at this conference. *Sigbovik*, 2021.

Gabriel Chuang and Brandon Wu. What Lothar Collatz Thinks of the CMU Computer Science Curriculum. *Sigbovik*, 2021a.

Gabriel Chuang and Brandon Wu. The Newcomb-Benford Law, Applied to Binary Data: An Empirical and Theoretic Analysis. *Sigbovik*, 2021b.

R Copley. A Note on the Consent Hierarchy. *Sigbovik*, 2021.

Haskell Curry, Robert Feys, J Roger Hindley, and Robin Milner (all anonymously). STOP DOING TYPE THEORY. *Sigbovik*, 2021.

J Devi and Chai-Tea Latte. Demystifying the Mortal Kombat Song. *Sigbovik*, 2021.

Diogenes. Winning the Rankings Game: A New, Wonderful, Truly Superior CS Ranking. *Sigbovik*, 2021.

Benjamin Efrati. Stone Tools as Palaeolithic Central Unit Processors. *Sigbovik*, 2021.

Bernhard Egger, Kevin Smith, ~~David Cox~~, and Max Siegel. openCHEAT: Computationally Helped Error bar Approximation Tool—Kickstarting Science 4.0. *Sigbovik*, 2021.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. SynthASR: Unlocking Synthetic Data for Speech Recognition. *CoRR*, abs/2106.07803, 2021. URL https://arxiv.org/abs/2106.07803.

Steven Frank. *cloud-to-butt*. 2013. URL https://github.com/panicsteve/cloud-to-butt/.

gallais. Dependent Stringly-Typed Programming. *Sigbovik*, 2021.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Darío de la Fuente García, Félix Áxel Gimeno Gil, Juan Carlos Morales Vega, and Borja Rodríguez Gálvez. On the dire importance of MRU caches for human survival (aginst Skynet). *Sigbovik*, 2021.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Shane Guan, Blair Chen, and Skanda Kaashyap. The Urinal Packing Problem in Higher Dimensions. *Sigbovik*, 2021.

Alex Hanna and Tina M Park. Against scale: Provocations and resistance to scale thinking. 2020. URL https://arxiv.org/pdf/2010.08850.pdf.

Nicolas Hurtubise. Unicode Magic Tricks. *Sigbovik*, 2021.

Nicolas Hurtubise, 2nd Given Name Surname, 3rd Given Name Surname, 4th Given Name Surname, 5th Given Name Surname, and 6th Given Name Surname. Refutation of the *"Failure to remove the template text from your paper may result in your paper not being published"* Conjecture. *Sigbovik*, 2021.

Raluca Jalaboi and Mads Eiler Hansen. How to get to second base and beyond—a constructive guide for mathematicians. *Sigbovik*, 2021.

S Kazmierski. Latin With No Dictionaries? 2009. URL http://latinteach.blogspot.com/2009/06/latin-with-no-dictionaries.html.

Marcin Konowaiczyk. Macro-driven metalanguage for writing Pyramid Scheme programs. *Sigbovik*, 2021.

David Krajewski and Eugene Li. Solving reCAPTCHA v2 Using Deep Learning. *Sigbovik*, 2021.

Akiva Leffert and Jason Reed. Oracle Types. *Sigbovik*, 2021.

Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. *CoRR*, abs/1712.09913, 2017. URL http://arxiv.org/abs/1712.09913.

Aman Madaan and Gary Yao. Yet Another Lottery Ticket Hypothesis. *Sigbovik*, 2021.

Jim McCann. Instruction Programs. *Sigbovik*, 2021.

Jim McCann and Mike McCann. RadicAI: A Radical, Though Not Entirely New, Approach to AI Paper Naming. *Sigbovik*, 2021a.

Jim McCann and Mike McCann. Story Time. *Sigbovik*, 2021b.

Robert McCraith. Tensorflow for Abacus Processing Units. *Sigbovik*, 2021.

Michael Mulet. A full video game in a font: Fontemon! *Sigbovik*, 2021.

Dr Tom Murphy VII PhD. Lowestcase and Uppestcase letters: Adventures in Derp Learning. *Sigbovik*, 2021.

usH nalA and eiX xelA. Inverted Code Theory: Manipulating Program Entropy. *Sigbovik*, 2021.

William F Nolan and George Clayton Johnson. Logan's run. 1967.

Vinay Uday Prabhu. Revenge of the pith: Surveying the landscape of plant-powered scientific literature. *Sigbovik*, 2021.

Freddie Rawlins. Spacecraft Attitude Determination and Control. *Sigbovik*, 2021.

Shalin Shah. Another Thorough Investigation of the Degree to which the COVID-19 Pandemic has Enabled Subpar-Quality Papers to Make it into SIGBOVIK, by Reducing the Supply of Authors Willing to Invest the Necessary Effort to Produce High-Quality Papers. *Sigbovik*, 2021.

Robert J Simmons. Build your own 8-bit busy beaver on a breadboard!, or, Look, it's clearly decidable whether any program on your computer terminates or not. *Sigbovik*, 2021.

Patrick Steinmann. `NetPlop`: A moderately-featured presentation editor built in NetLogo. *Sigbovik*, 2021.

Sam Stern. Soliterrible: Deterministically Unplayable Solitaire. *Sigbovik*, 2021.

Clayton W Thorrez. Deep Deterministic Policy Gradient Boosted Decision Trees. *Sigbovik*, 2021.

James Vincent. What a machine learning tool that turns Obama white can (and can't) tell us about AI bias. June 2020. URL https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias.

Zikuan Wang. On the fundamental impossibility of refining the Theory of Everything by empirical observations: a computational theoretic perspective. *Sigbovik*, 2021.

John Whelpton. Latin Speech Engines. 2020. URL https://web.archive.org/web/20201015163020/https://linguae.weebly.com/latin-speech-engines.html.

Cameron Wong. SIGBOVIK2021 isn't called SIGCOVID. *Sigbovik*, 2021.

Brandon Wu. If It Type-checks, It Works: FoolProof Types as Specification. *Sigbovik*, 2021.

Hesper Yin, Oscar Dadfar, Max Slater, Anne He, Alice Lai, Emma Liu, and Po Po. A Complete Survey of 0-Dimensional Computer Graphics. *Sigbovik*, 2021.