Tungnaá: a Hyper-realistic Voice Synthesis Instrument for Real-Time Exploration of Extended Vocal Expressions

Jonathan Chaim Reus*
EMUTE Lab
University of Sussex
Brighton, UK
j.reus@sussex.ac.uk

Victor Shepardson*
Intelligent Instruments Lab
University of Iceland
Reykjavík, Iceland
victorshepardson@hi.is

Thor Magnusson
Intelligent Instruments Lab
University of Iceland
Reykjavík, Iceland
thormagnusson@hi.is

ABSTRACT

This demo showcases Tungnaá, a new voice synthesis system and software instrument for real-time musical exploration of "Deep Voice Synthesis". The design of Tungnaá emphasizes real-time interaction and customization, enabling artists to manipulate various aspects of the synthesis process and to explore aesthetic artefacts unique to autoregressive neural synthesis. The synthesis engine achieves real-time streaming generation of paralinguistic and extended forms of vocal expression, while controlling them using symbolic text notations drawn from the entire unicode character set, allowing for the creation of new notation systems. The interface provides visual display and mouse- or OSC-controllable interventions into the machine vocalisations. The demo showcases Tungnaá on a laptop with headphones and a MIDI controller, allowing participants to explore the instrument via both a textual and physical interface.

Author Keywords

Speech Synthesis, Voice, Machine Learning, Paralinguistic, Autoregressive, Realtime

CCS Concepts

•**Applied computing** → **Sound and music computing;** •**Computing methodologies** → **Neural networks;** •**Human-centered computing** → *Interactive systems and tools;*

1. INTRODUCTION

Deep learning-based voice synthesis methods ("Deep VS") are lately capable of producing human-sounding voice with ever increasing realism, prosodic expressivity and computational efficiency. The bulk of research has been in the domains of text-to-speech (TTS) and singing voice synthesis (SVS).However, TTS systems are rarely designed with musicality in mind, and as a plethora of artists in experimental music and across cultures demonstrate, the human voice is capable of expressions far beyond the narrow focus of SVS research on clean and precise tonal singing styles.

This motivates us to design a voice instrument that leverages the hyper-realistic synthesis of Deep VS towards vocal expressions which may be non-tonal, paralinguistic and extended into arbitrary domains of human vocal expression, while also conferring the uncanny and intriguing artefacts of neural synthesis – what we call the "WaveNet Aesthetic" of hyperrealistic babbling first heard in a 2016 demo [19].

So motivated, we designed Tungnaá, a software NIME based on Deep VS. In its synthesis engine, a Tacotron2-style [23] alignment model predicts audio features from text, while a RAVE [6] vocoder streams them to audio. We also explore token-free text encoders [9] to allow the creation of arbitrary text-based notation systems for generating audio. Tungnaá combines this engine with a GUI exposing the underlying mechanisms of neural synthesis, like text-audio alignments and latent variables, for manipulation.

2. BACKGROUND

2.1 Voice Instruments

Given that the voice is such a prominent part of most forms of music across cultures, it's unsurprising that artists would seek out new and unique relationships with this "primal instrument" [10]. Voice research in NIME is so rich that Kleinberger et al [16] produced a taxonomy of vocal NIMEs, based on whether the voice is input or accompaniment, analyzed or synthesized, or live versus pre-recorded. The new wave of data-driven Deep VS techniques complicates this categorisation, where pre-recorded samples are used as training data, but analysis and synthesis are entangled with a degree of malleability suggesting reconsideration of the relationship between artist and voice.

Nevertheless, we feel several real-time interfaces for speech synthesis could be considered spiritual antecedents of Tungnaá. Glove-TalkII [12] was a seminal glove-based gestural controller using a neural network to map gestures to articulatory voice synthesis parameters in real-time, and which was later imagined as a NIME in GRASSP [20]. Another tangible interface for statistical parametric speech synthesis came with MAGE [2] in 2012. Indeed, instruments based on "old fashioned" articulatory synthesis may be closest in spirit to our work, because they open up the voice to its broadest sonic possibilities. The popular web-based articulatory synthesizer Pink Trombone [25] is notable.

What we call "the WaveNet aesthetic", the signature weirdnesses of hyper-real babbling, neural network distortion artefacts, and paralinguistic voice sounds can be heard in contemporary music, such as the 2021 album *AAI* by Mouse on Mars utilizing the `krach.ai` synthesis instrument, Holly Herndon's 2019 *PROTO* [13] which used a Deep VS sys-

*The authors contributed equally to this work

NIME'24, 4–6 September, Utrecht, The Netherlands.

tem to glitchy effect, and DadaBots's [31] collaboration with Jennifer Walshe on the 2020 *A Late Anthology of Early Music*. Such work might be placed within a tradition of composers working with vocal fragments such as Paul Lansky [15], and we are also inspired by artists in the traditions of Dadaist sound poetry, such as Jaap Blonk or Tomomi Adachi, who notably uses an "Infrared Sensor Shirt" instrument to live-sample improvised sound poetry. Finally, musicians such as Jonathan Chaim Reus have combined vocal improvisation with Deep VS and bespoke physical interfaces, bending voice conversion models to create uncanny vocalizations [22].

2.2 Deep Voice Synthesis

Deep VS can be seen as an extension of traditional statistical parametric speech synthesis to nonlinear models with large numbers of parameters, which can result in highly realistic and flexible models when fit to large datasets.

DNN-based SVS approaches have developed rapidly [8], but most research focuses on obtaining clean articulation and melody with a commercial music production environment in mind. In light of this, our research focus on adapting TTS methods into a NIME context.

2.2.1 Streaming TTS

Streaming TTS methods continuously generate speech with low latency, compared to offline methods which must work with large chunks. Many streaming methods [28][24] build on FastSpeech2 [21]. These rely on a text encoder to predict token durations, and causal layers to then decode to vocoder features. When durations are modeled as conditionally independent, it can lead to poor performance on expressive speech datasets. Recent work in the FastSpeech2 lineage [1] includes a more powerful generative duration model, but still requires durations estimated by an external alignment model when training.

Another family of streaming TTS methods based on Tacotron2 [23] learn text-audio alignment jointly with conditional generation. These models compute a distribution of attention over text tokens for each frame of audio, depending on all past audio frames, alignments, and the input text [5].

2.2.2 Vocoding

Most streaming TTS methods rely on a separate vocoder. Streaming neural vocoders include WaveRNN [14] and derivatives [26] [17]. While efficient, they require bespoke low-level implementations for fast sample-by-sample generation.

In contrast, block-level models can be implemented using high-level machine learning frameworks, as overhead is less problematic when block size is large. In such vocoders, causal convolutions or block-level RNNs support a generative model based on normalizing flows or GANs [27] [18].

RAVE [6] is an autoencoder for raw audio which learns a continuous latent space of audio features. An adversarial loss term allows high-fidelity audio reconstruction despite a highly compressed representation. In this regard, RAVE is similar to neural codec models [29], however RAVE's latent representations are continuous and relatively interpretable. RAVE is both high bandwidth (44.1-48 kHz) and streaming via cached causal convolutions [7].

3. TUNGNAÁ

Our instrument consists of a neural text-to-voice synthesis engine, and a software GUI.

3.1 Voice Synthesis Engine

Tungnaá's synthesis engine is designed to meet the following requirements:

1. Real-time performance on a CPU, with latency below 100ms, suitable for interactive use within a live-coding paradigm.
2. Interactivity, with human-in-the-loop manipulation of the synthesis process.
3. Controllability, exposing the underlying neural synthesis engine to allow nuanced explorations of effects such as alignment failures, glitches and babbling.
4. Flexibility, with production of any sounds a human voice might make, without limitation to fluent speech or a single singing style.
5. "Hi-Fi" audio, representing frequencies up to 20 kHz with dynamic range suitable for digital music applications.
6. Openness, for users to train their own models and design their own text notations with small datasets.

3.1.1 Alignment Model

Per requirement (4), we choose a Tacotron2-like architecture which can learn alignments from utterance-level text and audio pairs, avoiding a forced alignment step which might fail when text is annotated with unconventional symbols. Specifically, we use Dynamic Convolutional Attention (DCA) [5], which mitigates the instability of purely content-based or location-sensitive attention but allows for creative (mis)use per (3).

Additionally, we increase expressivity using a neural spline flow [11] to model the density of audio features. This model quickly learns alignments and models diverse prosody without further conditioning.

3.1.2 Tokenless Text Encoder

Considering requirement (6), a pre-trained CANINE language model [9] is used as a text encoder. Unlike most language models, CANINE represents text as unicode characters rather than variable-length tokens, which simplifies the text-audio alignments from a UI perspective.

3.1.3 Real-time Vocoder

RAVE [6] is used as a streaming vocoder. Per requirement (5), it attains a high sound quality. Per requirements (6, 2), RAVE has both a high-quality open implementation of the training code, and streaming inference which is integrated with computer music workflows. Because RAVE learns a latent space of audio features, almost any input will decode to a speech-like sound, making RAVE's latent space more suitable for exploratory manipulation than spectrogram-based vocoders (3).

3.2 Model Training

We designed Tungnaá with the idea of pre-training on public speech datasets, then fine-tuning on small artist-created datasets. Thus far, we've mainly explored the former, however, we expect to include artist-made datasets by the time of demonstration.

For development, we use part of the Hi-Fi TTS audiobooks dataset [3], specifically the segment recorded by speaker

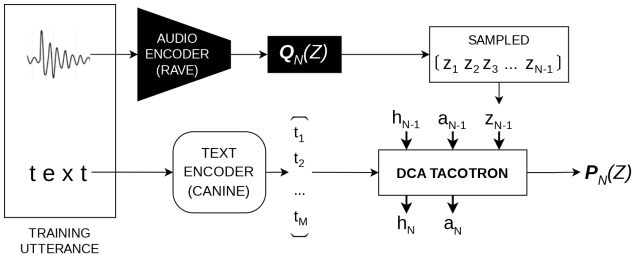

Figure 1: One audio frame of training. The pretrained audio encoder is frozen, while the pretrained text encoder is fine-tuned. A Tacotron2-like module predicts alignments a and audio feature distributions $P_N(z)$, conditioned on character encodings $t_1 ... t_M$ and previous audio features z_{n-1} via hidden state h_{n-1}.

9017 (John Van Stan) due to the large number of utterances, high quality of recordings, and expressivity of performance, which varies in style as the reader assumes different characters. We also experimented with the multi-speaker VCTK dataset [30].

To train a Tungnaá model, a RAVE vocoder is first fit to the audio part of the dataset. Then, the audio is preprocessed through the RAVE encoder. Finally, a text encoder and alignment model are fit to pairs of text and encoded audio (Figure 1).

3.3 Software Instrument

Tungnaá is a software instrument distributed as a Python package, combining our voice synthesis engine with GUI elements for input and display. It also exposes features to OSC control. A video demonstration is available in the accompanying materials.

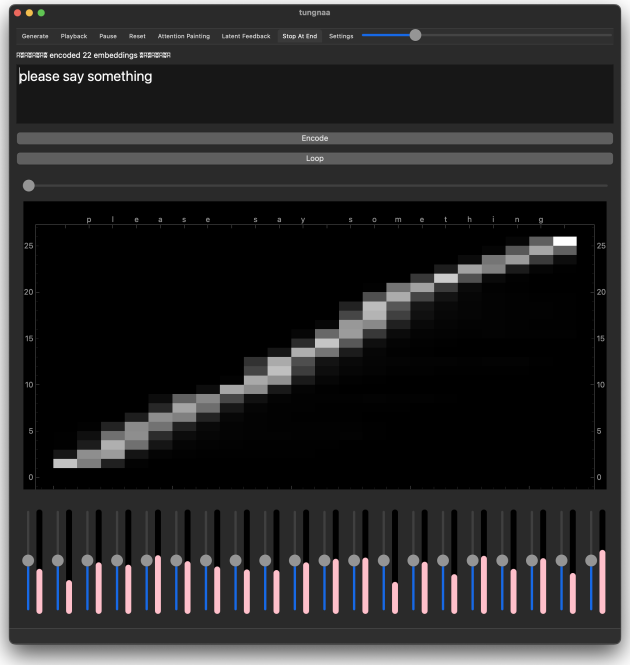

Figure 2: Tungnaá GUI. From top to bottom: option switches and temperature control; text entry field; encode and loop buttons; paint bar; scrolling text-audio alignment; vocoder meters and bias faders.

The Tungnaá GUI consists of three major elements: text entry, alignment, and vocoder (Figure 2). A text entry field allows the performer to prepare a short text for Tungnaá to vocalise. When ready, the performer can send it to the text encoder, which also resets the alignments.

A scrolling alignment graph depicts progress through the text over time. Once encoded, an input text appears along the horizontal axis of the graph, while time is on the vertical axis, with the present time at the top. Light pixels denote the portion of the text being used by the model at a given time. If *attention painting* mode is engaged, the *paint bar* allows the performer to directly manipulate this alignment. Otherwise, the model autonomously reads through the text according to its learned rhythms of speech.

A set of meters and faders display the RAVE latent vectors as they are produced by the alignment model, while a *temperature* control affects how variable they are. The performer can also apply a bias to each latent dimensions using the faders, to directly manipulate the sound. Each fader controls one aspect of voice sound as represented by the RAVE model. This is different for each model, but dimensions related to loudness, voicedness, frontalization of vowels and so forth can often be discerned. Biasing the vocoder this way does not affect the alignment model unless the *latent feedback* switch is engaged. Transformed latents are fed back into the alignment model, disrupting its progress through the text in aesthetically interesting ways.

We found that potentials of the autoregressive model came to light via the graphical interface, inspiring new features such as *forking paths*: with attention painting mode, the autoregressive generator continues in the context of previous manipulations. But with the forking paths feature, it can be rewound to a previous hidden state, to generate endless alternative takes of the utterance.

Finally, a sampler mode is in development, allowing previously generated material to be looped and sampled via text-based search over utterances.

4. CONCLUSION

Tungnaá represents a proof of concept for interactive artistic exploration of Deep VS, with several avenues for future research.

We have yet to fully explore the potential of the token-free text encoder for new unicode-based notation systems, which is exciting for areas where innovation in notation is important, such as live coding or spectromorphological analysis. While our design choices are deliberately style-agnostic, we have yet to demonstrate models for paralinguistic vocalisations or languages other than English.

As we have mainly focused on technical aspects of the system, future studies should address the creative dynamics of Tungnaá, asking how artists respond to a real-time version of their own or others' voices in different musical and cultural contexts. These are especially crucial questions to ask in this era of readily available vocal deepfake tools.

Finally, we wish to promote an open ecosystem around Tungnaá by providing tools to train models and design custom notations. This requires further research into techniques for efficient fine-tuning of small-dataset models from model pre-trained on larger corpora.

5. ACKNOWLEDGMENTS

This research is funded by the European Union from call CNECT/2022/3482066 – Art and the digital: Unleashing creativity for European industry, regions, and society under grant agreement LC-01984767. It is part of the S+T+ARTS

programme.

This research is also supported by the European Research Council (ERC) as part of the Intelligent Instruments project (INTENT), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101001848), and by an NVIDIA hardware grant.

6. ETHICAL STANDARDS

Tungnaá is dependent on voice recordings of actual people - a sensitive form of digital information, especially for vocal artists who are tradition bearers, have developed a particular craft or artistic brand. We endeavor to use only ethically sound voice data in the development of this work. Initial research was done using standard public English language speech datasets, such as VCTK [30] and HiFi TTS [4].

We are developing bespoke datasets in close collaboration with vocal artists, these data and models will not be released without their explicit consent.

7. REFERENCES

[1] A. Abbas, T. Merritt, A. Moinet, S. Karlapati, E. Muszynska, S. Slangen, E. Gatti, and T. Drugman. Expressive, variable, and controllable duration modelling in tts, Jun 2022. arXiv:2206.14165 [cs, eess].

[2] M. Astrinaki, N. d'Alessandro, and T. Dutoit. Mage –a platform for tangible speech synthesis. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Ann Arbor, Michigan, 2012. University of Michigan.

[3] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang. Hi-Fi Multi-Speaker English TTS Dataset, June 2021. arXiv:2104.01497 [eess].

[4] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang. Hi-fi multi-speaker english tts dataset. June 2021. arXiv:2104.01497 [eess].

[5] E. Battenberg, R. J. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby. Location-relative attention mechanisms for robust long-form speech synthesis, Apr 2020. arXiv:1910.10288 [cs, eess].

[6] A. Caillon and P. Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv:2111.05011 [cs, eess]*, Nov. 2021. arXiv: 2111.05011 version: 1.

[7] A. Caillon and P. Esling. Streamable Neural Audio Synthesis With Non-Causal Convolutions. *arXiv:2204.07064 [cs, eess, stat]*, Apr. 2022. arXiv: 2204.07064.

[8] Y.-P. Cho, F.-R. Yang, Y.-C. Chang, C.-T. Cheng, X.-H. Wang, and Y.-W. Liu. A survey on recent deep learning-driven singing voice synthesis systems. In *Proc. AIVR 2021*, page 319–323, Nov 2021.

[9] J. H. Clark, D. Garrette, I. Turc, and J. Wieting. CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Trans. ACL*, 10:73–91, Jan. 2022. arXiv:2103.06874 [cs].

[10] C. Czernowin. The primal, the abstracted and the foreign: Composing for the voice. *Contemporary Music Review*, 34(5–6):449–463, Nov. 2015.

[11] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural Spline Flows. *arXiv:1906.04032 [cs, stat]*, Dec. 2019. arXiv: 1906.04032.

[12] S. Fels and G. Hinton. Glove-talk ii - a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Transactions on Neural Networks*, 8(5):977–984, Sep 1997.

[13] H. R. Herndon. *PROTO*. PhD thesis, Stanford University, 2019.

[14] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu. Efficient Neural Audio Synthesis, Feb. 2018. arXiv: 1802.08435.

[15] M. Katz. Review of paul lansky. conversation pieces. *American Music*, 22(2):327–329, 2004.

[16] R. Kleinberger, N. Singh, X. Xiao, and A. v. Troyer. Voice at nime: a taxonomy of new interfaces for vocal musical expression. In *Proc. New Interfaces for Musical Expression*, Jun 2022.

[17] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai. Full-band lpcnet: A real-time neural vocoder for 48 khz audio with a cpu. *IEEE Access*, 9:94923–94933, 2021.

[18] A. Mustafa, J.-M. Valin, J. Büthe, P. Smaragdis, and M. Goodwin. Framewise wavegan: High speed adversarial vocoder in time domain with very low computational complexity, Mar 2023. arXiv:2212.04532 [cs, eess].

[19] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio, Sep 2016. arXiv:1609.03499 [cs].

[20] B. Pritchard and S. S. Fels. Grassp: Gesturally-realized audio, speech and song performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 272–276, Paris, France, 2006.

[21] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *arXiv:2006.04558 [cs, eess]*, Mar. 2021. arXiv: 2006.04558.

[22] J. C. Reus. iː goʊ weɪ. *AIMC 2023*, Aug 2023.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, Feb 2018. arXiv:1712.05884 [cs].

[24] G. Shopov, S. Gerdjikov, and S. Mihov. StreamSpeech: Low-Latency Neural Architecture for High-Quality on-Device Speech Synthesis. In *Proc. ICASSP*, pages 1–5, June 2023.

[25] N. Thapen. Pink Trombone.

[26] J.-M. Valin and J. Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *Proc. ICASSP*, page 5891–5895. IEEE, May 2019.

[27] B. Wu, Q. He, P. Zhang, T. Koehler, K. Keutzer, and P. Vajda. Fbwave: Efficient and scalable neural vocoders for streaming text-to-speech on the edge, Nov 2020. arXiv:2011.12985.

[28] C. Wu, Z. Xiu, Y. Shi, O. Kalinli, C. Fuegen, T. Koehler, and Q. He. Transformer-Based Acoustic Modeling for Streaming Speech Synthesis. In *Proc. Interspeech*, pages 146–150. ISCA, Aug. 2021.

[29] Y. Wu, I. D. Gebru, D. Marković, and A. Richard. Audiodec: An open-source streaming high-fidelity neural audio codec. In *Proc. ICASSP*, page 1–5, Jun 2023.

[30] J. Yamagishi, C. Veaux, and K. MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr

voice cloning toolkit (version 0.92). Nov. 2019.

[31] Z. Zukowski and C. J. Carr. Generating black metal and math rock: Beyond bach, beethoven, and beatles, Nov 2018. arXiv:1811.06639 [cs, eess].