Cloning voices and making kin: A multivocal AI approach to kinship

1st Ada Ada Ada, 2nd Stina Hasse Jørgensen, 3rd Jonas Fritsch
IT University of Copenhagen
Copenhagen, Denmark
adfo@itu.dk, shaj@itu.dk, frit@itu.dk

Abstract

Synthetic voice cloning tools are becoming increasingly accessible for artistic engagements. This paper contributes to unfolding the growing aesthetic potential of this development by 1) providing an overview of existing frameworks and approaches that might be used in artistic and design work, 2) situating voice cloning as a form of making kin by relating it to the concepts of kin networks, assembling and attachment sites and 3) presenting three speculative approaches to creating multivocal synthetic voices that seek to make kinships explicit.

Keywords

voice cloning, multivocal, making kin, speculative practice, AI in art, practice-based research

Introduction

Synthetic voices have been with us for a long time now. Initially as auxiliary and experimental features [5], before becoming the primary point of interaction in smart home products through Apple's Siri, Samsung's Bixby, Amazon's Alexa and more [29, 1, 3]. However, it is only recently that the design of these voices has become available outside of large companies and well funded research projects. AI voice cloning tools have rapidly changed this landscape, making it possible for artists, composers, producers and others to engage easily in creating and modifying synthetic voices. The potential ramifications for the creation of various kinds of media art and music are massive.

As voice cloning tools are becoming more and more accessible to artists, it becomes imperative to investigate the aesthetic potentials of these technologies. Research in speech and language processing currently seems to be focused on increasing the efficiency [38], emotionality [34], multilinguality [26] or "human-level quality" [33] of synthetic voices as also argued by Edward Kang [19] and Thao Phan [25]. These are not the primary concerns from an artistic perspective. An artistic investigation into voice cloning tools can be focused on investigating the potential for these technologies to create novel aesthetic experiences.

Synthetic voice design is not restricted to the anatomy of a human vocal tract, and therefore gives artists the space to play with and create radically different vocal outputs. In our research, we are interested in how the introduction of voice cloning tools to the artistic process can enable us to create what we term multivocal experiences[18] in our ongoing research project on vocal imaginaries [11]. By synthesizing multiple voices, we see the potential for representing multiple vocal identities in a single synthetic voice experience.

In this paper, we start by providing an overview of existing voice cloning frameworks and approaches that might be used in artistic work and research, including a distinction between text-to-speech and voice-to-voice tools. We also situate voice cloning as a form of kinmaking [14] in order to provide a potential frame for understanding the techno-aesthetic concerns inherent in voice cloning tools. Hereafter, we report on three experimental prototypes centered on kinmaking and voice cloning. Finally, we discuss some potential next steps in exploring the connections between making kin and cloning voices.

State of the art

The field of machine learning-based voice cloning has grown rapidly in the last few years. The available technologies can generally be divided into two areas: text-to-speech and voice-to-voice. In the following, we sketch out an overview to identify artistic openings and potentials in these developments.

Text-to-speech (TTS) voice cloning models are commonly trained on text-audio pairs. Specifically, this means that the training dataset consists of audio files and transcriptions of the utterances in those audio files. TTS models synthesize new utterances based on text input from the user. The development of text-to-speech machine learning-based voice cloning gained traction with the release of Tacotron [37]. This framework opened up the door for end-to-end text-to-speech synthesis. Since then, there has been a lot of development in the field with both the release of Tacotron2, Tortoise, and VITS [20, 4, 30]. Aside from these open source models, there has been a surge in commercially available voice cloning services from Microsoft, Google, Amazon, ElevenLabs and more [2, 8, 10, 12, 23]. Cloning a voice with TTS usually requires between 10 and 60 minutes of training data. Some research points to opening up the scope for TTS synthetic voices instead of seeking to make it more human-like, intelligible or efficient. The project [multi'vocal], for example, explores how synthesized voices can be produced differently [17]. In a similar vein, we find the "genderless" voice Q, which aims to provide an alternative to the standard TTS voices found in the home assistants of large tech companies

[31].

Voice-to-voice (V2V) technologies, also known as voice conversion (VC), work a little bit differently from the TTS approach, and are usually not trained on text-audio pairs, but instead purely on audio files. V2V software can be understood as a sort of voice filter that transforms the speech of an input audio file. V2V technologies come in both open source [27, 28] and commercial variants [35]. V2V approaches generally require less training data than TTS, and in some cases, they require no more than a minute to produce a decent clone. V2V technologies have been in use in the field of music production for a couple of years now. In 2021, the artist and researcher Holly Herndon announced Holly+, a freely available model of her own voice [15]. Voice cloning also made a big splash in mainstream news in April 2023, when ghostwriter977 released the song "Heart on My Sleeve" by cloning the voices of artists Drake and The Weeknd without permission [9]. A couple of weeks later, the artist Grimes announced that she would split her royalties with anyone making a song that uses her cloned voice [21].

Some software also allows users to produce other types of audio than speech via text-based interfaces. Yamaha's Vocaloid combines a piano UI with textual lyrics inputs, and has gained fame for being used to create the vocal identity of the virtual idol Hatsune Miku [36]. The software has recently been released with support for V2V-like features [39]. Another option for AI audio synthesis is a text-to-audio model like Bark [32], which can produce synthetic speech along with other paralinguistic and audio content.

A kinmaking approach to voice design

We argue that voice cloning tools can be situated as a way of making kin, a concept we borrow from ecofeminist scholar Donna Haraway. Kinmaking is a philosophical approach that emphasizes not just human connections, but also those that are more-than-human. Making kin is therefore not centered around biological relations, but rather on the attachments that emerge from being and belonging with each other. These attachments bring obligations, accountabilities and gratifications towards fellow kin. Nobody is kin to everything, so instead kinmaking sees kinships as existing in networks upon networks. We see three different ways in which these technologies can be viewed through a kinmaking lens in the Haraway sense; kin networks, assembling and attachment sites [13]. These three concepts are not distinct from each other, but overlap in a multitude of ways. In this paper, each of these concepts come to the fore in one of the following ways: through the process of fine-tuning a machine learning model, via our multivocal approach, and as an inherent part of the act of listening to voices.

Firstly, voice cloning generally happens through a process often referred to as fine-tuning. In practice, this works by calibrating a previously trained machine learning model to adapt to a new dataset. The primary reason for doing this is that the initial training of the model requires a large amount of data, while fine-tuning generally requires much less. However, this also means that a model trained through fine-tuning [40] builds on an understanding from the initial dataset. In other words, a fine-tuned voice clone consists not of one voice, but several different ones merged together relating to Haraway's thought on kin networks, which we understand as the multitude ways in which kin connect to each other.

Secondly, in our multivocal approach, we aim to make kinships explicit by designing a synthetic voice that represents multiple vocal identities. We thereby explore how engaging with one synthetic entity can present itself as an engagement with a multiplicity of identities and kinful relatives instead of merely a one-to-one experience. With several voice representations, we seek to make apparent the multifaceted relationships that synthetic voices put us in. This connects to the Haraway notion of assembling, in the sense that we assemble or gather multiple voices into one expressive entity. Through these sorts of multivocal experiences, we attempt to highlight that artificial intelligence voices are yet another tool that can make kin and tie us with our fellow earthlings.

Thirdly, we see the act of listening as a kinmaking experience and consequently synthetic voices become "attachment sites" [24]. Any type of voice should be understood as not just a singular entity, but rather a relation between listener and speaker. According to philosopher Adriana Cavarero [7], voice and listener become part of mutually defining each other as listening happens, and so even in terms of synthetic voices, we can therefore also expect to see kinful attachment forming between synthetic speaker and listener.

Designing a multivocal voice

In this section of the paper, we present our practice-based research process of designing a multivocal synthetic voice using TTS voice cloning tools. We define multivocal as that which is many-voiced, or speaking as multiple roles or identities. Like Jørgensen [18], we are interested in the paralinguistic materiality of multivocality.

We will introduce three different speculative approaches to creating multivocal synthetic voices. Each voice design type provides a unique set of opportunities, challenges and aesthetic potentials for making kin explicit. To make it easier to compare and understand their varying potentials, we have named each of the three approaches: The Choral Voice, The Pooled Voice and The Fluctuating Voice. [1]

All three multivocal voice types deal only with the dataset itself. Their differences appear in how we construct that dataset, whether through preprocessing or varied forms of data assemblage. We use the same neural network setup, i.e. the Tacotron2 Google Colab Notebook by justinjohn0306 [16], to train the voices. They were all created using data from the public domain audiobook website, Librivox [22]. The voice models build on approximately 30 minutes of speaker data.

The Choral Voice

This version of the multivocal voice is created by layering three editions of the same voice on top of each other. One of the voices is pitched down, another is pitched up and the third voice is the default pitch. This somewhat emulates the

[1] Examples of the synthetic voices as well as the underlying dataset can be heard here: `https://on.soundcloud.com/ntM6TgubDDyENXey8`

idea of a choir, which is why we named this approach The Choral Voice. This approach is similar to that of Unisong [6], but with two main differences: we use a machine learning approach instead of a concatenation one, and we focus on spoken voice synthesis instead of singing voice synthesis.

After a couple of hours of training, the result of the voice cloning is a synthetic voice that sounds quite a lot like the original data. One of the main downsides by this approach, however, is that the intelligibility of both the original voice data and the resulting synthetic voice is very low. It is hard to tell what is being said with this type of multivocal voice. However, we were positively surprised that the voice cloning software was able to faithfully reproduce audio with multiple voices inside it.

From an aesthetic point of view, The Choral Voice approach shows that the machine learning model is quite capable of picking up the connection between text and audio, even when the training data has low intelligibility. This opens up the opportunity for manipulating the original dataset even more to create novel ways of synthesizing voices. As long as it is possible to establish a statistical connection between utterance and text, the voice cloning tool seems capable of reproducing the original in the synthesized version.

In kinship terms, The Choral Voice embodies the concept of assembling. Both its dataset and its final synthesized version consists of multiple voices that have been assembled in time, as can be seen in figure 1. They speak not as one identity, but as a multitude.

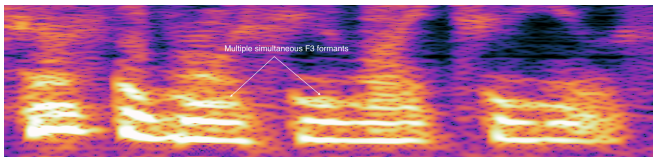

Figure 1: *A spectrogram of a Choral Voice utterance showing how the multiple formants of the choir are replicated in the synthesized speech.*

The Pooled Voice

In The Pooled Voice, the dataset consists of audio files from two different speakers. The two speakers are actually the same, but one of them is pitched down. With this approach, we have decided to go in a different direction than what is generally recommended for voice cloning. Generally, the recommendation is to have one speaker per dataset, because the machine learning model attempts to find statistical patterns in the provided utterances. By mixing multiple voices into the dataset instead, we encourage the software to consider both of the voices as part of the same synthetic voice. On the surface, this seems similar to how multi-speaker/multi-voice synthesis frameworks like Tortoise [4] work. The main difference is that, in those approaches, the different speakers are separated into different blocks inside the dataset.

The resulting synthetic voice does not manifest as a combination of both speakers. Instead, when synthesizing new utterances, the model tends to choose one of the two voices to speak with, as can be seen in the two spectrograms in figure 2. It seems as if the model first decides which voice is most likely to be speaking the initial segment of the generated audio, and this decision sets the precedent for the rest of the audio. This makes sense considering that a model like this is more or less an ”autocorrect for sound”. It starts by creating the first piece of audio, and then piece-by-piece adds more and more to match the provided sentence.

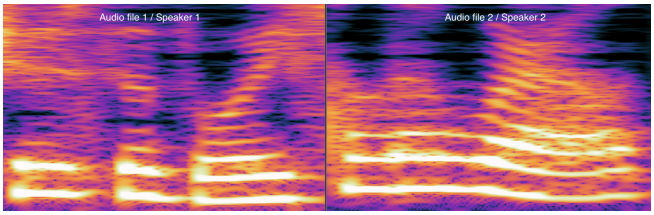

Figure 2: *Two spectrograms from synthesized Pooled Voice audio files. Speaker 1 is the pitched down speaker, while speaker 2 is the standard pitch.*

Artistically, The Pooled Voice has limited potential. In the end, it seems to act as a sort of random picker of voices. The artist has no direct control over which voice gets produced, and as such a certain level of control is given up to the machine learning model. Yet since machine learning is based on statistical inference, this loss of control cannot simply be replaced by an auxiliary random function. Its reliance on statistical likelihood means that the software chooses the voice based on the similarity between input text and training data. Therefore, adding more voices and more data could make the results more complex to an interesting degree.

As for its relation to making kin, The Pooled Voice reveals itself as a kinful attachment site with each new utterance. As every sentence starts with a new voice, the listener is put in contact with a new synthetic identity, emphasising that the interaction is manifold.

The Fluctuating Voice

Finally, we introduce The Fluctuating Voice. This voice design approach builds on a dataset with two different speakers. However, whereas The Pooled Voice splits the different speakers into separate audio files, The Fluctuating Voice instead puts both speakers into the same audio files. For our experiments into this approach, we used two completely different speakers reading different scripts. The audio files do not contain a complete 50/50 split between speaker 1 and speaker 2, but the total amount of audio from both speakers is more or less equal.

The end result of The Fluctuating Voice is a synthetic voice that switches between both speakers in the middle of an utterance. The speaker usually changes between words, as can be seen in figure 3, but in some cases, the shift occurs inside a word pronunciation. When the switch happens in the middle of a word, the shift can be audibly heard as a type of modulation between the two voices.

The Fluctuating Voice seems to have a lot of aesthetic potential. The way that the voice switches in the middle of an

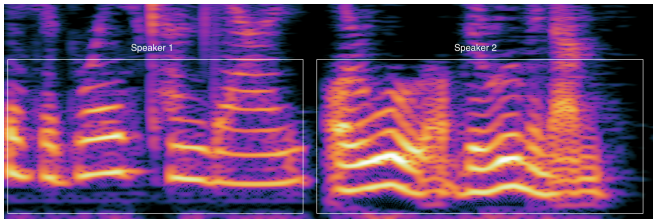

Figure 3: *An annotated spectrogram of a synthesized utterance with The Fluctuating Voice. In this case, the shift between speakers happens in between a word.*

utterance is quite unique to synthetic voices, and is hard to reproduce in traditional audio software. The artist does not really have any control of when and how the voice shifts from one to the other, but this loss of control can again be quite interesting as an artistic tool. Leaving the voice change up to statistical probability opens up opportunities for surprising and serendipitous vocal experiences.

This third multivocal approach builds kin networks concurrently throughout a synthesized utterance. It shifts, bends and moves between different voice expressions and identities, and as such it reveals not just that the synthetic voice is an attachment site for kinships, but also that it actively builds and restructures networks between kin.

Discussion & conclusion

The purpose of this paper has been to introduce three speculative approaches to making kin explicit with multivocal synthetic voices. Specifically we have focused on voice cloning through artificial intelligence text-to-speech technologies. In our multivocal approach, we make kinships explicit by designing a synthetic voice that represents multiple vocal identities. Each approach builds on a different dataset construction, and each construction provides significantly different aesthetic potentials. With our multivocal synthetic voice experiences, we argue that engaging with one synthetic entity is no longer presented as a one-to-one experience, but rather an engagement with a network of vocal identities and relatives. With several voice representations wrapped into one, we make apparent the multifaceted relationships that synthetic voices put us in. Through these sorts of multivocal experiences, we highlight that artificial intelligence voices are yet another tool that builds kinship and ties us with our fellow human and more-than-human kin. Our preliminary experiments in this area have shown promise in terms of artistic research potential. So far, we have only worked with two voices at a time, and it would be an interesting next step to build on this with more and different voices to see how we can make kinship even more explicit and wide reaching in AI voice cloning contexts.

References

[1] Amazon. 2020. Alexa Privacy: How Does Alexa Work? https://www.youtube.com/watch?v=esMOQgDMAeo.

[2] Amazon. 2023. Amazon Polly Features. https://aws.amazon.com/polly/features/.

[3] Apple. 2023. Siri. https://www.apple.com/siri/.

[4] Betker, J. 2023. Better speech synthesis through scaling.

[5] Black, A. W., and Lenzo, K. A. 2003. Building synthetic voices. *Language Technologies Institute, Carnegie Mellon University and Cepstral LLC* 4(2):62.

[6] Bonada, J.; Blaauw, M.; Loscos, A.; and Kenmochi, H. 2006. Unisong: A choir singing synthesizer. In *Proceedings of the 121st AES Convention, San Francisco, USA*. Citeseer.

[7] Cavarero, A. 2005. *For more than one voice: Toward a philosophy of vocal expression*. Stanford University Press.

[8] Coqui. 2023. Coqui. https://coqui.ai/.

[9] Coscarelli, J. 2023. An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. *The New York Times*.

[10] ElevenLabs. 2023. ElevenLabs - Generative AI Text to Speech & Voice Cloning. https://elevenlabs.io/.

[11] Fritsch, J.; Jørgensen, S. H.; Ada, A. A.; and Knudsen, S. L. 2022. Voice as a matter of design: a framework for novel vocal imaginaries. https://vocal.itu.dk/.

[12] Google. Text-to-Speech documentation. https://cloud.google.com/text-to-speech/custom-voice/docs.

[13] Haraway, D. 2015. Anthropocene, Capitalocene, Plantationocene, Chthulucene: Making Kin. *Environmental Humanities* 6(1):159–165.

[14] Haraway, D. J. 2016. *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

[15] Herndon, H. 2021. Holly+. https://holly.plus/.

[16] justinjohn0306. 2023. justinjohn0306/FakeYou-Tacotron2-Notebook.

[17] Jørgensen, S. H.; Baird, A.; Juutilainen, F. T.; Pelt, M.; and Højholdt, N. C. 2018. [multi'vocal]: reflections on engaging everyday people in the development of a collective non-binary synthesized voice. In *Proceedings of EVA Copenhagen 2018, Aalborg University, Copenhagen, Denmark, 15 - 17 May 2018*.

[18] Jørgensen, S. H. 2020. *Vocal Bodies: Performing Paralinguistic Stereotypes and Multivocalities in Art and Digital Media*. PhD Thesis, University of Copenhagen.

[19] Kang, E. B. 2022. Biometric imaginaries: Formatting voice, body, identity to data. *Social Studies of Science* 52(4):581–602.

[20] Kim, J.; Kong, J.; and Son, J. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.

[21] Legaspi, A. 2023. Grimes on AI Songs: 'Feel Free to Use My Voice Without Penalty'. *Rolling Stone*.

[22] Librivox. LibriVox | free public domain audiobooks. https://librivox.org/.

[23] Microsoft. 2023. Train your custom voice model - Speech service - Azure AI services. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-voice-create-voice.

[24] Paulson, S. 2019. Making Kin: An Interview with Donna Haraway.

[25] Phan, T. 2017. The Materiality of the Digital and the Gendered Voice of Siri. *Transformations (14443775)* (29).

[26] Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; Baevski, A.; Adi, Y.; Zhang, X.; Hsu, W.-N.; Conneau, A.; and Auli, M. 2023. Scaling Speech Technology to 1,000+ Languages.

[27] Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; and Hasegawa-Johnson, M. 2019. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss.

[28] RVC-Project. 2023. Retrieval-based-Voice-Conversion-WebUI. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI.

[29] Samsung. 2023. Bixby | Apps & Services. https://www.samsung.com/uk/apps/bixby/.

[30] Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R. J.; Saurous, R. A.; Agiomyrgiannakis, Y.; and Wu, Y. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.

[31] Simon, M. 2019. The Genderless Digital Voice the World Needs Right Now. *Wired*.

[32] Suno AI. 2023. Bark. https://github.com/suno-ai/bark.

[33] Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; Soong, F.; Qin, T.; Zhao, S.; and Liu, T.-Y. 2022. NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality.

[34] Tang, H.; Zhang, X.; Wang, J.; Cheng, N.; and Xiao, J. 2023. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis.

[35] Uberduck. 2023. Uberduck | Make Music with AI Vocals. https://www.uberduck.ai/.

[36] Verini, J. 2012. How Virtual Pop Star Hatsune Miku Blew Up in Japan. *Wired* 20(11).

[37] Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; and others. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[38] Xiao, Y.; Zhang, S.; Wang, X.; Tan, X.; He, L.; Zhao, S.; Soong, F. K.; and Lee, T. 2023. ContextSpeech: Expressive and Efficient Text-to-Speech for Paragraph Reading. In *INTERSPEECH 2023*, 4883–4887.

[39] Yamaha Corporation. 2022. Yamaha New Comprehensive Vocal Synthesis Software VOCALOID™6. https://www.yamaha.com/en/news_release/2022/22101301/.

[40] Yu, F. 2016. A Comprehensive guide to Fine-tuning Deep Learning Models in Keras (Part I) | Felix Yu. https://flyyufelix.github.io/2016/10/03/fine-tuning-in-keras-part1.html.