The Lena Singer Project: Simulating the Learning Experience of a Singer

Matthew Rice[1] and **Simon Colton**[1,2]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]SensiLab, Faculty of Information Technology, Monash University, Australia

m.rice@se22.qmul.ac.uk s.colton@qmul.ac.uk

Abstract

The Lena Singer project involves a generative AI process based on a recent singing voice synthesis system that iteratively produces audio to simulate a singer learning how to sing. Users can select an initial motivation and an initial ability for the singer, then, through a feedback-based process involving random elements, the singer may improve at singing, or they may get worse, which in turn boosts or diminishes its confidence, ability, and motivation. In this way, we aim to provide a simple model which simulates the learning experience of a human singer and demonstrate how it differs from standard machine learning approaches. We also explore the feedback loop that learning can have on internalized features, and exemplify how a machine might express the output of this learning. Finally, we discuss how the context provided by this process can be seen as relatable. A demo of this project can be found at `https://lena-singer.vercel.app`.

Introduction and Background

Since the introduction of artificial neural networks, the goal has been to emulate human learning by modeling the brain and the composition of its neural connections (Russell and Norvig 2010). However, even current deep learning or reinforcement learning systems can't model some of the complex ways particular factors or experiences can affect decision-making and learning (Simplilearn 2022). Moreover, emulation systems are designed to approximate some human capabilities, but there is a stark difference in how these systems are trained (by minimizing some objective functions) versus how a human would learn through various experiences and situations. This is particularly seen in the fundamental differences in biological and artificial intelligence (Korteling et al. 2021).

When people learn, important factors related to self-theories, i.e., people's beliefs about themselves, can have a huge impact. In the influential human psychology book "Self-theories: Their Role in Motivation, Personality, and Development", Dweck frames students' response to failure into two categories: 'helpless' and 'mastery-oriented' (Dweck 2000). The students in the helpless category, although at the same initial ability as the others, tended to lose motivation and give up, whereas the mastery-oriented students would tend to be resilient. However, sorting into these two categories seems to directly relate to the students' learning goals, which are determined before they even start learning. Other research in (Druckman and Bjork 1994) suggests self-confidence is related to one's perception of ability, and may play a central role in how one learns skills over time. Moreover, these variables seem to be part of a feedback loop, as confidence can increase motivation, which in turn can increase ability and this can cyclically increase confidence (Bénabou and Tirole 2005).

We aim here to simulate some of these internal factors and demonstrate how their initial state may play a substantial role in an agent's ability to learn a creative skill such as singing. Learning to sing well can take years, and often involves thousands of hours of repetitive performance and analysis of one's own singing voice as it would compare to others. This process can require a great deal of patience and continued motivation to keep practicing. Moreover, a lack of belief in oneself, due to low confidence, is a common cause of mistakes in singing (Ni Riada 2019).

Recent advances in generative deep learning have produced high-quality singing voice synthesis (SVS) systems. These systems can generate audio of a realistic human singing voice from musical scores and lyrics. In this project, we use VISinger2 (Zhang et al. 2022), a recent high-quality SVS system. We combine this with a step-based probabilistic learning model that we heuristically construct to converge to particular outcomes. The outputs of the probabilistic learning model are used to corrupt the inputs and outputs of the singing model to alter the singer's ability at each step. Via a web-based front end, users can define the starting state of the model and watch as it either improves at singing or ultimately performs worse. Our contributions include:

1. The development of a novel initial framework for simulating aspects of human learning, built on top of an existing controllable AI generation system.
2. The exploration of how a machine can make mistakes and how these may differ from those made by people.
3. A demonstration of how the learning process of a relatable creative AI system may reassure users and encourage them to build a narrative about the process.

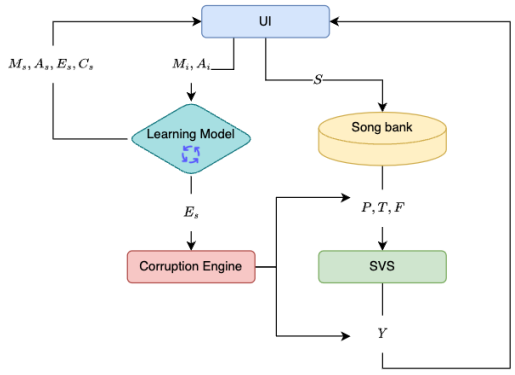

Figure 1: Overview of the Lena Singer system which shows the major components and data flow. P, T, F represent phonemes, phoneme timings, and frequencies respectively. M, A, E, C represent motivation, ability, mistakes, and confidence. Y represents the output audio.

System Description

The Lena Singer system is composed of four parts: the *singing model*, the *learning model*, the *corruption engine*, and the *user interface*. Figure 1 presents a diagram of how the component parts work together in the overall system.

Singing Model

The underlying SVS system is VISinger2 (Zhang et al. 2022), an end-to-end model that creates a realistic singing voice in Mandarin Chinese from an input of phonemes, phoneme timings, and phoneme frequencies. At its core, the model is a conditional variational autoencoder with a discriminator. The model encodes phonemes, phoneme timings, and frequencies to a latent space and similarly encodes the mel spectrogram (O'Shaughnessy 2000) of an associated audio file. During training, the objective is to minimize the distance between the latent encodings of related phonemes and spectrogram representations. The model also includes a decoder, which uses a DDSP (Engel et al. 2020)-inspired process to generate a mel spectrogram, which is converted to the waveform domain using a modified HiFiGan (Kong, Kim, and Bae 2020) model. Following the decoder, a discriminator is simultaneously trained to ensure high-quality outputs that are similar to the training examples.

The developers of VISinger2 used the Opencpop (Wang et al. 2022) dataset to train the model. This consists of 100 Mandarin songs sung by a professional singer with human-labeled annotations. The annotations include the lyrics, the phonemes, the notes, the note durations, the phoneme durations, and whether or not a note was a slur note. We obtained permission to use the Opencpop dataset and trained our own version of the model accordingly, using a single A100 GPU, for 200k steps. The fine-grained control of phonemes and pitches allows for precise and realistic manipulation of the resulting audio. As part of the Lena Singer system, this SVS system can generate one of five songs from the Opencpop test set, with each song being around seven seconds long. Importantly, the model is used purely to generate audio from the inputs given to it by the corruption engine and its outputs are not fed back into the learning model. Thus it could be easily swapped for any other SVS system and/or dataset providing it follows the same format as the Opencpop dataset.

Learning model

In overview, we aim to implement a system able to simulate the human learning process, using variables that represent relatable concepts. Specifically, we develop a simple stochastic model for how factors of motivation, confidence, ability, and mistakes during performance interplay over time, as an agent practices the creative act of singing. Overall, the learning model consists of seven variables with a range of 0-100 which are updated at every step, based on the values of the other variables. They represent both the internal state of the singer (ability, motivation, confidence, mistake factor, mistake history factor) and the external results (mistakes/mistake history). All variables are updated in a stepwise fashion and are normally distributed with a standard deviation of 5. This allows for some randomness, hence different results from the same starting point. Algorithm 1 shows the pseudocode for the variable updating scheme. We designed this model with three goals in mind:

1. To simulate how a large amount of initial motivation or ability should usually be able to overcome a low amount of the other. Low amounts of both variables should usually lead to failure, while high amounts should usually lead to success.
2. To ensure that the system is non-deterministic, so two sessions with the same initial values of ability and motivation may lead to different processes and outcomes.
3. To simulate how learning to sing well should take more timesteps than giving up and stopping.

The variables in the learning model are defined as follows: **Ability**: the singer's natural ability. It is inversely related to mistakes but also affected by current motivation. **Motivation**: The singer's interest in continuing to learn. A low motivation will cause the singer to stop trying to learn and give up. **Confidence**: how confident the singer is in its abilities. A high confidence will cause the singer to stop trying to learn because it believes it is good enough. **Mistakes**: how corrupted the output audio will be. **Mistake Factor**: how much the current mistakes value matters to the singer. **Mistake History**: how many past mistakes the singer remembers. **Mistake History Factor**: how much these past mistakes matter to the singer. The mistake factor, mistake history, and mistake history factor variables were designed to allow the model to learn to overcome and ignore their current and previous mistakes as a representation of resiliency.

Although the variables and update equations are not explicitly derived from any true academic model, they were initially inspired by human learning, then fine-tuned and weighted to fit our goals. They rely on a feedback model of learning where internal state variables are updated from combinations of other internal variables and external results.

Algorithm 1 Updating scheme for the learning model

let $N(x) = \mathcal{N}(x, 5)$
motivation $m \leftarrow m_{init}$
ability $a \leftarrow a_{init}$
confidence $c \leftarrow 0$
mistakes $e \leftarrow 0$
mistake factor $mf \leftarrow 0$
mistake history $h \leftarrow []$
mistake history factor $hf \leftarrow 0$
step $n \leftarrow 0$
while $n < 30$ **do**
  $e \leftarrow N(100 - 0.55a - 0.45m)$
  $h.append(e)$
  $h.resize(10 - c/10)$
  $c \leftarrow N(a - e * mf)$
  $a \leftarrow N(100 - e + m * n)$
  $m \leftarrow N(a - \sum h * hf)$
  $mf \leftarrow m * 0.01$
  $hf \leftarrow (100 - m/100) * 0.01$
  $n \leftarrow n + 1$
  if $c = 100$ *or* $m = 0$ **then**
    $break$
  end if
end while=0

For instance, confidence is derived from the agent's current ability as well as current mistakes multiplied by a mistake factor. Over a maximum of 30 steps, the model will converge to one of three outcomes. If confidence continues to stay near 100, the model will stop to signify that it has reached its goal. Or, if motivation hovers near 0, the model will stop to show it is giving up. Finally, if the model reaches the maximum number of steps, the model will stop to denote that it is finished learning. The model itself depends solely on these variables and does not rely on feedback from the corruption engine, SVS, or the UI.

Corruption Engine

The corruption engine produces the mistakes the singer is making. There are nine corruptions split into pre- and post-generation corruptions. The pre-generation corruptions affect the inputs to the VISinger2 model. These include overall speed change, inter-phoneme timing change, and random phoneme replacement. By reducing or increasing the phoneme timings, either by a global amount or a varying amount, the speed of the singing is affected without changing the pitch. The post-generation corruptions are high and low-pass filtering, compression, distortion, random pitch detune, and gain reduction. Besides random pitch detune, which is implemented directly, all other effects are implemented using Pedalboard (Sobot 2021). There is also a reverb effect added that inversely follows mistakes to frame the perfected singer as more professional. Uniform distributions are used to regulate each corruption's activation and intensity, which are parameterized by the mistakes metric or, in the case of gain reduction, the confidence metric. The ranges for these distributions are chosen heuristically.

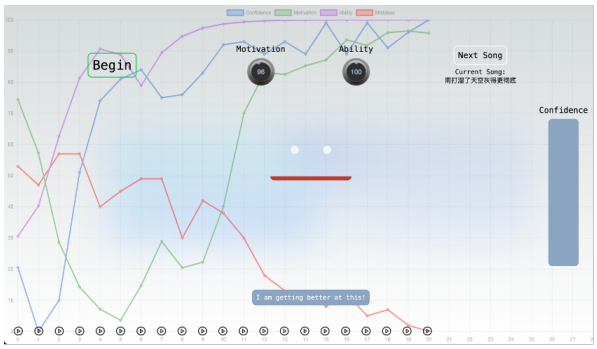

Figure 2: The web interface for the Lena Singer system. This run had an initial motivation of 74 and an initial ability of 31. Here, the agent successfully learned how to sing.

User Interface

The GUI of the system enables users to set the initial motivation, ability, and song and displays the output data from the learning model and SVS model. The motivation, ability, confidence, and mistakes metrics are plotted on a graph for each time step, while the audio output appears at the bottom of the screen as a button. The GUI will also display some text to reflect the current state of the learning run. Figure 2 shows an example session in the GUI.

Experimental Results

We evaluate here both our learning model and the corruption engine. For all evaluations, we used a default song, as song choice has no impact on the learning model. To evaluate the learning model, we ran 200 simulations of the model with random initial states of motivation and ability, sampled on a uniform distribution from 0-100. Figure 3 shows if the run was a success as well as how long it took based on the initial variables of ability and motivation. Overall, it seems like our initial goals for the learning model have been achieved in terms of sensible outcomes from certain starting conditions. In particular, a high ability or high motivation will likely lead to success, while mediocre or poor metrics lead to failure. Also, it seems there are many cases with a similar initial state that lead to different results. Furthermore, in general, it seems like achieving the task of singing takes longer than failure. However, it is interesting that having a high initial ability and motivation does not seem to generally converge to success quicker than in other cases.

As a straightforward way to evaluate the corruption engine, we measure the multi-resolution STFT (Steinmetz and Reiss 2020) error between the clean audio (no corruption) and the output audio at every time step for a particular run with initial motivation = 86 and initial ability = 25. Figure 4 plots these metrics. It seems that the mistakes graph somewhat correlates with the multi-resolution STFT error (PCC of 0.778). Therefore, the system reduces the quality to some extent depending on the mistakes metric. However, we argue that there is no requirement for a perfect linear relationship in this emulation system.

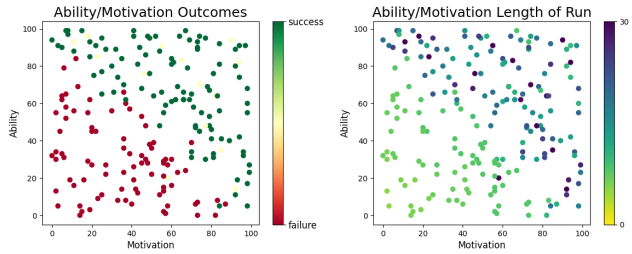

Figure 3: Success/failure and number of steps taken for 200 initial states of motivation and ability.

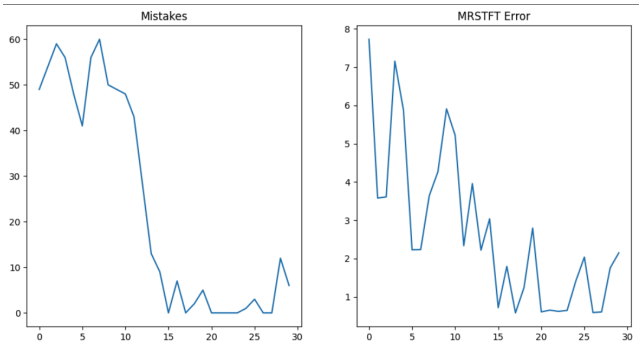

Figure 4: Mistakes and MRSTFT error between corrupted audio and clean audio on each learner model step.

Discussion

Although this project represents a toy example of human learning, it achieves a result that is empirically much closer and more relatable to a user's actual experience learning a skill than standard machine learning approaches. Each simulation run produces a unique result in terms of the learner model variable journeys, as well as the output audio on each step. When running a simulation with the system, a user can experience and empathize with the learning agent, potentially building up a mental story as to why a particular outcome might have occurred. For instance, a run that begins with high confidence and high motivation, but high mistakes, may see the confidence and motivation drop as the errors increase, but then slowly start to decrease the mistakes, leading to an increase in confidence. This could cause the user to imagine a singer initially frustrated with their mistakes, but with enough resilience to overcome them and gain confidence. These runs may cause the user to feel more connected to both the intermediate and final outputs. However, as suggested by (Colton, Pease, and Saunders 2018), barriers exist to truly make these "life experiences" challenging for people to accept, unless we are able to more accurately reflect the audience or alternatively use an enactive AI (Froese and Ziemke 2009) approach, with a completely different social and cultural environment.

While the learning model variables and update scheme were chosen arbitrarily, they seem to control the system such that the results are mostly expected, but sometimes surprising. The variables also present a feedback loop, as an increase in one usually leads to an increase in another. Although the learning model itself is significantly more interpretable than a standard deep learning approach, it is still not entirely clear how variable updates can cause particular outcomes. The corruptions selected for this system were primarily chosen for their ease of implementation. However, their outputs are significantly different from the consequences of human errors. While people cannot physically manipulate their voices with audio effects, some corruptions generate unnatural outputs. Nonetheless, there are some similarities between the two, e.g., it is common for a human singer to rush through a song or sing more quietly due to nervousness or lack of confidence. We didn't strive to completely model human mistakes but attempt to illustrate some initial examples to inspire future discussion.

Related Work

To the best of our knowledge, modeling motivation, ability, and confidence in a learning model for a creative skill like singing hasn't been studied from a computational creativity perspective. However, there is related work on the creative process and intrinsic motivation, e.g., (Salge, Glackin, and Polani 2014) explores intrinsic motivation in AI, and builds a 3D simulation to explore a mathematical definition of agent empowerment as an example. In addition, (Guckelsberger, Salge, and Colton 2017) describes an enactive framework to map an AI's intrinsic values and goals to its creativity. Unlike our system, they develop a non-anthropocentric model and aim to study creativity from the bottom-up.

Earlier work on music generation with creative agents in (Miranda 2003) explored granular synthesis through imitation agents. However, the agents themselves don't have any sense of embodiment or self-awareness. (Linkola et al. 2017) focus on self-awareness as it relates to metacreativity and "the capability to reflect on one's own creative processes and adjust them". Their model defines key aspects of self-awareness that are useful even for non-metacreative systems, namely artifact-awareness and goal-awareness. (Ford and Bryan-Kinns 2023) study the aspect of reflection in people using creativity support tools, and suggest that it is an important part of self-expression. Finally, (Cook et al. 2019) discuss the idea of framing in computational creativity, defined as "providing a narrative context for the actions and motivation of the software". They conclude that projects that include descriptions of the underlying processes can help audiences to relate to them.

Conclusions and Further Work

We presented a reasonable computational simulation of people learning how to sing. In particular, we developed a novel learning model and data corruption engine that attempts to model particular aspects of human learning in a feedback loop. We combined those modules with a recent controllable SVS model to synthesize realistic human singing, using the corruption engine to modify the inputs and outputs of the SVS model to portray the human learning process. We discussed the design decisions for our system and showed how the system meets our intended goals through the evaluation of experimental results. Furthermore, we described how such a system could be seen as relatable to a user. This

system is only a first step for future systems studying human and machine learning through the lens of creative practice.

In general, this framework itself needs refinement and iteration. As a first step, future systems could attempt to have a richer and more scientifically accurate model of human learning for creative tasks. For example, in the previously cited work (Dweck 2000), researchers found that, while confidence is a good predictor for academic achievement, it doesn't help students in difficult situations. Alternatively, a model could be developed that would be more in line with enactive AI with intrinsic motivation, with a clear design to allow the model to have intentional creative agency that adapts to its environment. Moreover, while the corruptions provide a solid baseline, future work could explore either the idea of closer modeling of human mistakes or could posit novel corruptions that non-anthropocentric creative agents could explore. We would also like to emphasize the conclusions of (Shneiderman 2020) which suggest that assumptions from tool-like application systems, such as virtual assistants, should not be directly applied to emulation systems like the Lena Singer system. These should be treated and designed separately to avoid poorly crafted designs since they usually have separate goals. Finally, due to the advent of AI systems that can realistically mimic particular human abilities like singing, painting, etc..., we are particularly interested in work that follows a similar framework with different underlying generative AI technologies.

Author Contributions

Matthew Rice theorized and developed the Lena Singer system and web demo. Simon Colton provided advice and direction for this work. Matthew Rice wrote the majority of the manuscript, while Simon Colton contributed to the writing and editing of the manuscript.

References

Bénabou, R., and Tirole, J. 2005. Self-Confidence And Personal Motivation. *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumptions* 19 – 57.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of Authenticity in Autonomously Creative Systems. *Proc. ICCC* 272–279.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing In Computational Creativity – A Survey And Taxonomy. *ICCC* 156–163.

Druckman, D., and Bjork, R. A. 1994. *Learning, Remembering, Believing: Enhancing Human Performance*. Washington, D.C.: National Academies Press.

Dweck, C. S. 2000. *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. arXiv:2001.04643.

Ford, C., and Bryan-Kinns, N. 2023. Towards a reflection in creative experience questionnaire. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery.

Froese, T., and Ziemke, T. 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* 173(3):466–500.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "Why?" in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency. *Proceedings of the 8th International Conference on Computational Creativity (ICCC'17)*.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv:2010.05646.

Korteling, J. E. H.; van de Boer-Visschedijk, G. C.; Blankendaal, R. A. M.; Boonekamp, R. C.; and Eikelboom, A. R. 2021. Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence* 4.

Linkola, S.; Kantosalo, A.; Mannisto, T.; and Toivonen, H. 2017. Aspects of Self-awareness: An Anatomy of Metacreative Systems. *Proceedings of the 8th International Conference on Computational Creativity (ICCC'17)*.

Miranda, E. R. 2003. On the evolution of music in a society of self-taught digital creatures. *Digital Creativity* 14(1):29–42.

Ni Riada, A. 2019. 3 Singing Mistakes That Are Killing Your Confidence. https://www.confidenceinsinging.com.

O'Shaughnessy, D. 2000. *Speech Communications: Human and Machine*. Wiley.

Russell, S. J., and Norvig, P. 2010. *18.7 Artificial Neural Networks*. Prentice-Hall, 3rd edition. 727–737.

Salge, C.; Glackin, C.; and Polani, D. 2014. Changing the Environment Based on Empowerment as Intrinsic Motivation. *Entropy* 16(5):2789–2819.

Shneiderman, B. 2020. Design Lessons From AI's Two Grand Goals: Human Emulation and Useful Applications. *IEEE Transactions on Technology and Society* 1(2):73–82.

Simplilearn. 2022. Artificial Intelligence vs. Human Intelligence | Simplilearn. https://www.simplilearn.com/artificial-intelligence-vs-human-intelligence-article.

Sobot, P. 2021. Pedalboard. 10.5281/zenodo.7817839. https://github.com/spotify/pedalboard.

Steinmetz, C. J., and Reiss, J. D. 2020. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*.

Wang, Y.; Wang, X.; Zhu, P.; Wu, J.; Li, H.; Xue, H.; Zhang, Y.; Xie, L.; and Bi, M. 2022. Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. http://arxiv.org/abs/2201.07429.

Zhang, Y.; Xue, H.; Li, H.; Xie, L.; Guo, T.; Zhang, R.; and Gong, C. 2022. VISinger 2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer. http://arxiv.org/abs/2211.02903.