Novelty Search on Vocal Models

Joris Grouwels[1], Nicolas Jonason[1], Bob L.T. Sturm[1]
[1]Division of Speech, Music and Hearing (TMH), KTH Royal Institute of Technology, Sweden
grouwels@*kth.se*, njona@*k*th.se, bobs@kth.se

Abstract

Knowing which sounds can be produced by a simulated vocal model is not trivial. Being able to map this out can be interesting for applications that make use of the extended capabilities of a voice, e.g. singing. Novelty search is proposed as a method to explore acoustic capabilities of articulatory models, leveraging the representations of a human-like auditory perception model.

Introduction

Knowing which sounds a voice can make is not an easy question to answer, but important for singers, especially when they are still in training. It can be a help in deciding on what you should spend your practice time. If you know for sure that your physiology does not afford the production of certain kinds of sounds, then you can stop striving for it, and conversely – maybe even more importantly – if you know that a sound is possible, but you do not know how to make it yet, it is an incentive for vocal exploration.

Despite computer-implemented articulatory-acoustic models being more observable than the vocal mechanism of a living person, their acoustic potentials are not easily deducible from their implementations because of the non-linear relationships between the articulatory control parameters, their acoustic consequences and human auditory perception. Thus, when studying these models in terms of their suitability to emulate singing, one is faced with a similar question as the living singer: What sounds can this model actually produce? Some limitations might be obvious from the model specifications (e.g. only vowels), while others (e.g. What timbres are within reach?) might be much more difficult to answer by only looking at the model design.

In contrast to methods that attempt to optimize model parameters to fit a certain auditory target (e.g. the vowel *a*) we propose using *novelty search* (Gomes, Mariano, & Christensen, 2015; Lehman & Stanley, 2011) to automate the exploration of the model parameter space to find sounds perceived as meaningfully different. To conduct a novelty search, we have to specify a dissimilarity measure. As we are aiming for the domain of (human) singing, this measure should preferably be informed by human auditory perception. Machine hearing as the modeling of human auditory perception is a field of study in itself, with a good introductory text in Lyon (2017). Finally, this method produces a corpus of sounds exhibiting the sonic variety that the model can produce, together with the input parameters that produce them.

There are several benefits to producing sound and parameter collections that attempt to acoustically characterize a vocal model. Most obviously, having such a corpus enables a characterization and possibly a taxonomy in terms of what the model can do. Furthermore, through the analysis of overlap and differences, these corpora can form a basis for comparing one model with other models as well as with corpora of human utterances. Studying whether similar sounds were produced by similar articulatory configurations and vice versa might also be revealing. If an articulatory voice model makes a sufficiently close approximation of a real human voice, this would hold the promise of giving a useful pointer to human singers as to how to achieve this sound with their own instruments.

In the following, we first shortly review some related work and then delve into the three components of this experiment: Articulatory vocal models, novelty search, and human-like perception models. Then, we give some specifics of the experiments we intend to perform and of which we might be able to report some preliminary results at Fonetik 2024. While some of the named benefits can apply to both articulatory and non-articulatory vocal models, we limit our scope to the former category. We end with a discussion and conclusion.

Related work

As an articulatory-acoustic model can be seen as a synthesizer with a specific parameter set, exploring the sound space of a musical synthesizer is conceptually the same problem as the one described above. Novelty search has been used in that context to generate diverse sounds (e.g. Masuda & Saito, 2023), albeit in a sound matching context.

In the realm of speech, there have been several attempts to emulate the process of baby-babbling, in which a baby explores its vocal endowments and learns to produce syllables. This process is simulated with the Maeda (1979, 1990) model by Moulin-Frier et al. (2014) and by Philippsen (2021) in the context of Vocal Tract Lab (Birkholz, 2013). While these systems implement goal-directed and intrinsically motivated exploration, they do not attempt to explore model potential as such.

Non-articulatory Singing Voice Synthesis has been around since at least 1977 with the KTH MUSSE system (Sundberg, 2006). In those early years it was mainly a research instrument for analysis by synthesis aimed at perception and musical performance research. While MUSSE started as an analog signal processing system driven by a rule-based performance system, nowadays the field is dominated by dataset driven neural approaches (Cho et al., 2021; Cui et al., 2024; e.g. Katahira, Adachi, Tai, Takashima, & Takiguchi, 2020; Shimizu et al., 2022; Sugahara et al., 2023) and achieves a very high degree of naturalness, which allows it to be used in digital music production.

Articulatory-acoustic models

Kröger (2022) surveys computer-implemented articulatory models for speech since the 1960s. Where these models originally had the goal of producing high-quality speech through replicating the human speech organs, they have long been outperformed by non-articulatory models in this domain. Nowadays they mainly serve as a research tool. These models have control parameters that correspond to configurations of the articulators, which can be either static positions or dynamic trajectories. Of the 21 articulatory models described in the survey, we are interested only in the 10 models with an acoustic component, i.e. that can produce sound. It should be mentioned that all these models are meant primarily for speech research, so certain design choices might have been made that do not suit singing very well. This would be another reason to explore and analyze their capabilities.

For our purposes, the models should have an existing, fast implementation that is available for research use, as novelty search requires many iterations. Examples of implementations are the Maeda model as a part of DIVA (Guenther, 2006; Tourville & Guenther, 2011)[i] and Vocal Tract Lab [ii], but also *gnuspeech*[iii] and Thapen's *Pink Trombone*[iv].

Novelty search

Lehman and Stanley (2008, 2010, 2011) promote novelty search as an alternative to objective driven optimization in deceptive problem spaces, i.e. problems where the solver tends to get stuck in local minima. In their evolutionary algorithms, they use *novelty* as a criterion instead of a typical fitness function that measures proximity to a goal. Just like a typical optimizing fitness function, novelty is defined in the phenotypical space. In our case we choose the auditory perception domain as the phenotypical space, further called “perception space”. Contrary to typical fitness, novelty is defined in relation to neighboring points in a space, not in relation to a fixed objective.

Let p be an articulatory parameter vector for a vocal model VM. Through the vocal model, we obtain a sampled acoustic waveform $x = VM(p)$. The perception model PM then projects x into the perception space $z = PM(x)$. We then define the novelty $\rho(z)$ at some evolutionary iteration as the sparseness of its neighborhood with respect to sounds so far generated by the model. We adapt the definition of sparseness by Lehman & Stanley (2011), thus defining the novelty z as the mean distance to its k nearest neighbors:

$$\rho(z) = \frac{1}{k}\sum_{i=1}^{k} dist\,(z, \mu_i) \quad (1)$$

where $dist(\cdot,\cdot)$ is a dissimilarity measure defined in the perception space. The neighbors $\mu_i$come from the current population generated by the evolutionary algorithm, as well as from a possible archive of sounds that has been kept over the course of the calculations. As $\rho(z)$ increases, the closest neighbors are at a larger distance, meaning that the current sound collection is sparser around point z in perception space.

The novelty search algorithm tries in every iteration to maximize the sparseness for every produced sound. In a genetic algorithm, this will make the population diverge to different regions of the perception space.

As mentioned before, one can keep an archive of previously calculated sounds, which can be included as neighbors in the sparsity calculation. Sounds can be included in the archive on different grounds e.g. based on a novelty threshold, or at random with a certain probability (Gomes et al., 2015).

The intended result of the novelty search algorithm is an exploration of the goal-space that should be significantly more effective than a random walk and that does not get stuck in local minima.

Novelty search might require a considerable number of evolutionary generations to give a good idea of a model’s capabilities. Initially, therefore, we must work with fast vocal and perception models, and parallelize the procedure as much as possible.

Human-like perception models

With our focus on singing, it is important that perception be focused on music, not on speech. For example, two or three formants might be enough to characterize speech vowels, but not for conveying the timbral resolution that is necessary for singing.

As mentioned in the previous section, we need a dissimilarity measure that is not defined in the articulatory space, but in perception space. In order to aim for sounds that are meaningfully different to humans, we need an auditory dissimilarity measure that reflects human perception.

One way to achieve this is to leverage methods that aim to directly emulate the signal processing in human hearing as measured by psychoacoustic experiments, e.g. the gammatone (Patterson, Allerhand, &Riguère, 1995) and gammachirp (Irino & Unoki, 1998) filterbanks. Another strategy is to model the functional components of the auditory system, like Lyon’s (2017) CAR-FAC and SAI models, that emulate the cochlea and the auditory representation in the brain stem. More recently, deep learning strategies with the same goal have appeared (e.g. Baby, Van Den Broucke, & Verhulst, 2021).

In the previously mentioned studies on baby-babbling (Moulin-Frier et al., 2014; Philippsen, 2021), the authors make use of ad-hoc auditory perception models. Moulin-Frier et al. (2014) devise a strategy that includes only the intensity and the first two formant frequencies, while Philippsen (2021) uses three formants for her experiments with static sounds. When working with dynamic vocal trajectories, she starts out with MFCC features. As with the filterbanks above, this entails that samples of different length also tend to be considered as very different. Philippsen mitigates this by embedding these features with a small RNN and a subsequent dimensionality reduction by PCA and LDA.

The recent appearance of the deep learning-based CLAP-model (Elizalde, Deshmukh, Al Ismail, & Wang, 2023; Elizalde, Deshmukh, & Wang, 2024; Wu et al., 2023) trained on massive amounts of sound and text captions might offer another way of tackling this problem. Its embeddings (fixed length vector representation of audio) offer an audio representation that is cheap to compute and relates to human perception through the human involvement in the selection of the sound recordings in its dataset as well as through their text captions. This might be one of the first things to try.

Forthcoming experiments

As this study is still in an early stage, we cannot show any experimental results yet. We hope to be able to show some preliminary results in the conference presentation. Similar to Lehman & Stanley (2010) we intend to implement novelty search based on a genetic algorithm, for which we will have to define the mapping between the genome and the articulatory control parameters that affords the genetic operations of crossover and mutation in a meaningful way. We will start with static vowels, and subsequently possibly expand our scope to consonants and more dynamic situations. Apart from listening to the generated sound samples, more rigorous evaluation could comprise statistical analyses of the audio features present in the generated corpus, as well as the extent to which the articulatory input parameter ranges have been used.

Discussion

One question that only the experiments can answer is whether novelty search will work in high-dimensional perceptive goal spaces. Lehman & Stanley (2011) show that increasing the dimensionality does not hurt performance in their use case, but the NEAT algorithm they are using (Stanley & Miikkulainen, 2002) evolves behavioral neural architectures. In their case, it is the behavior exhibited by these architectures that should be novel. In addition, NEAT also has a built-in tendency to generate architectures that increase in complexity over time.

Secondly, our hope is that the generated corpora could be a way to investigate the quantal theory of speech (Stevens, 1989). According to this theory, there are certain regions in the articulatory space that are acoustically stable in the sense that larger parameter changes do not entail big changes in sound. Conversely, there are also regions where smaller parameter changes have a big impact. As a result, aiming well into the middle of the stable regions makes speech, and possibly even singing, robust to inevitable noise in vocal control. Last but not least, we hope that the resulting corpora might become good datasets to train inverse dynamics for these articulatory models, meaning that a machine learning model that is given a sound efficiently could generate the parameters that would produce that sound.

Conclusion

In this report on work in progress we have presented the motivation behind and the ingredients of our coming experiments with novelty search to characterize the singing potential of articulatory-acoustic vocal models.

Acknowledgements

This work is an outcome of a project that has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (MUSAiC, Grant agreement No. 864189).

References

Baby, D., Van Den Broucke, A., & Verhulst, S. (2021). A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nature Machine Intelligence*, *3*(2), 134–143. doi: 10.1038/s42256-020-00286-8

Birkholz, P. (2013). Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *PLOS ONE*, *8*(4), e60603. doi: 10.1371/journal.pone.0060603

Cho, Y.-P., Yang, F.-R., Chang, Y.-C., Cheng, C.-T., Wang, X.-H., & Liu, Y.-W. (2021). A Survey on Recent Deep Learning-driven Singing Voice Synthesis Systems. *2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 319–323. doi: 10.1109/AIVR52153.2021.00067

Cui, J., Gu, Y., Weng, C., Zhang, J., Chen, L., & Dai, L. (2024). Sifisinger: A High-Fidelity End-to-End Singing Voice Synthesizer Based on Source-Filter Model. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11126–11130. doi: 10.1109/ICASSP48485.2024.10446786

Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023). Clap learning audio concepts from natural language supervision. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Elizalde, B., Deshmukh, S., & Wang, H. (2024). Natural language supervision for general-purpose audio representations. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 336–340. IEEE.

Gomes, J., Mariano, P., & Christensen, A. L. (2015). Devising Effective Novelty Search Algorithms: A Comprehensive Empirical Study. *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, 943–950. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2739480.2754736

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Irino, T., & Unoki, M. (1998). A time-varying, analysis/synthesis auditory filterbank using the gammachirp. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, *6*, 3653–3656 vol.6. doi: 10.1109/ICASSP.1998.679675

Katahira, K., Adachi, Y., Tai, K., Takashima, R., & Takiguchi, T. (2020). Opera Singing Voice Synthesis Considering Vowel Variations. *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, 865–866. doi: 10.1109/GCCE50665.2020.9291895

Kröger, B. J. (2022). Computer-Implemented Articulatory Models for Speech Production: A Review. *Frontiers in Robotics and AI*, *9*, 796739. doi: 10.3389/frobt.2022.796739

Lehman, J., & Stanley, K. O. (2008). Exploiting Open-Endedness To Solve Problems Through The Search For Novelty. *Artificial Life XI: Proceedings of the 11th International Conference on the Simulation and Synthesis of Living Systems, ALIFE 2008*. Retrieved from https://stars.library.ucf.edu/scopus2000/9505

Lehman, J., & Stanley, K. O. (2010). Efficiently evolving programs through the search for novelty. *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, 837–844. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1830483.1830638

Lehman, J., & Stanley, K. O. (2011). Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation*, *19*(2), 189–223. doi: 10.1162/EVCO_a_00025

Lyon, R. F. (2017). *Human and Machine Hearing: Extracting Meaning from Sound*. Cambridge: Cambridge University Press. doi: 10.1017/9781139051699

Maeda, S. (1979). An articulatory model of the tongue based on a statistical analysis. *The Journal of the Acoustical Society of America*, *65*(S1), S22–S22. doi: 10.1121/1.2017158

Maeda, S. (1990). Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 131–149). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-009-2037-8_6

Masuda, N., & Saito, D. (2023). Quality-diversity for Synthesizer Sound Matching. *Journal of Information Processing*, *31*, 220–228. doi: 10.2197/ipsjjip.31.220

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2014). Self-organization of early vocal development in infants and machines: The role of intrinsic motivation. *Frontiers in Psychology*, *4*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2013.01006

Patterson, R. D., Allerhand, M. H., & Giguère, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, *98*(4), 1890–1894. doi: 10.1121/1.414456

Philippsen, A. (2021). Goal-Directed Exploration for Learning Vowels and Syllables: A Computational Model of Speech Acquisition. *KI - Künstliche Intelligenz*, *35*(1), 53–70. doi: 10.1007/s13218-021-00704-y

Shimizu, S., Matsubara, K., Adachi, Y., Tai, K., Takashima, R., & Takiguchi, T. (2022). Comparative Evaluation of Neural Vocoders for Speech Synthesis of Operatic Singing. *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, 28–29. doi: 10.1109/LifeTech53646.2022.9754796

Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, *10*(2), 99–127. doi: 10.1162/106365602320169811

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*(1), 3–45. doi: 10.1016/S0095-4470(19)31520-7

Sugahara, A., Kishimoto, S., Adachi, Y., Tai, K., Takashima, R., & Takiguchi, T. (2023). Operatic Singing Voice Synthesis Using Diff-SVC. *2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, 776–777. doi: 10.1109/GCCE59613.2023.10315526

Sundberg, J. (2006). The KTH synthesis of singing. *Advances in Cognitive Psychology*, *2*(2), 131–143.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981. doi: 10.1080/01690960903498424

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. doi: 10.1109/ICASSP49357.2023.10095969

[i] https://sites.bu.edu/guentherlab/software/diva-source-code/

[ii] https://www.vocaltractlab.de

[iii] https://www.gnu.org/software/gnuspeech

[iv] http://dood.al/pinktrombone or reimplemented on https://github.com/zakaton/Pink-Trombone