Effects of Added Vocals and Human Production to AI-composed Music on Listener's Appreciation

Liam Dallas and Fabio Morreale

School of Music
University of Auckland
New Zealand

liamdallas.nz@gmail.com
f.morreale@auckland.ac.nz

Abstract. We investigated the extent to which listeners' appreciation for AI-composed music would change with the addition of vocals and human-production. By combining a variety of commercially available and ad-hoc created software, we created four songs entirely composed by forms of machine intelligence. We then created 4 different conditions: with vs. without vocals (with lyrics generated by an AI), and human vs. AI production. We performed an experimental study with 40 participants and found that the added vocals did not improve listeners' appreciation. Human-produced songs were also not better appreciated than AI-produced ones. By analyzing participants' comments, we discuss possible reasons behind these results.

Keywords: AI-music; music perception; listener appreciation; vocals

1 Introduction

In recent years, the quality of AI-generated music has taken large steps forward. Experiments as Magenta,[1] Flow Machines,[2] and Open AI's MuseNet,[3] as well as commercial solutions like Jukedeck[4] and Landr[5] are arguably bringing the quality of AI systems close to that of human instrumentalists and producers. However, despite the steady growth in quality of AI-composed music, AI-created vocals are yet to believably replicate a human vocalist. Many advancements are being made, including the recent DeepSinger, but "the synthesized singing voices do not have as rich and diverse expressiveness and emotion as human voices" (Ren et al., 2020). These AI voices can also trigger unease in listeners with the audio equivalent of the Uncanny Valley (Avdeeff). We hypothesized that this limitation, combined with the quality of

[1] https://magenta.tensorflow.org/
[2] https://www.sonycsl.co.jp/tokyo/2811/
[3] https://openai.com/blog/musenet/
[4] https://www.jukedeck.com/
[5] https://www.landr.com/

commercial AI-productions, might hinder the success of AI music in mainstream media. We set up an experimental study to test this hypothesis. We used an array of existing and ad-hoc developed AI and software systems to compose four pop-rock pieces. For each piece, we created 4 versions: with human-sang-AI-written vocals vs. without vocals, and with AI-production vs. human-production. We then conducted an online survey with 40 participants to test whether the tracks with added vocals were more enjoyable to the listener. The quantitative results disproved our hypothesis: listeners preferred the tracks with no vocals. The qualitative results offered a rationale for this finding: even if an AI system could convincingly create human sounding vocals, there is one more important element missing: the lyricist. The results also showed that AI-produced songs were generally more appreciated than human-produced ones, though, as we reflect in the Discussion, this result might have to do with inherent limits of our productions. We conclude suggesting that the roadmap for AI songwriting should include new methods to develop *credible* lyrical content.

2 Experiment

2.1 Song preparation

In order to generate the experimental material for this study, we needed solutions that allowed us to create compositions with and without vocal parts while having as little human input as possible. To the best of our knowledge, currently no systems exist that cover all aspects of an AI-composed and -produced song with vocals. SampleRNN-based sytems such as Dadabots (Carr & Zuckowski, 2018) can generate pieces of music with vocal sounds, but the lyrics are mostly "nonsensical syllables, as the model does not learn a language model", making these unsuitable for creating versions with a human vocalist. Therefore, we had to use multiple systems that could work together.

We eventually chose the following systems for their ability to write parts that fit a recognized verse-chorus structure, so that each part could be developed independently from each other and yet being able to be integrated together. AIVA[6], a commercial application that generates short songs in different genres, provided both the algorithmic compassion and the algorithmic production and seemed to offer the most reliably structured tracks that would be suitable for the addition of vocals. We then used "TheseLyricsDoNotExist.com" to generate the lyrics, which provided lyrics in a verse-chorus structure that could align with AIVA's A-B section structure. This tool requires a theme to be chosen so we generated four *love* songs to provide consistency between compositions. This consistency avoids the subject matter standing out to participants, as love songs make up 50-60% of popular music (Keen & Swiatowicz, 2007).

[6] https://www.aiva.ai/

The next task was to algorithmically create a melody to fit a given text and give chord progression. We were unable to find any off-the-shelf tool, so we developed our own simple algorithm. Some related work (Genchel, Pati & Lerch, 2019) use harmonic information along with RNN to generate melodies, but they don't use syllabic input to fit to text. The algorithm we devised takes 4 lines of text and fits it over 8 bars of music to fit the structures of generated material. It then generates a rhythm based on the syllabic content of the text. It subsequently fits chord tones to strong beats based on the harmony that is manually provided, and then fills the gaps with various melodic devices. The algorithm then adapts to tonal areas by choosing notes from the current and surrounding chords. In cases where neither a natural note nor its alteration are present, the algorithm chooses the note that is most consonant with the sounding chord based on a given hierarchy. An example of a melody generated from this algorithm is shown in Figure 1, formatted as the sheet music we provided to a vocalist (see details below).

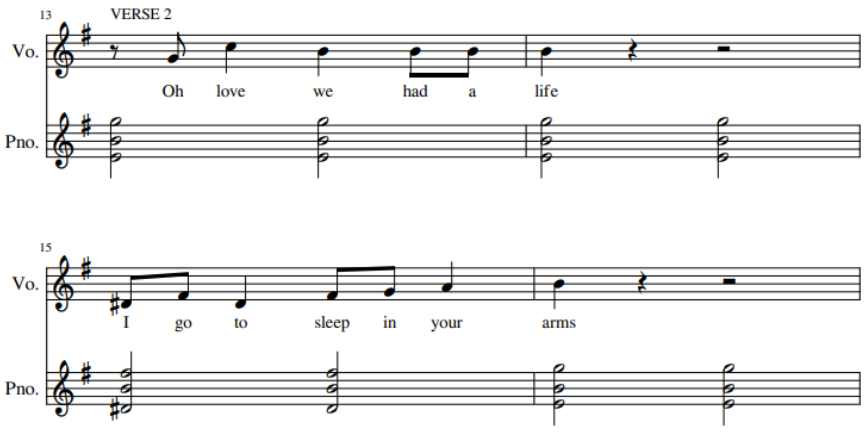

Fig. 1. Excerpt from the vocal sheet music for track 4 displaying a melody generated by the algorithm developed for this study. It contains the melody, lyrics and block harmony.

In order to avoid any possible selection bias, no generated tracks were discarded from this experiment. The human-produced versions of the tracks were recorded using the same categories of instruments as the AIVA-generated samples – i.e. programmed drums, human-performed bass and electric guitars. The production style was pop-rock focused. We also performed a few typical production operations, such as using stereo guitars to balance the spectral and stereo images. All versions of the tracks are available at the following link tinyurl.com/y52mlyuv. The below letter codes correspond with the different production modes and the numbers correspond to the composition.

- AN : AIVA-produced song, instrumental
- PN: Human-produced song, instrumental
- AV: AIVA-produced song, with vocals
- PV: Human-produced song, with vocals

2.2 Experimental study

After having sought and obtained consent from the University of Auckland Human Participants Ethics Committee (Ref 024616), we conducted an online survey to find out which versions of the tracks were preferred by listeners. We created four variations of the survey: each included a song from all four production modes, and each song was heard in its four modes across the groups. This means that each song was heard independently of its other modes, and each mode was heard independently of the other songs in that mode
Forty participants were directed to one of the four versions of the survey. The first question asked them to self-identify whether they are a musician or non-musician. Having participants to self-identify their musical abilities has a few advantages over empirical music sophistication indexes for more user-friendly experience and for the difficulty of quantifying some parameters like years of musical experience can be difficult (Rickard & Chin, 2017).

The survey asked the participants to rank each piece of music on five categories using 1-7 likert scales, where 1 is least so and 7 is most so. The scales had the following labels:

1. Enjoyable;
2. Emotional;
3. Unique;
4. Memorable;
5. Overall Quality.

Finally, in order to obtain qualitative insight into the data, we offered participants an opportunity to write a short comment after rating each song to motivate their answers. Notably, our participants were unaware of the nature of the study. Nowhere in the study description or in the Participant Information Sheet, did we mention that an AI system was used in the creation of the music.

2.3 Hypothesis

We expected the presence of a human vocalist to provide a relatable, human aspect to the music that an AI system may not be able to supply. Thus, we expected the addition of a vocals to improve the scores of each piece of music on all scales, especially those of Enjoyable and Emotional. We had similar hypothesis with respect

to the human-produced versions of the songs, which we believe would have exceeded AI-produced songs in most categories.

2.4 Quantitative analysis

A one-way ANOVA (assuming equal variances) was used to test for statistical significance of responses to the 1-7 scales. We found two statistically significant results. The first was that the tracks with vocals were less enjoyable than those without ($p < 0.005$). The mean for tracks with no vocals was 4.26, and for with vocals was 3.09.

In the Unique category, the tracks with vocals were found to be more unique ($p < 0.05$) and in the Overall Quality category, tracks with vocals scored lower ($p < 0.0005$). No significant results were identified between the two groups (musicians vs. non-musicians). The mean for tracks with no vocals was 3.35, and for with vocals was 4.20.

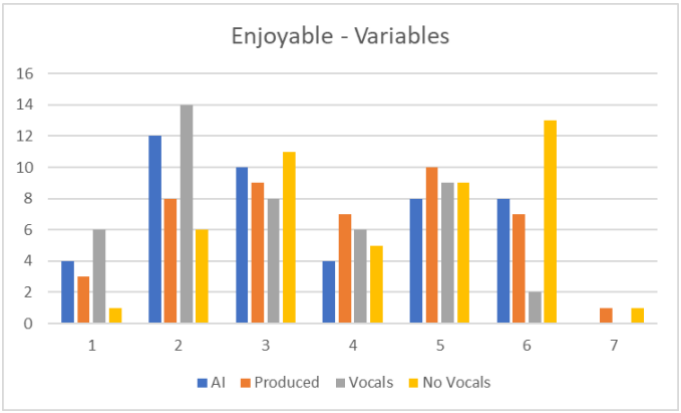

Fig. 2. Chart of enjoyable scores with AI-Produced and Vocals-No Vocals grouped

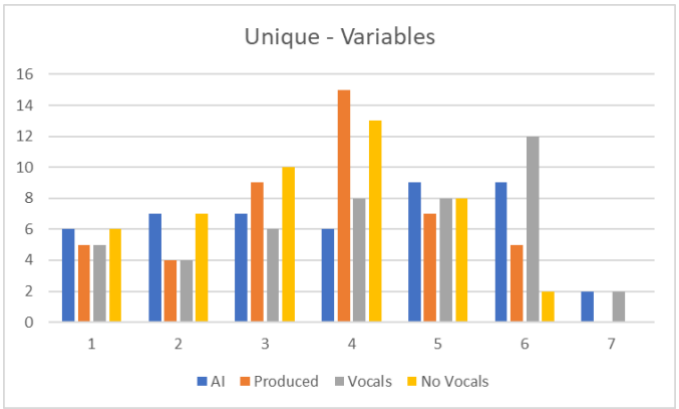

Fig. 3. Chart of unique scores with AI-Produced and Vocals-No Vocals grouped

2.5 Qualitative analysis

The comments offered by the participants provided some insight into the reasons behind participants' ratings. Over a third of all comments mentioned the lyrics in a negative light. This is particularly important as we did not prompt them to comment on the lyrics. Some of these comments included: "Saying emotional things doesn't make it emotional", "Love was mentioned so repetitively that it lost meaning", and "sounds like an alien observing mankind and trying to make a sad garage rock song". This quote is an apt metaphorical description of what was happening, as machine learning models do, in a way, watch mankind and imitate what they find to be important.

3 Discussion

The quantitative analysis disproved our initial hypothesis about adding vocals that would improve the ratings. The qualitative results offered a partial explanation of reasons behind it. These results, though preliminary, reveal an aspect that is possibly lacking in AI music creation, especially in relation to mainstream genres such as pop or rock music: the lyrics.

When designing the experimental study, we overlooked the semantic and emotional aspect of the lyrics, which carry value on its own, independent of the vocal performance. An alternative study would have been to compare the vocal performance to current singing voice synthesis models that would be able to perform our written music. This alternative might have led to more unbiased comparison as participants would have arguably reduced the weight of lyrics' semantic on listeners' appreciation. A similar study also might offer new knowledge about priorities for this domain. For example, the timbral characteristics may be less important than the "sequential nature of speech" (Nakamura et al., 2020). This sequential nature is noticeable in how models begin and end sung phrases as well as the transitions between sung notes.

Our failure to recognize this distinction proved valuable however, allowing us to reveal the paramount importance of lyrical content in this style of music.

Another unexpected result was the lack of a significant difference between the AI and Produced tracks. However, our qualitative data offered little evidence about why that was the case despite our original assumption that the produced tracks would perform better. We hypothesize it might have to do with the coherency of the AIVA-produced tracks, which have a straightforward and consistent musical style due to their MIDI libraries, and this coherency could play across to listeners as musical intent and confidence. Another important factor impacting the enjoyability of the human-produced tracks may have been production quality. First, due to Covid-19, the production studios at the University of Auckland were closed so the tracks had to be home-produced. Second, due to these circumstances, session musicians could not be hired, which would have increased the overall quality of the played parts (especially

programmed drums comparing to a professional drummer). Most comments about the vocals focused on lyrical content rather than performance itself, so this would be unlikely to change the results regarding the addition of vocals. While there was no significant result here, as always one may have arisen with a larger sample size also.
Finally, it is unsurprising that the tracks with vocals were found to be more unique, considering the idiosyncratic lyrics that were used. The lyric generation tool we used was a machine learning system, and as with all machine learning systems, "caution must be taken when discussing what these systems have actually learned to do" (Sturm, Bentahal, Monagham & Collins, 2018). To what level this system understands how lyrics are written is unknown, but it seems to lack long-term coherency past the level of words repeated across different phrases. Each lyric indeed on its own makes sense, but when all taken together, they have very weird interactions, or lack thereof. It is safe to assume that our participants have so far almost exclusively listened to music written by humans, and the lyrics being the musical element that least accurately mimics a human lyricist, makes them the most unique (but not in a good way) to listeners. With respect to the overall quality, we postulate the results being directly connected to the enjoyment factor.

4. Conclusion and future work

In this study we investigated whether the presence of absence of vocal tracks impact listeners' appreciation of AI-composed tracks. Given the abovementioned limitations with our experimental set-up we could not offer definitive results on that respect. However, we serendipitously found that the content of the lyrics overtakes the presence of vocals when determining listeners' appreciation. Thus, we propose that lyric generation is as one of the most important areas of development for the future of any music entirely composed by AI.
Future investigations into the effects of vocals into AI-music appreciation might have different results if the tracks were professionally produced. A variant of the study could be conducted with lyrics written by human lyricists. This adaptation would possibly subtract some emphasis from the lyrics themselves and add more to the effect of the presence of vocals. Future work could involve variants of this study with human lyricists, professional production, and more participants. Another variation could be the use of a singing voice synthesis program to remove the variable of lyrics entirely. This would focus the results on the presence of a human vocalist and possibly provide insight into the value of development around singing voice synthesis.

Acknowledgements

Thanks to the University of Auckland for funding this research (Faculty of Creative Arts and Industries FRDF grant 3719326).

References

Ren, Y., Xu Tan , Tao Qin , Jian Luan , Zhou Zhao & Tie-Yan Liu (2020). DeepSinger: Singing Voice Synthesis with Data Mined From the Web. *In Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD '20),

Keen, C., & Swiatowicz, C. (2007). Love still dominates pop song lyrics, but with raunchier language. *University of Florida News, 31.*

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv:2005.00341*.

Rickard, N.S, & Tanchyuan Chin. (2017) Defining the Identity of "Non Musicians". *Handbook of Musical Identites*. Oxford University Press.

Genchel, B., Pati, A., & Lerch, A. (2019) Explicitly *Conditioned Melody Generation: A Case Study with Interdependent RNNs*. From Proceedings of the 7th International Workshop on Musical Metacreation (MUME 2018)

Carr CJ & Zuckowski, Z. (2018) Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands. Proceedings of the 6th International Workshop on Musical Metacreation (MUME 2018).

Avdeeff, M. (2019). Artificial intelligence & popular music: SKYGGE, flow machines, and the audio uncanny valley. In *Arts* 8:4. Multidisciplinary Digital Publishing Institute.

Sturm, B., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D. et al. (2018). Machine Learning Research that Matters for Music Creation: A Case Study. *Journal of New Music Research*, Routledge.

Nakamura, K., Takaki, S., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2020). Fast and High-Quality Singing Voice Synthesis System based on Convolutional Neural Networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7239-7243). IEEE.