Predicting and Critiquing Machine Virtuosity: Mawwal Accompaniment as Case Study

Fadi M Al-Ghawanmeh, Melissa J Scott, Mohamed-Amine Menacer, Kamel Smaïli

Predicting and Critiquing Machine Virtuosity: Mawwal Accompaniment as Case Study

Fadi M. Al-Ghawanmeh
Music Department, University of Jordan, Jordan
SMarT Group, LORIA, F-54600, France
f_ghawanmeh@ju.edu.jo

Melissa J. Scott
UC Berkeley Department of Music, USA
melissascott@berkeley.edu

Mohamed-Amine Menacer
SMarT Group, LORIA, F-54600, France
mohamed-amine.menacer@loria.fr

Kamel Smaïli
SMarT Group, LORIA, F-54600, France
kamel.smaili@loria.fr

ABSTRACT

The evaluation of machine virtuosity is critical to improving the quality of virtual instruments, and may also help predict future impact. In this contribution, we evaluate and predict the virtuosity of a statistical machine translation model that provides an automatic responsive accompaniment to mawwal, a genre of Arab vocal improvisation. As an objective evaluation used in natural language processing (BLEU score) did not adequately assess the model's output, we focused on subjective evaluation. First, we culturally locate virtuosity within the particular Arab context of tarab, or modal ecstasy. We then analyze listening test evaluations, which suggest that the corpus size needs to increase to 18K for machine and human accompaniment to be comparable. We also posit that the relationship between quality and inter-evaluator disagreement follows a higher order polynomial function. Finally, we gather suggestions from a musician in a user experience study for improving machine-induced tarab. We were able to infer that the machine's lack of integration into tarab may be due, in part, to its dependence on a tri-gram language model, and instead suggest using a four- or five-gram model. In the conclusion, we note the limitations of language models for music translation.

1. INTRODUCTION

This paper examines the evaluations of, and predictions for, a machine translation model for automated responsive accompaniment to Arab musical improvisation. In this study, we focus on subjective evaluations of the model because, as we will demonstrate, objective measures (such as the BLEU score) are insufficient for this particular case study. We still acknowledge the need for estimating quantitative measurements of quality in order to develop an artificially intelligent model, and offer suggestions for improvement in the conclusion. We further mitigate the shortcomings of this approach by incorporating a qualitative user experience study.

We specifically focus on *mawwal*, a vocal improvisatory genre that is non-metrical, based on a particular maqam, and with a fixed poetic text. In this form, instrumentalists accompany the vocalist by either providing a drone or following the singer's melodic path[1]. Once the singer completes a phrase or sentence, the instrumentalists recapitulate the singer's melody, and this response is referred to as "tarjama," or literally "translation." This project focuses on developing a machine translation model for producing an automated "tarjama," or automatic translation or accompaniment, to Arab vocal improvisation.

Given the call-and-response format of *mawwal* performance, our research requires a parallel corpus of vocal sentences and their corresponding instrumental responses. We have had to build our own corpus due to the lack of available transcriptions of accompanied Arab improvisations. Technical reasons have further prevented us from transcribing commercially produced recordings with high accuracy[2]. When building this corpus, we recorded many hours of vocal and instrumental sentences. We transcribed vocal sentences automatically using a transcriber designed for this vocal style [3], however instrumental sentences were originally recorded in MIDI and did not need this transcription. Given the small size of our corpus, we built a statistical machine translation (SMT) model rather than neural machine translation, in line with work on under-resourced natural languages[4]. In analyzing this data, we considered each vocal idea, regardless of length, as a distinct musical sentence. Each instrumental response is similarly considered one instrumental sentence. Within these sentences, our model quantifies musical notes by their pitch, represented by scale degree, and their quantized duration. The corpus then trains the model to transform vocal sentences into instrumental sentences, imitating the style of the oud, or Arab lute [2].

Subjective assessments of the model's output have been key to guiding the development of this project. This paper analyzes listening tests and an interview in order to evaluate plans to expand the corpus, and to better understand the model's role in, and broader potential for, music performance. The paper is organized as follows: we begin with a discussion of the specificity of Arab musical aesthetics in order to make explicit the terms of subjective evaluation and their cultural context. Specifically, we

discuss *tarab* (modal ecstasy), its relationship to virtuosity, and its significance within Arab music performance. We then briefly examine an objective evaluation, followed by a discussion of two different approaches to subjective evaluation. First, we examine data collected from surveys of expert evaluators who assessed the model's output over three expansions of the corpus. This data has allowed us to formulate equations that predict the subjective evaluation of quality in relation to both the corpus size and the degree of disagreement among evaluators. In situating "quality" within culturally specific understandings of virtuosity, we understand this equation as a method to predict the model's ability to act as "virtuoso." Second, we analyze an in-depth user experience interview that illuminates how the translation model is subjectively experienced within Arab music performance contexts that privilege *tarab*. We consider how this translation model, as well as technological devices more broadly, shape interaction as well as perceptions of success within Arab music settings.

2. VIRTUOSITY AND *TARAB* IN ARAB MUSIC PERFORMANCE

Virtuosity is often understood as the display of high technical skill, particularly in an art form [5]. We culturally locate "virtuosity," however, within a particular Arab context. While there is no direct translation of "virtuoso" or "virtuosity" in Arabic, performers in classical Arab styles are often appraised for culturally specific demonstrations of skill. Vocal music, for example, is often assessed based on a singer's deftness in improvisation; mastery of maqamat (melodic modes) and their paths and modulations; pronunciation; attention to the relationship between text and melody; complexity and appropriateness of vocal ornamentations; and physical production of sound [6, 7]. Instrumental music is often judged based on similar criteria, such as knowledge of maqamat and use of ornamentation. The efficacy of these skills are typically assessed, however, by their ability to elicit ecstasy, or *tarab*, among listeners. Some schools of Arab instrumental performance privilege technical dexterity, such as faster melodies and modulations. *Tarab*, however, refers to both a form of emotional response and a set of aesthetics (emphasizing slower tempi, repetition, and so forth) that was dominant in Arab-majority societies for much of the 20th century [8]. In this framework, the aesthetic parameters of musical performance are deeply connected to emotions emerging from layers of interaction.

We conceptualize virtuosity within the context of *tarab*. Conceptions of "virtuosity" that privilege technical dexterity in line with Western European aesthetics do not share the historical and social significance that *tarab* carries for the musicians and listeners who work on this project, interact with this model, and listen to its performances.

In the *Tarab* framework, the aesthetic parameters of musical performance are deeply connected to emotions emerging from layers of interaction between singer, instrumentalists, and audience, which A. J. Racy terms an "aesthetic feed-back model" [9, 8]. In this structured interaction, instrumentalists interact with the vocalist, which cultivates modal ecstasy among musicians (referred to as saltanah). The audience provides feedback through clapping, sighing, and calling out encouraging words, which all indicate a state of *tarab*. These behaviors then influence how the performers experience ecstasy and inform their decisions in music performance. Situating the aesthetic goals and parameters of *mawwal* within *tarab* allows us, then, to explore the culturally contingent complexities of human-computer interaction (HCI) and, specifically, the role of a computer interactant in an Arab music setting.

3. OBJECTIVE EVALUATION AND THE BLEU SCORE

In this study, we attempted to assess the progress and accuracy of the translation model by using both objective and subjective evaluations. We at first used an objective measure originally designed to calculate the accuracy of natural language translation, known as the BLEU score [10]. We trained a Statistical Machine Translation model with a tri-gram language model, using 5% of the corpora for validation and another 5% for testing. Using the same model's configuration, we trained two other models to return to the previous expansions of the corpus, such that we had three models running on 2.8K, 4K, and 7K parallel corpora. The BLEU scores for these models are presented in the table below. The table indicates that, from moving from the 2.8K to the 4K corpus, the BLEU score decreased from 18.81 to 17.70, indicating a reduction in quality. This may be due to the fact that we introduced new musicians into the corpus, and each musician may have a different performance style, and thus perform different patterns of successive musical notes, even when performing the same musical form (mawwal). When the corpus expanded from 4K to 7K, however, the BLEU score increased from 17.70 to 22.12, indicating a considerable rise in quality. It is worthy of mention that no new musicians were introduced in the last expansion.

Corpus size	2.8k	4k	7k
BLEU	18.8	17.7	22.1

Table 1. Objective quality of computer translations measured along corpus expansion.

As seen in the table above, the BLEU score of this model fluctuated over a series of corpus expansions, however we felt the quality of machine music translations steadily improved across corpus expansions. We therefore found it useful to apply subjective listening tests as an alternative measure, providing more robust results for musical translation that are not limited to the metrics for quality used in natural language processing.

4. SUBJECTIVE EVALUATIONS: TOWARDS A FORMULA FOR PREDICTING QUALITY

In order to gather subjective assessments of the translation model, we used the Mawaweel computer application to create listening tests and further examined its functionality in a user experience interview. The Mawaweel application was built in Java, using JSyn API for audio synthesis [11] and JMSL API [12] for algorithmic composition and performance. In order to carry out these evaluations, we replaced Mawaweel's knowledge-based accompaniment model with the statistical machine translation model. The modified application uses the SMT engine Moses [13] in both training and translation. It also retains the original

Mawaweel transcriber that was designed to transcribe the basic melody of a vocal signal while also applying musical rules specific to the maqam used (Al-Ghawanmeh 2012).

In order to create the listening tests, we fed vocal sentences into the model and exported them individually with their instrumental translation in one audio file. The audio files were listened to one at a time so that evaluators could input their assessment immediately after listening. In our user experience interview, which we discuss later, the musician experimented with the application directly through its graphical user interface. See figure 1.

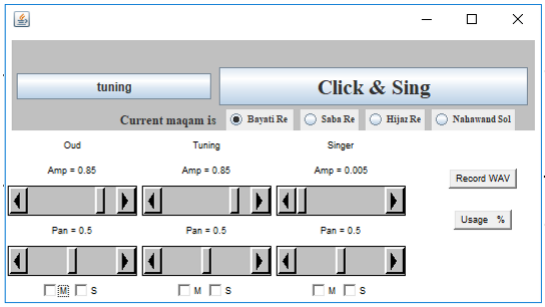

Figure 1. Interface of *Mawaweel* application. See www.mawaweel.com

We then asked three professional practitioners of classical Arab music to complete extensive listening tests. We conducted listening tests four times: once for human-performed translations, and three times for computer-generated translations that drew from corpus sizes available at the time of the tests: 2.8K, 4K, and 7K. In each test, the evaluators listened to randomly selected parallel sentences: fifty of these sentences were human-performed instrumental translations and fifty were computer-generated translations. The evaluators rated each translation from 1 (very poor) to 5 (excellent). We asked evaluators to consider pitch and rhythm, ignoring dynamics, tempo, register, and timbre as these qualities are not considered in the study. We then calculated for each corpus: 1) the average score of each evaluator across all sentences, 2) the mean of the averages of the three evaluators, and 3) the range of averages (highest-lowest). We examined the range of averages in order to compare the degree to which our evaluators disagree with one another when evaluating human accompaniment versus machine accompaniment. Table 2 presents the means of averages (M. AVGs.) and the range of averages (RNG. AVGs) for each of the four sets of tests. The following subsections will provide further elaborations.

	Human	CMP 2.8k	CMP 4k	CMP 7k
M. AVGs.	4.03	2.56	3.13	3.29
RNG. AVGs.	3.91-4.17	2.21-2.89	2.64-3.74	2.85-3.85

Table 2. Subjective quality of human translations as compared to computer translations measured before and after corpus expansion.

The idea of quantifying quality is challenging and possibly controversial, especially within the context of tarab. Proposing an artificially intelligent model, however, requires such quantification in order to better understand the performance of the proposed logic and algorithms. Such quantification also helps us, in this particular study, to predict the required resources (here corpus size) for reaching satisfactory model performance.

4.1 Perceived Quality vs Corpus Size

Figure 2 demonstrates the relationship between corpus size (x axis) and the subjectively perceived quality of computer translation (y axis). The figure presents this preliminary equation that represent the relationship between corpus size and subjective results while assuming a logarithmic pattern of such a curve:

$$subjective\_quality = -3.2 + 1.7 \times \log_{10} corpus\_size \quad (1)$$

This equation will help determine how much the corpus will need to increase in order to obtain automatic accompaniment comparable to human accompaniment. The three blue points in the figure are the means of averages presented in the Table 2. The equation's coefficient of determination (R^2) equals 0.838, indicating a high confidence in predictions based on this formula. Accordingly, the equation suggests that the corpus size would need to increase to 18K in order for human and machine accompaniment to be comparable. We are mindful, however, of the fact that evaluations are a limited measure, particularly as our evaluators only considered pitch and rhythm.

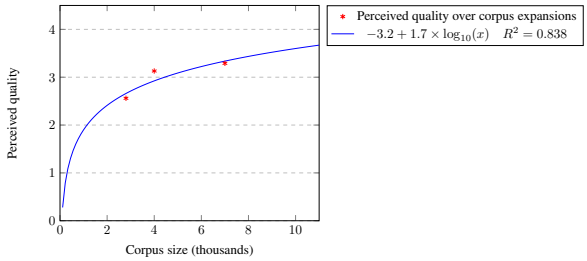

Figure 2. Subjective quality vs. corpus size.

4.2 Perceived Quality vs. Sentence Length

In the following graph (Figure 3) we compare the subjective quality (represented as a mean of averages) and the length of vocal sentences. The blue triangles represent human translated sentences, and the red circles represent computer translated sentences using the most recent translation model, based on a corpus size of 7K parallel sentences. It is interesting to note that the relationship between length and quality for human translation appears to be distinctly different from (if not completely the opposite to) computer translations. For human translations, longer sentences are more likely to receive better assessments among our evaluators. As shown in this figure, 4 out of 5 blue triangles sit on a line with a positive slope. The case is the opposite for computer translation: 4 out of 5 red circles sit on a line with a negative slope, indicating a linear decrease in subjective evaluation with the increase of sentence length. In both cases, the point outside the line represents medium-length sentences, representing an exaggeration of the same trend: the quality increases with length in human translation (above the line) and decreases with length in computer translation (below the line.)

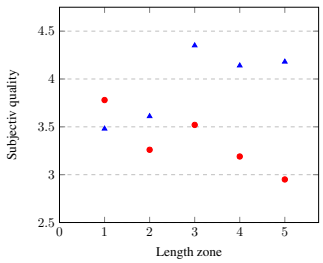

Figure 3. Subjective quality vs the length of vocal sentences. Red circles for computer translated sentences (7k corpus) and blue triangles for human translated sentences. Length zones 1 to 5 are respectively in seconds: <= 2.5, (2.5-5.0], (5.0-7.5], (7.5-10], >10.

We suggest that our evaluators considered longer vocal sentences more melodically complete, leading to a deeper sense of *tarab* and, therefore, more mature instrumental responses. This specific aesthetic framework can at least partly account for the increase in subjective quality evaluations for longer human translations. As for computer translations, translation quality decreases with longer sentences, likely due to the fact that the model is built upon a relatively small corpus. The quality is therefore lower for longer sentences, which are more likely to have a more complex melodic path in any given maqam.

It is also worth highlighting that machine translation actually outperformed human translation for very short sentences. We suggest that very short sentences are often used within a quick, highly interactive call-and-response exchange, where melodic content can be highly variable. When short, human-translated sentences are removed from this performance context, evaluators may perceive human translations as disjointed and departing from the aesthetic expectations that inform longer sentence translation. In other words, evaluators may expect shorter sentence translations to behave like longer sentence translations, however the particular performance context of human translations may alter performer and audience expectations for shorter sentences. Our model design does not consider previous sentences or performance context as a human performer would, and instead systematically treats shorter sentences similarly to longer sentences. This likely accounts, at least partially, for why evaluators were more convinced with computer accompaniment than human accompaniment for very short sentences.

4.3 Perceived Quality vs Evaluators' Disagreement

The next figure (Figure 4) depicts the relationship between the subjectively perceived quality of SMT, represented by means of averages (x axis), and the inter-evaluator disagreement, represented by the range of averages (y axis). We consider data from different stages of research that utilized three different corpus sizes, particularly 2.8K (red circles), 4K (yellow squares), and 7K (green pentagons), and compare this data with evaluations of human-performed translation (blue triangles). Within each group, test sentences are divided into five sub-groups based on duration, as seen previously in Figure 3. In Figure 6, each point in this Cartesian plane is considered valuable and represents the average subjective evaluation of an average of 10 sentences, categorized by length, among three different expert musicians. If we study the pattern of each group of points by color, and then proceed to compare this subsection to the others, it is difficult to find one clear pattern common to all the groups. However, if we consider all these twenty points together as one group, as depicted in Figure 4, we can have a better understanding of the relationship between evaluators' assessments of quality and the degree to which they disagree with one another.

Let us consider the points located in the region within 2.40 to 4.15. Within this intermediate quality range, it is clear that most of the points are located around a parabolic curve where the most disagreement (maxima) is around 3.25, and the least disagreement (minimas) are at the beginning and end of this range.

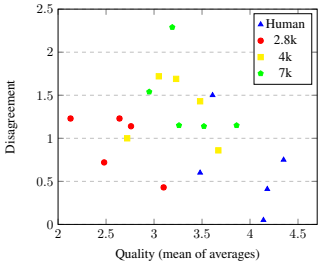

Figure 4. The relationship between subjectively perceived quality and inter-evaluator disagreement. different colors refer to different corpora 2.8K (red circles), 4K (yellow squares), and 7K (green pentagons), and human performed translation (blue triangles). Each point refers to the several sentences located wthin one length zone.

Beyond this range, however, two points (2.13 and 4.35) indicate an increase in disagreement, departing from the parabolic curve discussed above. If we include these two points, they imply an additional maxima before 2.40 and another after 4.15. In an attempt to predict the relationship between quality and disagreement in these extreme regions of the figure, we will add two more theoretical points. First, we add one point indicating an average quality of 0 with no disagreement; in other words, all evaluators agree that the quality is 0. Conversely, we add a second point indicating an average quality of 5 with no disagreement; again, here all evaluators agree that the quality is 5. Figure 5 presents these additional points in the same color as the other collected data.

It is apparent that the overall relationship is of a higher degree of polynomial function. Still, we want to derive an appropriate equation with a high R^2, and will thus need to suggest places for the other two maximas. When the mean of averages is between 4 and 5, non-zero range of disagreement is either 1 (5-4) or 2 (5-3). For the former range, the mean of averages is either 4.67 (with votes of 5, 5, 4) or 4.33 (5, 4, 4). For the latter range, the mean of averages must be 4.33 (with votes of 5, 5, 3). There are then 3 possible points to include on the curve: (4.33, 2), (4.66, 1), or (4.33, 1). When applying the same process on the other side of the graph, between 0 and 2 on the x-axis exclusively, there are 3 more possible points: (1.33, 1), (1.67, 1), or (1.67, 2). We experimentally found that the curve will have the highest R^2 (0.62) when using the points (1.67, 2) and (4.67, 1) as maximas, those points are black-filled in Figure 5.

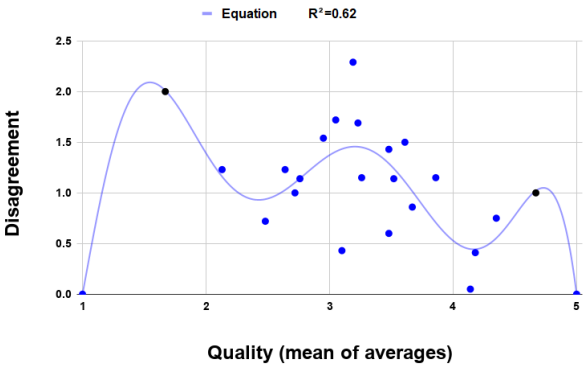

Figure 5. Subjective quality vs. corpus size.

This relatively high R^2 helps us predict the future attitudes of human evaluators. Disagreement between evaluators is expected to rise gradually with the expansion of the corpus, until reaching a maxima when the average quality is 4.67. Beyond this, however, disagreement dramatically falls towards zero as quality scores approach the maximum (5). This prediction is important when considering future plans for expanding this project's corpus, which can be quite costly. Equation 1 suggests that it would be necessary to expand the corpus to 41K in order to raise quality evaluations above 4.67, however this is a very difficult task. We ultimately face two options: 1) focus on achieving higher quality translations while accepting increasing disagreement among evaluators and the broader music community, or 2) accept lower quality ($\sim$4) as well as disagreement ($\sim$0.5), expected when the corpus size reaches 17k.

5. EXPERIENCING TECHNOLOGY IN ARAB MUSIC PERFORMANCE

In order to subjectively evaluate this model, we also interviewed Muhannad al Khateeb, a professional vocalist and oudist. Our interview took place in person and over WhatsApp Messenger voice messages in December 2019. Muhannad extensively used the translation model embedded in the Mawaweel application, and performed with the computer in a previously published musical artwork [14]. Of primary concern was whether this model elicits, or has the potential to elicit, *tarab*, the primary goal of classical Arab music performance. Muhannad described *tarab* as emerging from interactions between audience, singer, and instrumentalists, similar to A. J. Racy's model of *tarab* as an "ecstatic feedback model" [9, 8].

Muhannad's framework for assessing audience engagement was markedly different, however, from the shouts, gestures, and gasps that scholars typically note as indicative of *tarab* [7, 9, 8, 15]. When describing the ways he typically interacts with audiences, Muhannad pointed to smartphones as significant to interpreting audience reactions and feedback:

> The more mobile phones are recording, the better the work is. Recording means they're documenting something nice حِلو *ḥiluw*. If they're not interested in documenting [it], it means it's not nice مِش حِلو *miš ḥiluw*. Although it's a contradiction because I want them to listen, but still, it's nice.

Here Muhannad expresses disappointment that audiences do not "listen" while using their smartphones to document performance. In this discussion he emphasized, however, that smartphones represent audience satisfaction with his performance. This suggests that contemporary technologies, and smartphones in particular, have already come to reconfigure the performance and experience of *tarab*. While they may, in Muhannad's view, disrupt traditional *tarab* interaction by distracting audiences and preventing attentive listening, smartphones also occupy a central interactive role by mediating and representing audience satisfaction. They partly replace, or perhaps complement, other demonstrations of attention and appreciation typical to *tarab* (e.g. calls, eye contact).

Muhannad further emphasized the significance of concentration and listening when describing his interactions with the machine translation model. He drew on the Arabic proverb "one with two minds is a liar" صَاحِب بَالَين كَذَّاب *ṣāḥib bālayn kaḏḏāb* to illustrate the difficulty of singing and accompanying himself on oud at the same time. He explained that he prefers to dedicate himself to singing so that he does not perform with divided مَقسُوم *maqsūm* attention, however he will at times accompany himself on oud at gigs in order to receive a higher payment. The translation model, in his view, allowed him to concentrate more fully on his own performance, even bypassing the need to interact with other musicians. This lack of interaction, however, also prevented his cultivation of modal ecstasy سَلطَنَة *salṭanah*. Muhannad suggested that the model lacked the spirit رُوح *rūḥ* and expression of feelings إِحسَاس *ʾeḥsās* necessary for him to experience ecstasy. He did concede, however, that he at times loses ecstasy when performing with human musicians when they, too, do not correctly demonstrate feeling and attention [16].

Muhannad's concern regarding listening and attention is implicated within broader generational anxieties regarding attentive listening practices once revered in Arab-majority societies (e.g. [15]). The role of technology, and its potential to disrupt attention, is a particular source of discomfort for practitioners and audiences of Arab music. Throughout the course of pursuing this research, we have found that practitioners of Arab music often react negatively to the idea of a computer producing Arab music. Certainly, as [17] discuss, incorporating music technology within traditional music settings requires a careful consideration of notions of cultural ownership. Muhannad's discussion of smartphones, however, demonstrates how Arab musical performance should not be viewed as untouched by technology. Following Muhannad's lead, we view this model as working within broader histories of the adaptation of sound recording, re/production, and transmission technologies within Arab-majority societies [18, 19, 20, 21].

Throughout our interview, Muhannad's assessment shifted from centering emotion (spirit and feeling) towards a discussion of musical structure, outlining some of the musical elements that encourage modal ecstasy. While Muhannad initially said he was unable to experience ecstasy when per-

forming with a computer interactant, he continued to offer specific suggestions to improve his experience. First, he requested a drone for the improvisatory singing section, which would imitate standard Arab music practice and provide a stable reference pitch. Second, he recommended offering different regional accompaniment styles, such as Egyptian and Iraqi, as options. He argued that every *mawwal* has a spirit رُوح *rūḥ*, similar to the "spirit of the law" رُوح القَانُون *rūḥ ālqānūn* that must also be intuited, and that attending to regional style or color لَون *lawn* would help personalize or individualize each *mawwal* performance. Third, he recommended providing prepared, preset musical accompaniment for popular songs and canonical pathways of improvisation. When implemented, these suggestions will provide additional musical and cultural specificity to both the translation model and the Mawaweel application. Muhannad's suggestions may help the model further integrate into, or reconfigure, the *tarab* feedback model discussed here. There remain, however, a number of limitations. The use of this model in Arab music performance may have drawbacks similar to what Muhannad described regarding the audience's use of smartphones. The machine may indeed constitute another form of distraction for audiences, however Muhannad noted that he was not distracted by the model's accompaniment as a performer. The model may also challenge musicians and audiences to fundamentally reconsider what, exactly, constitutes spirit روح *rwḥ* and feelings إِحسَاس *ā'ḥsās* in performance. This point questions where spirit and feelings can be found, whether in musical content, performance atmosphere جَو *ğaw*, or the presence and performance of a human body. Indeed, dividing musical "content" from "context" is problematic given its reification of music as a product removed from social, cultural, and embodied contexts, whereas approaching music as "process" may more fully capture the complexities of musical performance. It remains to be seen whether machine participation can imitate, produce or reconfigure "spirit" and "feelings," and whether audiences and musicians will allow for machine participation to modify and extend the *tarab* feedback model.

Muhannad's broader discussion also provides valuable insights from a technology perspective. First, he offers an account of perceived quality that differs from the subjective listening tests, as these tests reported a higher level of overall satisfaction. This may be accounted for by the fact that subjective listening evaluations were made only on separate pairs of sentences (vocal and instrumental), listened to in isolation rather than within the context of a full improvisation. The user experience study therefore confirms that more work needs to be done for the machine to better respond to and build upon each vocal sentence within the context of a full improvisation, and take into account the previously performed sentences.

Another key issue in the proposed model, as discussed previously, was the consistent lower quality of longer machine-translated instrumental sentences. This issue may be due to the model's limited ability to analyze the detailed melodic path سَير *sayr* of each instrumental sentence. We used a tri-gram language model, where only 3 adjacent notes are considered at a time. Compare this to a natural language context, where the translation of a word is more nuanced given the context of two other words. While this may be sufficient for natural languages to understand and model words in context, the melodic context requires a length longer than three, perhaps requiring even five or six notes. While this issue only directly addresses musical content, using higher-order n-grams will be an important step to improve machine performance and virtuosity in Arab music settings.

6. CONCLUSIONS

Throughout this discussion, we have noted the limitations of language models for machine music translation. The BLEU measure, designed to calculate the accuracy of natural language translation, was inconsistent in assessing our music translation model over the course of a series of corpus expansions. The results from our evaluators' listening tests, as well as our own assessments of the model, contradicted the measurements of the BLEU score. The BLEU score may, however, remain a useful indicator when testing corpora with larger differences in size. Furthermore, the use of the tri-gram language model, common in natural language processing, is limited in its ability to understand melodic context. This demonstrates the importance of distinguishing the needs of music translation from language translation for developing more precise approaches to music translation. By focusing on subjective evaluations, we were also able to predict translation quality across corpus expansions, allowing us to estimate 18K as an ideal corpus size. Furthermore, our discussion highlighted the importance of examining cultural and social context when developing and implementing music translation models. We considered machine virtuosity within the context of *tarab*, interaction, and attention, which allowed us to further attend to performer and audience expectations in a specifically Arab music context. This holistic and interdisciplinary approach afforded new, and at times unexpected, insights into this ongoing research.

Acknowledgments

The authors acknowledge financial support of this work, part of TRAM (Translating Arabic Music) project, by the Agence universitaire de la Francophonie and the Deanship of Scientific Research in the University of Jordan.

7. REFERENCES

[1] A. Hamam, *Al-Hayah al-musiqiyah fi al-urdun. (Musical Life in Jordan)*. Ministry of Culture, 2008.

[2] F. Al-Ghawanmeh and K. Smaïli, "Statistical Machine Translation from Arab Vocal Improvisation to Instrumental Melodic Accompaniment," *Journal of International Science and General Applications*, vol. 1, no. 1, pp. 11–17, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01873788

[3] F. Al-Ghawanmeh, "Automatic Accompaniment to Arab Vocal Improvisation "Mawwāl"," Master's thesis, New York University, 2012.

[4] M. A. Menacer, D. Langlois, O. Mella, D. Fohr, D. Jouvet, and K. Smaïli, "Is statistical machine translation approach dead?" 2017.

[5] O. Jander, *Virtuoso*. Grove Music Online, 2001, ed. Deane Root.

[6] J. Farraj and S. A. Shumays, *Inside Arabic Music: Arabic Maqam Performance and Theory in the 20th Century*. Oxford University Press, 2019.

[7] V. Danielson, *The Voice of Egypt: Umm Kulthum, Arabic Song, and Egyptian Society in the Twentieth Century*. University of Chicago Press, 1997.

[8] A. J. Racy, "Making Music in the Arab World the Culture and Artistry of Tarab," 2003.

[9] ——, "Creativity and ambience: an ecstatic feedback model from Arab music," *The World of Music*, vol. 33, no. 3, pp. 7–28, 1991.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[11] P. Burk, "JSyn-A Real-time Synthesis API for Java." in *ICMC*, 1998.

[12] N. Didkovsky and P. Burk, "Java Music Specification Language, an introduction and Overview," in *ICMC*, LaHaban, Cuba, 2001.

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.

[14] F. Al-Ghawanmeh and K. Smaïli, "Andalusian Fragrance," http://www.aiartonline.com/music/fadi-al-ghawanmeh-and-kamel-smaili/, 2018, accessed: 2020-01-30.

[15] J. H. Shannon, *Among the jasmine trees: Music and modernity in contemporary Syria*. Wesleyan University Press, 2006.

[16] R. Banerji, "Phenomenologies of Egalitarianism in Free Improvisation: A Virtual Performer Meets its Critics," Ph.D. dissertation, UC Berkeley, 2018.

[17] B. L. Sturm, O. Ben-Tal, Ú. Monaghan, N. Collins, D. Herremans, E. Chew, G. Hadjeres, E. Deruty, and F. Pachet, "Machine learning research that matters for music creation: A case study," *Journal of New Music Research*, vol. 48, no. 1, pp. 36–55, 2019.

[18] M. A. Frishkopf, *Music and media in the Arab world*. American Univ in Cairo Press, 2010, no. 4108-4109.

[19] C. Hirschkind, *The ethical soundscape: Cassette sermons and Islamic counterpublics*. Columbia University Press, 2006.

[20] A. K. Rasmussen, "Theory and practice at the 'Arabic org': digital technology in contemporary Arab music performance," *Popular Music*, vol. 15, no. 3, pp. 345–365, 1996.

[21] M. Stokes, *The republic of love: Cultural intimacy in Turkish popular music*. University of Chicago Press, 2010.