How can machine translation help generate Arab melodic improvisation?

Fadi Al-Ghawanmeh, Alexander Refsum Jensenius, Kamel Smaïli

How can machine translation help generate Arab melodic improvisation?

Fadi Al-Ghawanmeh
Music Dept., Univ. of Jordan
SMarT Group, LORIA, F-54600
RITMO Centre, Univ. of Oslo
fadi.al-ghawanmeh@loria.fr

Alexander Refsum Jensenius
RITMO Centre
Department of Musicology
University of Oslo, Norway
a.r.jensenius@imv.uio.no

Kamel Smaili
SMarT Group
LORIA, F-54600
University of Lorraine, France
smaili@loria.fr

Abstract

This article presents a system to generate Arab music improvisation using machine translation (MT). To reach this goal, we developed a MT model to translate a vocal improvisation into an automatic instrumental oud (Arab lute) response. Given the melodic and non-metric musical form, it was necessary to develop efficient textual representations in order for classical MT models to be as successful as in common NLP applications. We experimented with Statistical and Neural MT to train our parallel corpus (Vocal $\rightarrow$ Instrument) of 6991 sentences. The best model was then used to generate improvisation by iteratively translating the translations of the most common patterns of each maqām (n-grams), producing elaborated variations conditioned to listener feedback. We constructed a dataset of 717 instrumental improvisations to extract their n-grams. Objective evaluation of MT was conducted at two levels: a sentence-level evaluation using the BLEU metric, and a higher level evaluation using musically informed metrics. Objective measures were consistent with one another. Subjective evaluations by experts from the maqām music tradition were promising, and a useful reference for understanding objective results.

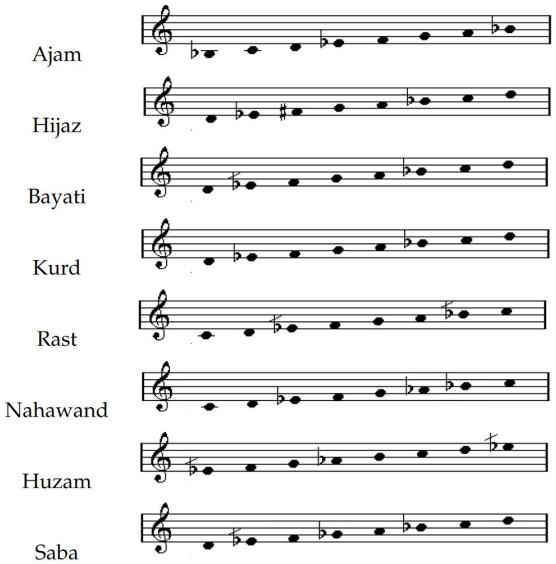


Figure 1: The most common maqāmāts in Arab Music (Al-Abbas, 1986).

1 Introduction

The purpose of this paper is to present a method for using machine translation (MT) to generate automatic instrumental improvisation in Arab maqām music, particularly in two contexts: responsive accompaniment to vocal improvisation (mawwāl), and free instrumental improvisation (taqāsīm). This method could then be adapted to other melodic musical traditions. We situate this project within the efforts to maintain, preserve, and develop these musical forms in Arab music using (MT) paradigms. We construct our own corpora and tools for their collection and processing, explore neural and statistical methods, and test them using the bilingual evaluation understudy (BLEU) measure, which is a common MT metric (Papineni et al., 2002). The study presents the results of using a BLEU score along with musically informed

metrics (Yang and Lerch, 2020) and subjective evaluations. More broadly, we view this music-based project as a MT challenge within the broader context of under-resourced languages (Krauwer, 2003) (Berment, 2004).

In Arab music, mawwāl is a non-metric vocal improvisation and is often applied to narrative poetry. Upon the completion of each vocal sentence, the instrumentalist performs a recapitulation, or a translation of that sentence (Racy, 1998) (Farraj, 2007). In other words, the duo (singer, instrumentalist) play a musical conversation. At time t_i, the singer produces an improvisatory phrase $p(t_i)$, then the instrumentalist produces a musical answer corresponding to $p(t_i)$, which we will call $a(t_{i+1})$. When the instrumentalist finishes, the singer responds with a new improvisatory sentence $p(t_{i+2})$, and this process is repeated until the end of the improvisation. In this paper, we first review our approach for using MT to propose an instrumental responsive accompaniment to mawwāl. We then explain how to use the same model to generate a full instrumental improvisation in the maqām context using iterative translation. We particularily aim to answer the following questions:

- How can reducing the dimensions in the symbolic (textual) representation of a small parallel (vocal and instrumental) dataset help training statistical and neural MT models?
- How can a vocal-to-instrumental MT model serve as a basis to generate real-time instrumental improvisation conditioned to listener feedback?
- What is the significance of an objective measure for the evaluation of MT (BLEU) in this particular application, especially in relation to musically informed objective metrics and expert subjective evaluation?

2 Background

2.1 Introducing the maqam

Arab music is based on the concept of maqām. It is a system of scales, melodic patterns, modulation possibilities, ornamental standards as well as aesthetic conventions that together form a rich melodic framework and artistic tradition. Maqamat (plural of maqām) are organized by principles that establish common patterns, developments, and relationships between the different maqamat. The most related counterpart in Western music is the mode (Boulos, 2021). Each maqām is based on a scale; figure 1 illustrates the most important maqāmāt. The first note in the ascending stepwise scale is the first scale degree, the second note is the second scale degree, etc.

Traditional Arabic music compositions and improvisations are based on the maqām system. Improvisations are non-metric forms and can be performed in vocal music as well as instrumental music. These are called mawāwīl (plural of mawwāl) in the former case, and taqāsīm (pl. of taqsimah) in the latter. The mawwāl exhibits the vocalist's virtuosity when singing narrative poetry, and taqāsīm demonstrates the instrumentalist's virtuosity and the instrument's beauty and capabilities. Both forms are tightly connected to a sense of modal ecstasy (Racy, 2004). In practice and before the start of the mawwāl, an instrumentalist may set the stage for the singer by performing a taqsīmah on the same maqām.

2.2 Related work

There are several recent contributions to generating musical compositions and accompaniments in Western music. In (Rao and Lau, 2018), hidden Markov models were used to follow the musical score in expressive performance, and also to play and possibly adjust the chordal accompaniment based on the soloist's interpretation of the score. Similarly, (Mo, 2022) used these models for piano accompaniment, and (Asesh, 2022) utilized them in order to reproduce and synthesize both monophonic and polyphonic music selected from vintage 8-bit video games. Finite state transducers were used in (Forsyth, 2016) to develop a data-driven method for automatic harmonic accompaniments to melodies.

In (Ren et al., 2020), an accompaniment model was built for pop music with an encoder-decoder. In so doing, they encoded multi-track MIDI events from each musical measure into one larger sequence. In order to capture long-term dependencies, a transformer was used as a backbone for both the encoder and decoder. The model was trained on MIDI datasets with sizes that ranged widely, from 5k to 21k musical pieces, where each dataset included tens of thousands of measures (bars). Using transformer-based NMT in (Kalonaris et al., 2020), a model to generate contrapuntal musical accompaniment was developed based on a total

dataset of 17K+ four-bar parallel sequences. For testing, they conducted both objective and subjective evaluations and reported that the objective BLEU score (Papineni et al., 2002) – typically used to evaluate MT of natural languages – corresponded with human subjective evaluation.

Early contributions towards automating the instrumental musical accompaniment started in the mid-1980s (Dannenberg, 1984) (Vercoe, 1984). However, researching automatic accompaniment in the context of Arab music is relatively recent (Al-Ghawanmeh, 2012). To our knowledge, this is the only project focused on generating instrumental improvisation in the Arab maqam idiom using machine learning. As asserted in (Magnusson, 2021), the musical ideas of a given place are imbricated with its music technologies. We thus understand our work as contributing to broader efforts to maintain, preserve, and develop the music practices of the Arab world in an AI driven era.

3 Machine translation of mawwal

Data structuring and representation are key in this application in order for statistical and neural MT models to be as satisfying as in NLP applications. We therefore begin this section by describing these details before discussing the details and evaluation of the MT models.

3.1 Dataset

For MT, we need a parallel corpus for training and fine-tuning the models. In our case, each sentence improvisation should be presented as follows: $p(t_i)$, $a(t_{i+1})$, $p(t_{i+2})$, $a(t_{i+3})$, ... $p(t_{i+n})$, $a(t_{n+1})$. From this structure, we should build a parallel corpus that will respect the format given in Table 1.

Source	Target
$p(t_i), \dots p(t_{i+n})$	$a(t_{i+1}), \dots a(t_{i+n+1})$

Table 1: The format of the parallel mawwāl corpus.

This kind of corpus does not exist, so we created it. In order to do so, we gathered own singers, MIDI keyboard instrumentalists, and equipped recording rooms. Indeed, the MIDI keyboard can emulate Arab instruments to a sufficient degree, and many singers today are accompanied by electronic keyboards rather than acoustic instruments. The singer sings a sentence $p(t_i)$ and the instrumentalist produces an oud answer ($a(t_{i+1})$). This protocol standardized the recording process and circumvented the need to transcribe existing mawāwīl, a consuming task. Vocal signals were transcribed automatically using a transcriber that was developed and tested for the mawwāl (Al-Ghawanmeh, 2012), allowing similar adjacent notes to merge for the better presentation of melodic patterns (Al-Ghawanmeh and Smaïli, 2018).

3.2 Data representation

The vocal sentence and the instrumental response are represented by scale degree and duration as in Table 2. The scale degree is represented by the letter s and the duration by t. The scale degrees of the vocal sentence, respectively, are: 7^{th} degree (octave lower) and 1^{st} degree. The instrumental response is a descending four-note motive. The scale degrees of this instrumental response are respectively: 3^{rd}, 2^{nd}, 1^{st} and 1^{st} degree (octave lower). The notation t_7 means that the duration of the previous note is of rank 7 on a scale of 8.

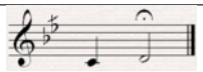
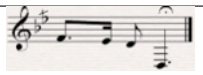
Type	Musical score	Text representation
V.		$s_7t_7s_1t_8$
I.		$s_3t_6s_2t_3s_1t_5s_1t_8$

Table 2: Data representation for MT. "V"is for vocal and "I" for instrumental sentences.

We recorded a corpus of 6991 parallel sentences whose statistics are given in Table 3. Instrumental sentences are usually longer than vocal sentences due to the acoustic features of plucking instruments. The dataset is available for use for research purposes. [1]

	Vocal	Instr
Sentence count	6991	6991
Duration	12.46h	10.96h
Note count (NC)	88947	176279
Average NC per sentence	12.75	25.27
σ of NC per sentence	10.53	20.60
Sentences within 1 octave	91.12%	45.27%

Table 3: Statistics on the parallel corpus.

Since our corpus is small in comparison to corpora for MT between natural languages, the

[1]The dataset is available via this link: https://github.com/FadiGhawanmeh/AMICOR

amount of vocabulary should be small in order to have a good coverage of the melodic sentences. In fact, the pitch range of both the vocal improvisation and the instrumental accompaniment can exceed two octaves. If we decide to use pitches as letters in our corpus, the total count of letters can exceed 48 (24 pitches per octave with a minimum interval of $\frac{1}{4}$). When using pitch-class representation, which equates octaves, the total count of letters does not exceed 24 pitches. This number remains high relative to the small size of the corpus. Given this issue, and the complication of incorporating different maqāmāt in varying keys, we decided to use scale degree representation. Arab maqāmāt are often based on seven scale degrees, allowing us to have the total number of letters as low as seven. Consequently, in our MT, we use a vocabulary of 15 different words ($s_1 \dots s_7, t_1 \dots t_8$).

3.3 Statistical and neural MT

For Statistical MT (SMT), we utilized the 2017 stable release of the Moses engine (Koehn et al., 2007) in order to train our models. This process utilized conventional phrase-based modeling, with bidirectional lexical and phrase translation probabilities, a word and phrase penalty, a distortion model, and a 3-gram language model with smoothing (witten-bell).

For neural MT (NMT), we trained our models with the OpenNMT system (Klein et al., 2020). We utilized sequence-to-sequence modeling (Sutskever et al., 2014). We obtained the best NMT results with the following configuration: one embedding layer, two bidirectional RNN (precisely: LSTM) encoder layers, two RNN decoder layers, and a softmax layer, with an RNN size of 512. It is worth noting that we also experimented with the transformer (Vaswani et al., 2017) as a potential substitute to RNN, however the results did not outperform the RNN. While transformers are typically used with larger corpora, ours is small, diverse, and accounts for few dimensions. In particular, the data only incorporates scale-degree and quantized duration, with a 15-word vocabulary and an average sentence length of 12.75 in the source sequence and 25.27 words in the target sequence.

In developing the models, we used 90% of the dataset for training, 5% for validation, and 5% for testing. As our dataset is small, we also applied cross-validation. In NMT, we applied data augmentation using transcriptions of time-stretched copies of the dataset. We used the BLEU measure (Papineni et al., 2002) as an objective method to compare, at the sentence level, the generated translation to the human translation. The BLEU scores for SMT and NMT are given in Table 4. The results for SMT, NMT (LSTM), and NMT (transformer) were 22.12, 18.29 and 12.6, respectively.

MT model	SMT	NMT (LSTM)	NMT (Transformer)
BLEU	22.12	18.29	12.60

Table 4: BLEU Results of MT.

To compare our SMT and best NMT models to human improvisation beyond the sentence level, we present in Table 5 five musically informed objective metrics adapted from (Yang and Lerch, 2020), and provide a generic statistical overview. We calculated the value of each metric for each sentence, then found the average and the standard deviation over the whole test set. Over the five metrics, the average distance between machine and human translation is 8.54% and 14.34% for SMT and NMT (LSTM), respectively.

Metric ($p.s$)	Human		SMT		NMT (LSTM)	
	$\overline{x}$	σ	$\overline{x}$	σ	$\overline{x}$	σ
M1	24.7	20.1	24.0	17.5	14.6	6.8
M2	5.1	1.6	5.1	1.5	4.7	1.5
M3	1.7	0.9	1.7	0.9	1.8	0.9
M4	0.9	0.4	0.8	0.2	0.9	0.4
M5	4.0	1.2	3.2	0.6	4.50	1.2

Table 5: Comparing best MT models to human accompaniment using musically informed objective metrics: note count (M1), scale degree count (M2), scale degree range (M3), mean scale degree interval (M4), and mean quantized-duration (M5). Metrics were calculated per sentence ($p.s$), then the average ($\overline{x}$) and standard deviation (σ) were calculated on all sentences of the test set.

We can conclude that for this particular application and within the presented conditions, SMT provided somewhat better results than NMT. This is probably because our dataset is relatively small, diverse, and incorporating relatively few dimensions or parameters. Many datasets targeting Western styles, such as mainstream pop in (Ren et al., 2020), are much larger and incorporate multiple parameters in order to address polyphony and meter. Our small dataset, however, only addresses

the parameters of pitch and duration. NMT results, however, are expected to outperform SMT with further expansion of the dataset. The two methods can then co-exist because each presents a musically different response as indicated by the above-presented musically informed metrics, as if each response comes from a different musical instrument of different characteristics. The example presented in table 6 also illustrates this different response. In the absence of baseline results in this style, we consider our results as baseline for future research.

Sentence:
$s_1 t_1 s_3 t_1 s_4 t_1 s_3 t_1 s_4 t_8 s_5 t_1 s_4 t_5 s_5 t_1 s_4 t_5 s_6 t_1 s_4 t_6 s_3$ $t_1 s_4 t_2 s_3 t_1 s_4 t_1 s_2 t_2 s_3 t_1 s_4 t_4 s_3 t_1 s_4 t_5 s_6 t_1 s_5 t_1 s_3 t_1$ $s_4 t_1 s_2 t_1 s_3 t_1 s_4 t_1 s_3 t_1 s_2 t_2 s_1 t_8$
Translation SMT:
$s_1 t_2 s_2 t_3 s_3 t_3 s_4 t_3 s_3 t_3 s_4 t_4 s_4 t_3 s_4 t_3 s_4 t_3 s_3 t_3 s_4 t_3 s_3$ $t_3 s_4 t_3 s_6 t_2 s_6 t_2 s_4 t_2 s_3 t_3 s_2 t_3 s_3 t_3 s_2 t_3 s_2 t_4 s_3 t_3 s_4 t_3$ $s_3 t_3 s_4 t_3 s_6 t_3 s_5 t_4 s_4 t_4 s_4 t_3 s_3 t_3 s_2 t_6 s_1 t_6 s_1 t_6 s_1 t_8$
Translation NMT (LSTM):
$s_1 t_2 s_2 t_2 s_3 t_2 s_4 t_3 s_3 t_2 s_3 t_3 s_2 t_1 s_1 t_3 s_5 t_3 s_4 t_2 s_3 t_2 s_2$ $t_3 s_3 t_4 s_4 t_1 s_3 t_2 s_2 t_2 s_1 t_4 s_1 t_2 s_1 t_4 s_3 t_1 s_5 t_2 s_4 t_1 s_5 t_2$ $s_2 t_4 s_3$

Table 6: Example of outputs of the best MT models.

3.4 Subjective evaluation

In natural languages, there are established conventions for determining successful and inadequate BLEU scores, however these standards do not necessarily apply to this application. It is therefore unclear whether or not a BLEU score of 22.12 could be considered good. We therefore applied subjective listening tests as an alternative measure. We asked three professional practitioners of classical Arab music to complete extensive listening tests for human-performed translations and computer-generated translations. 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 them to focus on pitch and rhythm, and to ignore dynamics, tempo, register, and timbre because these qualities were not considered in the study.

As shown in Table 7, we note that even for human responses, the experts were not totally satisfied with the performance of the instrumentalists. This is normal for experts of this music tradition; the automatic responses produced by the SMT, however, received an average score of 3.29, with a minimum of 2.65 and a maximum of 3.85. For further subjective and musicological discussion, we refer to relevant work related to this project that leans more towards humanistic musicological approaches (Al-Ghawanmeh et al., 2021) (Al-Ghawanmeh et al., 2019), focusing primarily on subjective evaluation, speculative discussions regarding the possibilities for machine virtuosity, and the holistic impact of artificially intelligent compositions for musical experience writ large.

	Human	**SMT**
Mean average	4.03	3.29
Range of averages	[3.91, 4,17]	[2.85, 3.85]

Table 7: Subjective evaluation of the human and SMT responsive improvisation

4 Taqasim Generation

Tarab music, to which taqāsīm belongs, emphasizes repetition (Racy, 2004). While repetition may be important for any musical work, it does not necessarily involve exact replication, and can incorporate variations and elaborations (Dai et al., 2022). In this section, we tackle the issue of taqāsīm generation. The main idea of our method starts from the definition of the maqām. As previously noted, the maqām is a set of pitches as well as characteristic melodic motives and formulas of their use (Nettl, 2007). Technically speaking, characteristic melodic motives are the frequently-repeated melodic patterns in a representative sample (corpus) of improvisations.

We thus constructed a representative taqāsīm corpus (C_{mi}) on several maqāmāt. We then extracted the frequently repeated patterns (n-grams) from each C_{mi} and used them afterwards as seeds to create and develop new musical sentences in new improvisations. This was inspired by (Ünal et al., 2014) who used n-grams efficiently within an algorithm for an automatic classification of Turkish maqām from symbolic data.

To construct the taqāsīm corpus, we requested two practitioners to perform improvisations of several lengths on eight main maqāmāt (see Figure 1). We collected 717 improvisations. Statistics concerning this dataset are presented in Table 8.

Musical detail	Value
Total number of improvisations	717
Total duration	22.09h
Total note count	631201
Average note count	880.34
σ of Note Count	690.68

Table 8: Statistics on the taqāsīm instrumental corpus

After construcing the taqāsīm corpus and extracting the frequently repeated n-grams, we then used the MT model presented in Section 4 but for a different task. Instead of translating a vocal sentence into an instrumental response, the new task was to translate a given n-gram into an elaborate variation of itself.

The process of generating music is based on Algorithm 1. Its main idea is to select an n-gram from a maqām's corpus C_{mi}, then in an iterative process we translate it into an elaborated variation of itself. This means we translate the translation to have more elaborated variations.

Algorithm 1 Process of generating sentences in a specific maqām

$S(0) \leftarrow Select(C_{mt}, ngram)$
$i \leftarrow 1$
while $count(s_1, S(i-1)) \leq \alpha$ **do**
$NewSent \leftarrow Trans(S(i-1))$
  if $MotionCapture(NewSent) = 0$ **then**
    $S(i-1) \leftarrow Select(C_{mt}, ngram)$
    $NewSent \leftarrow Trans(S(i-1))$
  else
    $S(i) \leftarrow NewSent$

  end if
  $i \leftarrow i + 1$
end while=0

This algorithm takes into account the user's feedback by analyzing the time series signal produced by a motion-capture tool connected to the headset. Listeners satisfaction in this musical style is obtained by either producing music that meets their expectations or by pleasantly surprising them (Racy, 1998) (Kahel, 2021). Their satisfaction is expressed by a response that generally corresponds to a movement of the body. In response, the musician answers by emphasizing what led to the satisfaction of the listener. Taking this interaction as inspiration, we analyze the motion-capture signal in order to determine whether the movement was actually caused by listening to the automatic generated sentences rather than any other external reason. Consequently, in the algorithm, if the response of the motion-capture is 0, this indicates that no pleasant movement related to music was detected, then we select another n-gram to produce new translation with the wish that this one will produce more effect on the listener. To allow for a smooth melodic development, the new n-gram will typically have some similarity – whether close or loose – to the previous n-gram. The musically-informed objective metrics that we presented earlier form a basic measure for n-gram similarity. Basic domain knowledge is also considered when selecting n-grams because characteristics of musical sentences change along the improvisation (Kisserwan, 2016).

The iterative translation of a given n-gram is repeated until the tone center s_1 dominates the sequence $S(i)$, or in other words when the number of s_1 in $S(i)$ is greater than a fixed threshold α. Table 9 illustrates an example of musical sentences produced by the model proposed in this section.

N.	Sequence	Description
S_1	$s_3t_3s_4t_2s_4t_3s_4t_1$	
S_2	$s_3t_3s_2t_2s_1t_2s_2t_2s_1t_2s_2t_2$ $s_1t_3s_2t_2s_1t_4s_2t_3s_3t_8$	Trans(S_1)
S_3	$s_1t_4s_1t_4s_1t_3s_1t_3s_1t_2s_1t_2$ $s_2t_2s_1t_3s_1t_3s_1t_3s_2t_3$ $s_2t_3s_1t_3$	Trans(S_2)
S_4	$s_1t_7s_1t_7s_1t_7s_1t_5s_1t_4s_1t_3$ $s_1t_2s_1t_2s_2t_2s_1t_3s_1t_3s_1t_3$ $s_1t_3s_1t_3s_1t_3s_1t_3s_1t_3s_1t_3$ $s_2t_3s_1t_3s_6t_2s_1t_3s_1t_3$	Trans(S_3)

Table 9: Musical An example of the iterative translation.

4.1 Evaluation

We performed both objective and subjective evaluations. In Table 10 and using the five musically informed objective metrics, we present a generic statistical overview of machine-generated taqāsīm and a set of ones of comparable length in the dataset. Results are very good for the metrics: scale-degree count and average scale degree interval. There is potential for further improvement in the other measures. In the absence of baseline results for taqāsīm, we consider these results as a baseline for future research.

Metric	Human		SMT	
$(p.s)$	$\overline{x}$	σ	$\overline{x}$	σ
M1	25.1	13.9	17.4	11.0
M2	4.8	1.2	4.5	1.6
M3	1.8	0.9	1.4	0.8
M4	0.7	0.2	0.7	0.3
M5	3.1	0.6	2.9	0.4

Table 10: Comparing iterative translation to human improvisation using musically informed objective metrics: note count (M1), scale degree count (M2), scale degree range (M3), mean scale degree interval (M4), and mean quantized-duration (M5). Metrics were calculated per sentence ($p.s$), then the average ($\overline{x}$) and standard deviation (σ) were calculated on all sentences of the test set.

For the subjective evaluation, we recruited two expert practitioners in the maqām music tradition. They listened to 102 improvisatory sentences situated within 34 groups of iterations. Each group consisted of a motivic n-gram that was repeated twice, then followed by three iterative translations. As this contribution is concerned mainly with the MT part of the model, we asked the experts to evaluate only the development of the musical motives throughout the iteration. Just like in Section 3.4, the evaluators considered pitch and rhythm when rating each translation from 1 to 5. Results for this MT task as shown in Table 11 are promising and experts noted their appreciation of the quality of the automatic improvisations.

Mean	min	max
4.03	3.81	4.25

Table 11: Subjective evaluation of iterative translation in taqāsīm generation.

5 Conclusion

We proposed a MT system for automatic instrumental improvisation in maqām music. By reducing the dimensions of the textual representation of musical sentences to only scale degree and quantized duration, it was possible to train SMT and NMT models using a parallel dataset (vocal and instrumental) that is both relatively small (6991 sentences) and diverse (8 different maqamat). The superior MT model was then used as a basis to generate real-time instrumental improvisation conditioned to listener feedback. To this end, we constructed a fully instrumental dataset of 717 improvisations from which we extracted frequent representative patterns (n-grams) for each maqam. MT was then applied, iteratively, starting first with the n-grams and then conditioned to listener feedback. Results were found promising based on subjective evaluations by experts from the maqām music tradition, as well as objective evaluation applied at two levels: the sentence level using the BLEU measure, and a higher level using statistical, musically informed metrics. The two objective measures were found consistent with each other. Future work will include investigating the influence of the following factors on the performance of music MT models: musical quality, size, and average sentence length of the (sub-)dataset.

6 Acknowledgements

Thanks to the Association of Francophone Universities (AUF) and the Deanship of Scientific Research at the University of Jordan for their contribution to funding this project. This work was also partially supported by the Research Council of Norway through its centres for excellence scheme, project number 262762.

References

Al-Abbas, Habib Thaher. 1986. *Nathariat Al-Musiqa al-Arabia [Arab Music Theories].* Ministry of Information, Baghdad.

Al-Ghawanmeh, Fadi and Kamel Smaïli. 2018. Statistical Machine Translation from Arab Vocal Improvisation to Instrumental Melodic Accompaniment. *Journal of International Science and General Applications*, 1(1):11–17.

Al-Ghawanmeh, Fadi, Mohamed-Amine Menacer, and K Smaïli. 2019. Accompaniment to arab vocal improvisation based on statistical machine translation: Objective and subjective evaluation. In Schiavio, Andrea, E Xypolitaki, C Scuderi, A Seither-Preisler, and Richard Parncutt, editors, *CIM19: Conference on Interdisciplinary Musicology-Embodiment in Music. Book of Abstracts.* University of Graz, Austria.

Al-Ghawanmeh, Fadi M, Melissa J Scott, Mohamed-Amine Menacer, and Kamel Smaïli. 2021. Predicting and critiquing machine virtuosity: Mawwal accompaniment as case study. In *International Computer Music Conference*.

Al-Ghawanmeh, Fadi. 2012. Automatic accompaniment to arab vocal improvisation "mawwāl". Master's thesis, New York University.

Asesh, Aishwarya. 2022. Markov chain sequence modeling. In *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, pages 1–6. IEEE.

Berment, Vincent. 2004. *Méthodes pour informatiser les langues et les groupes de langues peu dotées*. Ph.D. thesis, Université Joseph-Fourier-Grenoble I.

Boulos, Issa. 2021. Inside Arabic Music: Arabic Maqam Performance and Theory in the 20th Century. By Johnny Farraj and Sami Abu Shumays. *Music and Letters*, 102(1):171–172, 07.

Dai, Shuqi, Huiran Yu, and Roger B Dannenberg. 2022. What is missing in deep music generation? a study of repetition and structure in popular music. *arXiv preprint arXiv:2209.00182*.

Dannenberg, Roger B. 1984. An on-line algorithm for real-time accompaniment. In *ICMC*, volume 84, pages 193–198.

Farraj, Johnny. 2007. Arabic musical forms (genres). *Maqam World, Accessed on Jan. 5, 2023 from: http://www.maqamworld.com/forms.html*.

Forsyth, Jonathan P. 2016. *Automatic musical accompaniment using finite state machines*. Ph.D. thesis, New York University.

Kahel, Darin. 2021. Music is feeling: Tarab: a phenomenon of arab musical culture. *Independent thesis Basic level, Uppsala University*.

Kalonaris, Stefano, Thomas McLachlan, and Anna Aljanaki. 2020. Computational linguistics metrics for the evaluation of two-part counterpoint generated with neural machine translation. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 43–48.

Kisserwan, Ali. 2016. *Taqasim*. Muntada Al-Ma'aref, Beirut.

Klein, Guillaume, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion*, pages 177–180.

Krauwer, Steven. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, page 15.

Magnusson, Thor. 2021. The migration of musical instruments: on the socio-technological conditions of musical evolution. *Journal of New Music Research*, 50(2):175–183.

Mo, Ying. 2022. Designing an automatic piano accompaniment system using artificial intelligence and sound pattern database. *Mobile Information Systems*, 2022.

Nettl, B. 2007. taqsīm. *Encyclopedia Britannica. accessed on Oct. 18, 2022 from: https://www.britannica.com/art/taqsim*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Racy, Ali Jihad. 1998. Improvisation, ecstasy, and performance dynamics in arabic music. *In the course of performance: Studies in the world of musical improvisation*, pages 95–112.

Racy, Ali Jihad. 2004. *Making music in the Arab world: The culture and artistry of Tarab*. Number 17. Cambridge University Press.

Rao, Anyi and Francis Lau. 2018. Automatic music accompanist. *arXiv preprint arXiv:1803.09033*.

Ren, Yi, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1198–1206.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ünal, Erdem, Bariş Bozkurt, and M Kemal Karaosmanoğlu. 2014. A hierarchical approach to makam classification of turkish makam music, using symbolic data. *Journal of New Music Research*, 43(1):132–146.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vercoe, Barry. 1984. The synthetic performer in the context of live performance. In *Proc. ICMC*, pages 199–200.

Yang, Li-Chia and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784.