Statistical Machine Translation from Arab Vocal Improvisation to Instrumental Melodic Accompaniment

Fadi Al-Ghawanmeh, Kamel Smaïli

Statistical Machine Translation from Arab Vocal Improvisation to Instrumental Melodic Accompaniment

Fadi Al-Ghawanmeh[1], *Kamel Smaïli*[2]

[1]Music Department, University of Jordan, Jordan
[2]SMarT Group, LORIA, F-54600, France
[1]f_ghawanmeh@ju.edu.jo, [2]kamel.smaili@loria.fr

Abstract

Vocal improvisation is an essential practice in Arab music. The interactivity between the singer and the instrumentalist(s) is a main feature of this deep-rooted musical form. As part of the interactivity, the instrumentalist recapitulates, or translates, each vocal sentence upon its completion. In this paper, we present our own parallel corpus of instrumentally accompanied Arab vocal improvisation. The initial size of the corpus is 2779 parallel sentences. We discuss the process of building this corpus as well as the choice of data representation. We also present some statistics about the corpus. Then we present initial experiments on applying statistical machine translation to propose an automatic instrumental accompaniment to Arab vocal improvisation. The results with this small corpus, in comparison to classical machine translation of natural languages, are very promising: a BLEU of 24.62 from Vocal to instrumental and 24.07 from instrumental to vocal.

Index Terms: Arab music, Statistical machine translation, Automatic accompaniment, Maqam, Mawwal.

1. Introduction

Vocal improvisation is a primary musical form in Arab music. It is called Mawwal in the eastern part of the Arab world and istikhbar in the Maghreb. It is a non-metric musical practice that shows the vocalist's virtuosity when singing narrative poetry. It is tightly connected to the sense of saltanah, or what can be referred to as modal ecstasy. In performance, an instrumentalist set the stage for the singer by performing an improvisation on the given Maqam of the Mawwal. Then, the aesthetic feedback loop between the singer and the accompanying instrumentalists goes on. The Instrumentalists interact with the singer by playing along with him or her throughout every vocal sentence, then by recapitulating that sentence instrumentally upon its completion [1][2]. The audience takes part of this loop of aesthetic feedback especially by reacting to performers' expressiveness and virtuosity. This can be expressed by clapping or other means of showing excitement. Early contributions toward automating the instrumental musical accompaniment started in the mid-eighties [3][4]. However, researching automatic accompaniment in the context of Arab music started just recently [5], and has not yet been introduced to the capabilities and complexities of machine learning. Toward proposing an improved automatic accompaniment to Arab vocal improvisation, we studied the part of the accompaniment in which the instrumentalist recapitulates, or translates, the singer's musical sentence upon completion. To handle this challenge, we imagined it as a statistical machine translation problem, and then make use of techniques previously used in computational linguistics. Accordingly, our experiments require a parallel corpus consisting of vocal sentences and corresponding instrumental responses. Building our own corpus has been a necessity due to the lack of available transcriptions of accompanied Arab improvisations, also because selecting accompanied improvisations from the web and transcribing them automatically can be challenging for a variety of reasons. In our work we applied automatic transcription indeed, but on our own recordings performed by our singers and instrumentalists in equipped recording rooms. This was to ensure decent machine transcription. The remaining of this paper is organized as follows: we present related work in section two. In section three we discuss the idea of looking at the challenge of automating the melodic accompaniment from the perspective of statistical machine translation. In section four we present out corpus, and we apply machine translation experiments on it in section five, then results are presented in section six.

2. Related Works

Several harmonic accompaniment models have been proposed for different musical styles, such as jazz [6] and chorale style[7]. More generic models were also proposed, such as [8], which considered rock and R&B, among others. Several techniques were applied in the context of harmonic accompaniment, such as musical knowledge, genetic algorithms, neural networks and finite-state methods [9]. Fewer contributions considered non-harmonic accompaniment, including [5] and [10], who proposed Arab and Indian style melodic accompaniment, respectively. These last models used musical knowledge rather than machine learning methods. The model in [5] suggested a knowledge-based accompaniment to Arab vocal improvisation, Mawwal. The melodic instrumental accompaniment lines were very simple and performed slightly modified, or simplified, versions of vocal figures, all in heterophony with the vocal improvisation. Then and upon completion of each vocal figure, there was an instrumental imitation that repeated a full or partial parts of the vocal figure at a speed that could vary slightly from the speed of the vocal. In [11] and when analyzing scores of vocal improvisations along with correspond-

ing oud accompaniment, it was illustrated that, although at times the melodic lines of the instrumental accompaniment might follow the progression of the vocal lines, the particular melodic contour might twist in a way that is challenging to model. Indeed, such results encourage experimenting with corpus-based approaches to improve the automatic accompaniment. In [12] a corpus for arab-Andalusian music was built for computational musicology. The corpus consisted of audio materials, metadata, lyrics and scores. The contribution accented on the importance of the task of determining the design criteria according to which corpora are built. In [13] a research corpus for computational musicology was presented and consisted of audio and metadata for flamenco music. The contribution stressed on the idea that the distinctiveness of melodic and rhythmic elements, as well as its improvised interpretations and diversity in styles, are all reasons making flamenco music still largely undocumented. In [14] a parallel corpus of music and corresponding lyrics was presented. Crowdsourcing was used to enhance the corpus with notes on six basic emotions, annotated at line level. Early experiments showed promising results on the use of such corpus for song processing, particularly on emotion classification.

3. Melodic accompaniment and language models

Statistical Language Models work toward estimating the distribution of a variety of phenomena when processing natural languages automatically. These models seek regularities as a mean to improve the performance of applications [15]. In this contribution we investigated applying techniques common in statistical machine translation to handle the problem of automating the accompaniment to Arab vocal improvisation. In other words, we investigated translating the vocal improvisation into an instrumental accompaniment. We handled this translation problem sentence by sentence. Each vocal idea – whether as short as a motive or as long as a sentence – was considered a distinct musical sentence. The same was applied to instrumental responses; each response to the singer's previous sentence was considered one instrumental sentence. Indeed, In the Mawwal practice, the singer separates vocal sentences with relatively long rests, and accompanying instrumentalists fill these rests by recapitulating the singer's previous sentence. This type of instrumental response is referred to as "tarjama," literally meaning "translation" [2].

In general, each musical sentence consists of several musical notes, and each note has two main features: pitch and duration. In our proposed approach we represented them as scale degree and quantized duration, respectively. Section 5 justifies this choice of representation with further clarification. For each sentence, whether vocal or instrumental, we considered the degree as an element and the quantization step as another element. Elements might also be called words, as in natural languages. Figure 1 shows a score of a musical idea (or sentence, as in natural languages). It is a descending four-note motive in the maqam bayati that has its tonic on the note *D*. So the scale degrees of this sentence, respectively, are: 3^{rd} degree, 2^{nd} degree, 1^{st} degree and 1^{st} degree (one octave lower). In our approach we neither document nor process musical sentences in their traditional graphical music transcription. We rather use textual representations so we can apply statistical techniques common in natural languages processing directly to text files. Now for the textual representation of the musical idea, or sentence, in figure one; the first two elements are (dg_3) and (dr_6), both belong to the first note and tell its scale degree and quantized duration, respectively. This means that the scale degree of this note is 3, and the duration is of rank 6. The full textual sentence for this musical sentence is: $(dg_3)(dr_6)(dg_2)(dr_3)(dg_1)(dr_5)(dg_1)(dr_8)$.

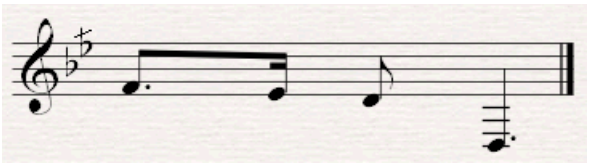

Figure 1: Example of a short musical idea in maqam bayati

4. The corpus

We built our own corpus with initial size of 2779 parallel sentences (vocal and instrumental). The goal is to use it to construct a statistical language model and apply a statistical machine translation paradigm. In this section we justify the need for building our own corpus and explain the procedure of building it. We also present some statistics about our corpus.

4.1. Why build it ourselves?

There are two main reasons that led us to build the corpus ourselves. Firstly, there is a lack of available transcriptions of Arab vocal improvisation, and it is much more difficult to find instrumentally accompanied improvisations. This is while taking into consideration that machine learning usually needs thousands of musical figures, not tens nor hundreds even. Secondly, although there are plenty of recordings of accompanied Mawaweel (plural of Mawwal) available on several audio- and video-sharing websites, transcribing such Mawaweel automatically is very challenging for a variety of reasons, including:

- The challenge of automatically transcribing the vocal improvisation with several instrumental melodic lines that are improvising accompaniment in a non-metric context.
- This musical form is highly interactive; so clapping and shouting from the audience can make the process more challenging.
- Arab music has many different Maqamat, and the same Maqam can have differences in microtonal tuning across different regions, especially for neutral tones. It is also common for the Mawwal to include modulations from a particular Maqam to others. Transcribing unknown audio files would

require a robust Maqam-finding algorithm. This is a different research problem that is tackled, yet not completely solved, by other researchers [16]. Indeed, automatically selecting and transcribing quality Mawaweel performances with instrumental accompaniment from YouTube and other online sources is a research challenge that needs further research. Unfortunately this was not within the scope of this project. For the reasons above, neither relying on available transcriptions nor transcribing Mawaweel from the Internet could have been a viable solution for building our parallel corpus at this time. We therefore decided to build our own corpus with our own singers, MIDI keyboard instrumentalists, and equipped recording rooms. Standardizing the recording process allowed us to avoid the issue of transcription quality in this research.

4.2. Procedure of building the corpus

To build the parallel corpus, we decided to use live vocal improvisation and Arab keyboard accompaniment. Indeed, the keyboard can emulate Arab instruments to a sufficient degree, and many singers today are accompanied by keyboardists rather than acoustic instruments. Moreover, transcribing keyboard accompaniment has perfect accuracy. This is because we only export the MIDI file that includes the transcription details, such as pitch and duration, as opposed to applying signal processing tasks to convert audio to transcription. In the latter approach, accuracy is decent, yet not perfect. In other words, when we sequence a MIDI score derived from a keyboard instrument, we hear the exact transcribed performance, but when we sequence a score of automatically transcribed audio, we are more likely to hear a deformed version of the original performance. Our choice reproduced the real-life scenario of the desired Mawwal automatic accompaniment, where the input is a vocal signal transcribed with a decent, yet not perfect, accuracy, and the output is an instrumental accompaniment that recapitulates the vocal input and its score is generated and reproduced audibly with perfect accuracy. Accordingly, building instrumental corpora using MIDI instruments would allow for incorporating instrumental accompaniment signals without deformity caused by transcription inaccuracy.

4.3. Corpus statistics

Statistics on the parallel corpus as a whole are presented in Table 1. As shown in the table, the vocal improvisation is in general longer than the instrumental accompaniment. This is although the number of instrumental notes is bigger. This is normal because the keyboard instrument imitated a plucked string instrument, the oud. Thus, the sound does not sustain for a long time, and this requires the instrumentalist to keep plucking in order to keep the instrument sounding. For both vocal and oud, the ranges of durations of notes are very wide. The table also shows that the overwhelming majority of vocal sentences lay within one octave; also half of the instrumental sentences lay in this pitch range. Table 2 presents corpus statistics at sentence level. For both vocal and instrumental sentences, it is clear that the sentence length may vary extremely. The sentence can be as short as one note or as long to have tens of notes.

	Vocal	**Instrumental**
Total duration	17907 s	13787 s
Note count	35745	55667
Total number of sentences	2779	2779
Percentage of sentences with tone range within octave	83.62	49.44
Maximum note duration	7.7 s	4 s
Minimum duration	0.14 s	0.002 s
Mean of durations	0.5 s	0.24 s
STD of durations	0.45	0.21

Table 1: Statistics on the parallel corpus as a whole

	Vocal corpus	**Instrumental**
Maximum note count	82	140
Minimum note count	1	1
Averages note count	12.86	20.03
STD of note count	10.70	17.96

Table 2: Statistics on the parallel corpus within one sentence

5. Data representation

The development of quality NLP models requires very large corpora. Our corpus, however, is both small and diverse. It is important, then, to represent this musical data with minimal letters and words from our two proposed languages, vocal improvisation and instrumental response. Yet it is also crucial that such minimization not deform the essence of the musical data. We analyze two main musical elements in this corpus, pitch and duration, and represent them as scale degree and quantized duration. The following two sub-sections discuss this process in detail.

5.1. Scale degree

Our corpus draws from a wide variety of Maqamat (musical modes), including Maqamat with neutral tones (tones with $\frac{3}{4}$ interval), and transpositions of Maqamat to less keys. Furthermore, the pitch range of both the vocal improvisation and the instrumental accompaniment can exceed two octaves. When using pitches as letters in our proposed language, 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 Maqamat in varying keys, we decided to use scale degree representation. Arab Maqamat are often based on seven scale degrees, allowing us to have the total number of letters as low as seven. One drawback to this method, however, is the inability to distinguish accidentals, the pitches that deviate from the given Maqam. Applying this configuration to the

automatic transcriber of vocal improvisation, however, allows for a significantly improved transcription quality [10] that outweighs the necessity to track accidentals.

5.2. Quantized duration

Here we present two histograms of note durations, one for vocal improvisation and the other for oud accompaniment. Analyzing the histograms helps determine the best total number of quantization steps, and also the duration range of each step. We need to have as few steps as possible in order to have better translation results, but it is crucial to retain the quality of the translation. Figure 2 shows the histogram of note durations of the vocal improvisation. We adjusted the value of the pin size to 0.139 seconds, and this is the minimum note duration (MND) in our adopted solution for the automatic transcription of vocal improvisation. Figure 3 depicts the percentage of notes located within or below each pin in the vocal improvisation. As shown in this figure, 89.3% of the note durations are within or below the first 7 pins. The remaining durations, which are relatively very long, are concentrated along other upper pins. It therefore follows to group these long (upper) durations into two bigger pins, each of which holds about half of these long durations. While taking into consideration that the first pin is empty because no note can be below the MND of the transcriber, the total count of used pins, or language letters, for the vocal corpus is 8.

Figure 4 shows the histogram of note durations of instrumental accompaniment. We adjusted the value of the pin size to 0.07 second. This is half of the vocal pin size, because in our corpus, the average duration of oud notes is half of the average duration of vocal notes. Figure 5 illustrates the percentage of notes located within or below each pin in the instrumental accompaniment. As can be noticed from the figure, about 89.9% of the note durations are within or below the first 6 pins. The remaining durations, the relatively very long ones, are concentrated along other upper pins. We group these long durations into two bigger pins, each of which incorporates about half of these long durations. Accordingly, the total count of used pins, or language letters, for the oud corpus is 8.

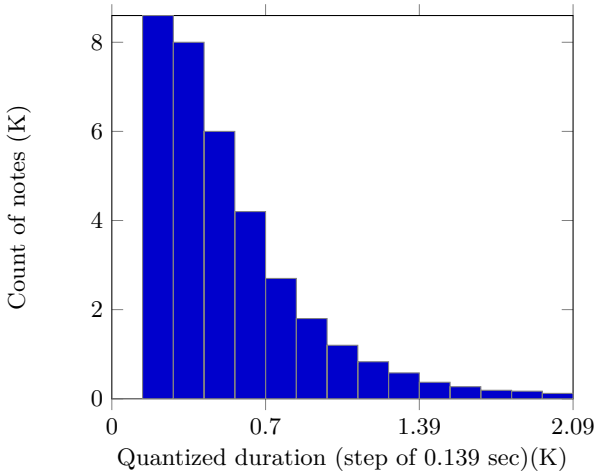

Figure 2: Note durations of the vocal improvisation

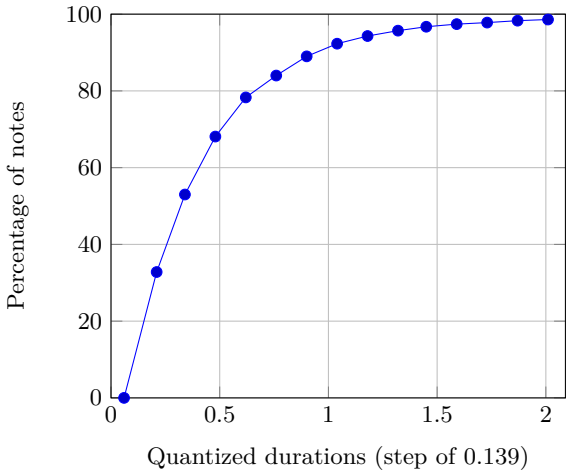

Figure 3: Percentage of vocal notes with durations below or equal each quantization step

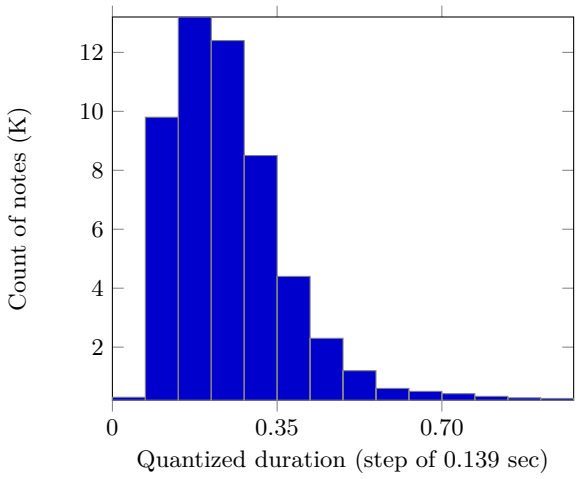

Figure 4: Note durations of instrumental accompaniment

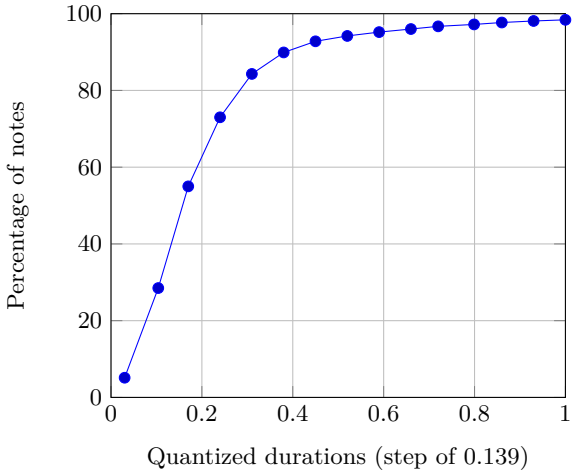

Figure 5: Percentage of instrumental notes with durations below or equal each quantization step

6. Machine Translation Experiments

Machine translation has been used to translate improvisation in both sides Vocal to Instrumental and Instrumental to vocal. In order to find the best model, we tested several representations of the music format. The MT system is a classical one with default settings: bidirectional phrase and lexical translation probabilities, distortion model, a word and a phrase penalty and a trigram language model. For the development and the test we used corpora of 100 parallel sentences for each of them. We used the bilingual evaluation understudy measure (BLEU)[17] to evaluate the quality of the translation. The formats and the BLEU scores are given In Table 3. Each format has different settings of three types of choice:

- Score reduction: this means that the music score was simplified using the formula in reference [5] in order to make the musical sentences shorter (with less notes). We used two representations for the reduced score:
  - Reduced Sustain: means that each unessential note was removed and its duration was added to its previous essential note, i.e., sustaining the essential note.
  - Reduced Silence: means that adjacent unessential notes were replaced by anew silent note that incorporates the durations of these unessential notes.
  - Unreduced: means no score reduction was applied. Apparently score reduction did not give good results, possibly because the reduction oversimplifies the patterns of melodic sentences and makes regularity ambiguous.
- Merging adjacent similar notes:
  - Merged: replace each two similar adjacent notes by one longer note to minimize the size of the musical sentences.
  - Unmerged: do not apply merging adjacent similar notes.
- Note representations:
  - Scale degree
  - Quantized duration
  - Scale degree and quantized duration

The best results have been achieved by merging adjacent similar notes, but without applying score reduction. Results are promising as the BLEU is 19.03. We also listened to the automatic accompaniment, and we believe it does have potential. Better BLEU score for this format was achieved when considering only one part of the musical information: either the duration or the scale degree, results were 21.27 and 24.62, respectively. The results of translating features separately (degrees alone and durations alone) could not be used to create accompaniment sentences, or translations, because creating a music notation need durations and degrees to have equal count. However, when separating the vocal sentence before translation into two parts, the number of resulting instrumental durations after translation does not necessarily equal the number of scale degrees. For example, when applying separated translation on a vocal sentence of 20 notes, i.e. 20 scale degree and 20 durations, the count of resulting instrumental translation can be 28 scale degrees and 32 durations. We cannot make a meaningful music notation in this case. Nevertheless, the results of translating musical features separately give an idea on where to apply more improvement in future research.

7. Conclusions

As part of efforts to improve the automated accompaniment to Arab vocal improvisation (Mawwal), in this contribution we considered the type of melodic accompaniment in which the instrumentalist(s) responses to, or translates, each vocal sentence after its completion. We built a relatively small parallel corpus; vocal and instrumental. We explained why we needed to construct this corpus ourselves. Then, we discussed data representation, also some statistics gathered from the corpus. After that we experimented with statistical machine translation. Results were positively surprising with a BLEU score reaching up to 24.62 from Vocal to instrumental, also 24.07 from instrumental to vocal. In addition, listening to translated music assured that this approach of automatic accompaniment is promising. Future work will include expanding the parallel corpus and introducing subjective evaluation side by side with the objective BLEU.

8. Acknowledgements

The authors acknowledge financial support of this work, part of TRAM (Translating Arabic Music) project, by the Agence universitaire de la Francophonie and the Arab Fund for Arts and Culture (AFAC).

Format of data	Vocal → Oud	Oud → Vocal
Unreduced Unmerged Scale Degree and quantized Duration	7.87	14.01
Reduced Merged Sustain Scale Degree	7.95	11.25
Reduced Merged Silence Scale Degree and quantized Duration	9.21	7.30
Reduced Merged Silence Scale Degree	9.94	15.92
Reduced Merged Sustain Scale Degree and quantized Duration	11.58	9.07
Reduced Merged Silence quantized Duration	14.10	8.40
Unreduced Unmerged quantized Duration	15.66	18.76
Unreduced Unmerged Scale Degree	15.66	24.41
Reduced Merged sustain quantized Duration	16.93	11.16
Unreduced Merged Scale Degree and quantized Duration	19.03	9.66
Unreduced Merged quantized Duration	21.27	22.38
Unreduced Merged Scale Degree	24.62	24.07

Table 3: BLEU score for each format data

9. References

[1] A. J. Racy, "Improvisation, ecstasy, and performance dynamics in arabic music," *In the course of performance: Studies in the world of musical improvisation*, pp. 95–112, 1998.

[2] "Arabic musical forms (genres)," 2007. [Online]. Available: http://www.maqamworld.com/forms.html

[3] R. B. Dannenberg, "An on-line algorithm for real-time accompaniment," in *ICMC*, vol. 84, 1984, pp. 193–198.

[4] B. Vercoe, "The synthetic performer in the context of live performance," in *Proc. ICMC*, 1984, pp. 199–200.

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

[6] D. Martín, "Automatic accompaniment for improvised music," Ph.D. dissertation, Master's thesis, Département de technologies de l'information et de la communication, Universitat Pompeu Fabra, Barcelone, 2009.

[7] J. Buys and B. v. d. Merwe, "Chorale harmonisation with weighted finite-state transducers," in *23rd Annual Symposium of the Pattern Recognition Association of South Africa,*, 2012.

[8] I. Simon, D. Morris, and S. Basu, "Mysong: automatic accompaniment generation for vocal melodies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 725–734.

[9] J. P. Forsyth and J. P. Bello, "Generating musical accompaniment using finite state transducers," in *16th International Conference on Digital Audio Effects (DAFx-13)*, 2013.

[10] P. Verma and P. Rao, "Real-time melodic accompaniment system for indian music using tms320c6713," in *VLSI Design (VLSID), 2012 25th International Conference on*. IEEE, 2012, pp. 119–124.

[11] F. Al-Ghawanmeh, M. Al-Ghawanmeh, and N. Obeidat, "Toward an improved automatic melodic accompaniment to arab vocal improvisation, mawwāl," in *Proceedings of the 9th Conference on Interdisciplinary Musicology-CIM14*, 2014, pp. 397–400.

[12] M. Sordo, A. Chaachoo, and X. Serra, "Creating corpora for computational research in arab-andalusian music," in *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*. ACM, 2014, pp. 1–3.

[13] N. Kroher, J.-M. Díaz-Báñez, J. Mora, and E. Gómez, "Corpus cofla: a research corpus for the computational study of flamenco music," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 2, p. 10, 2016.

[14] C. Strapparava, R. Mihalcea, and A. Battocchi, "A parallel corpus of music and lyrics annotated with emotions." in *LREC*, 2012, pp. 2343–2346.

[15] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[16] M. A. K. Sağun and B. Bolat, "Classification of classic turkish music makams by using deep belief networks," in *INnovations in Intelligent SysTems and Applications (INISTA), 2016 International Symposium on*. IEEE, 2016, pp. 1–6.

[17] 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.