N-GRAM BASED STATISTICAL MAKAM DETECTION ON MAKAM MUSIC IN TURKEY USING SYMBOLIC DATA

Erdem Ünal
TÜBİTAK-BİLGEM
unal@uekae.tubitak.gov.tr

Barış Bozkurt
Bahçeşehir University
baris.bozkurt@bahcesehir.edu.tr

M. Kemal Karaosmanoğlu
Yildiz Technical University
kkara@yildiz.edu.tr

ABSTRACT

This work studies the effect of different score representations and the potential of n-grams in makam classification for traditional makam music in Turkey. While makams are defined with various characteristics including a distinct set of pitches, pitch hierarchy, melodic direction, typical phrases and typical makam transitions, such characteristics result in certain n-gram distributions which can be used for makam detection effectively. 13 popular makams, some of which are very similar to each other, are used in this study. Using the leave-one-out strategy, makam models are created statistically and tested against the left out music piece. Tests indicate that n-gram based statistical modeling and perplexity based similarity metric can be effectively used for makam detection. However the main dimension that cannot be captured is the overall progression which is the most unique feature for classification of close makams that uses the same scale notes as well as the same tonic.

1. INTRODUCTION

The makam/maqam/mugam concept is very central to music of a very large geographical region from Balkans to Kazakhstan, Iran, and North Africa. Automatic classification of makam is hence very important for music information retrieval technologies though not widely studied.

Computational studies on makam music can be very broadly classified into two categories based on the type of data being processed: symbolic or audio. While some works such as [1], [6], [11] propose systems for makam recognition from audio data, works on symbolic data appear to be much more limited, probably due to lack of machine readable data. In [16], Şentürk and Chordia use Variable-Length Markov Models (VLMM) to predict the melodies in the *uzunhava* (long tune) form, a melodic structure in Turkish folk music. In [2] (which is a shorter version of [8]), Alpkoçak and Gedik present the first and only study on n-grams for makam recognition. Unfortunately, due to several deficiencies, reliability of their results is questionable. The paper presents classification results without cross-validation, uses limited and questionable data (20 pieces for each of 10 makam where data is represented with 12 notes in an octave while today's notation uses 24 notes in an octave). Due to such deficiencies, a new work needs to be conducted exploring the potential of n-grams in automatic makam recognition from symbolic data. Our main contributions in this study are: in addition to the 12-TET (Tone Equal Temperament) representation used in [2], [9], we also used data represented using the official theory of makam music in Turkey (TMMT) which uses 24 tones (unequally spaced) in an octave, and holding a larger database, and challenging makam sets, we were able to test the potential of n-gram based statistical approach in makam recognition more reliably. We also tested makam detection performance using comma level intervallic movements, showing how this system can be used in real life applications using audio data only.

In the MIR literature, makam recognition can be considered, to some level, as a key finding or a mode finding problem. However, there appears to be important differences between the concepts of key, mode and makam (a detailed discussion can be found in [3]). In the makam system, different makams can be constructed using the same set of pitches, the same set of tetrachord - pentachord formulation and the same tonic. Two examples are presented in Figure 1; the scale for makam *Hüseyni* and makam *Muhayyer* (top figure) and makam *Uşşak* and makam *Beyati* (bottom figure). Then, pitch hierarchy, melodic direction, typical phrases and typical makam transitions appear to be the discriminating features for makams having the same set of pitches and tonic.

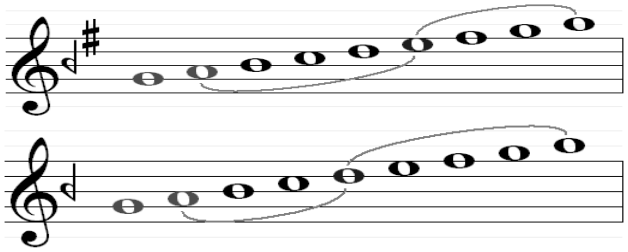

Figure 1. Scale used for makam Hüseyni and makam Muhayyer (top), makam Beyati and makam Uşşak

The listed characteristics have important influences on the pitch-class distribution of a given piece in a given makam, as in the case of key or mode in Western music

[13]. For that reason, the processing of pitch (class) histograms appears as the most common approach for computational studies of makam music [8]. N-grams can be considered to be an extension to this approach, where distributions of fixed length note-sequences are used in addition to single note pitch distributions.

The n-grams approach [10], from the text retrieval domain, have been widely used in computational studies on Western music. Various applications exist including indexing [6], query processing and music similarity computation [7], [17]. This study targets filling the gap for makam music in Turkey in the context of makam recognition. We first present the data, then details of implementation, results and discussions.

2. DATA

The current notation system (The Arel Theory (notation)) [4] used for TMMT assumes 24 notes in an octave. Although highly-criticized, almost all scores being used today are written in that system. While a large number of scanned scores are available on internet, machine readable data is very limited. Recently, we have announced the largest symbolic database of TMMT containing 1700 pieces in 155 makams [12]. Due to the availability at the time of the experiments, this study uses the following subset from [12]:

Makam name	*Total # of Songs*	*Total # of Notes*
Beyati	39	16,172
Hicaz	112	45,905
Hicazkar	48	17,950
Hüseyni	70	28,292
Hüzzam	63	26,842
Kürdilihicazkar	49	20,993
Mahur	51	22,037
Muhayyer	51	21,718
Nihavent	79	31,143
Rast	83	32,636
Saba	42	17,255
Segah	75	26,757
Uşşak	85	31,704
TOTAL	**847**	**339,404**

Table 1. Makam coverage and note statistics

The makam selection is based on three criteria: commonness, similarity and having sufficient number of samples in the database. For a classification study, it is beneficial to include similar classes and study the effects of such similarities in the classification performance. We have included in our set, makam couples which are stated to be differing only in melodic progression namely *Uşşak* - *Beyati* and *Muhayyer* - *Hüseyni* [14]. These couples share the same set of pitches, the same tonic and dominant (which is considered to be the boundary of tetrachord-pentachord division of the octave) as shown in Figure 1. As previous classification work on audio data showed, *Hüseyni* is also confused with *Uşşak* [8]. Therefore, the set includes challenging examples of classes.

2.1 Arel representation compared to 12-TET

In this work we test our system with two different representations of the symbolic data. The first one is the official theory of TMMT [4] and the second is the well-known 12-TET representation. The tonal space of the two systems are compared in Figure 2.

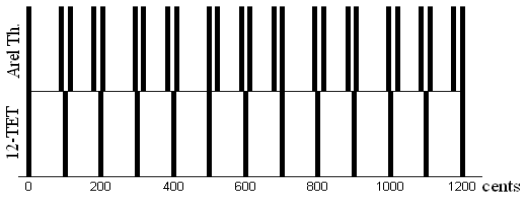

Figure 2. Tonal spaces of the Arel Theory Notation and 12-TET.

Being based on Pythagorean tuning, the 24 tones of Arel Theory are indeed close to 12-TET tones. While being a better representation (than 12-TET) for TMMT, Arel system is known to be insufficient in representing the practice. In this work we use the Arel Theory representation due to its wide use and the 12-TET representation, to be able to compare our results with [2],[9].

Arel theory uses two different but close formulations to represent notes and musical intervals, the first one being frequency ratios (such as 3/2, 9/8, etc.) and second one being the interval in integer multiples of Holdrian-commas (obtained by equal 53 divisions of an octave). The second is indeed a quantized version of the first and is more practical in explaining accidentals of the notation system as in Figure 3.

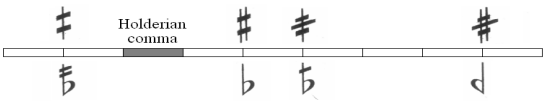

Figure 3. Accidentals used in the Arel Theory notation system

As shown in Figure 3, a whole-tone is composed of 9 Holdrian commas (will be referred as comma hereafter) in the Arel Theory system. In a machine readable format, it is convenient to name the notes using the comma steps such as *B4b1*, which corresponds to *B4* with a flat of single comma size, where *B4b4* would have a flat of 4 comma size. Alternatively each note can be represented using its distance to a reference note, for example *C1*, in commas. Such a representation makes it possible to easily obtain interval sizes (by simply subtracting the values assigned to each note as a distance in commas to a common reference) between consecutive notes which can further be used in modeling the progression (as in Section 4.4).

3. STATISTICAL MODELING

3.1 N-gram models

N-grams are widely used in computational linguistics, probability, communication theory and computational biology as well as music information retrieval [6], [7], [17]. N-grams predict X_i based on $X_{i-(n-1)}$, ..., X_{i-1}. In theory this is the information calculated by $P(X_i|_{Xi-(n-1)}$, ..., $X_{i-1})$. Given sequences of a certain set, one can statistically model this set by statistically counting the sequences that belong to it. In this study, according to the given note sequences that belong to the same makam, n-grams will be used to statistically model the pitch and intervallic space, as well as short melodic motifs to define makams.

The main hypothesis to be tested here is that, the short-time melodic contour and the frequency of makam specific notes are selective features for defining makams. This is why n-gram models are selected for training makam models using the Arel Theory notation. Given a microtonal notation sequence, using perplexity, the system will define how well the input sequence can be generated by the makam models in the database. The makam model that has the maximum similarity score is selected as the output of the system.

3.2 Smoothing

In practice, it is necessary to *smooth* the probability distributions by assigning non-zero probabilities to unseen words or n-grams. The reason is that models derived directly from the n-gram frequency counts have severe problems when confronted with any n-grams that have not been seen before which is called the "zero frequency problem". Different smoothing techniques are introduced in order to solve this problem [10]. Written-Bell smoothing technique available in the SRILM toolkit is used in our experiments [15].

3.3 Perplexity

Perplexity is a metric that is widely used for comparing probability distributions. The perplexity of a random variable *X* can be stated as the perplexity of the distribution over its possible values of *x*. Given a proposed probability model *q* (in our case: a makam model), evaluating *q* by asking how well it predicts a separate test sequence or set *x1, x2, ...,xN* (in our case: a microtonal note sequence) also drawn from *p*, can be performed by using the perplexity of the model *q*, defined by:

$$2^{\sum_{i=1}^{N}\frac{1}{N}log_2 q(x_i)} \qquad (1)$$

For the test events, we can see that better models will assign better probability scores thus a lower perplexity score which means it has a better potential to compress that data set. The exponent is the cross entropy per definition:

$$H(p,q) = -\sum_x p(x)log_2 q(x) \qquad (2)$$

The cross entropy thus the perplexity is the similarity measure between the test input and the makam models in the database. For each of the makam models defined, the system calculates the similarity metric to evaluate which makam is the most similar to the input sequence given.

4. EXPERIMENTAL SETUP

4.1 Leave-one-out

The experimental setup can be found in Figure 4, explaining how the leave-one-out strategy is inherited.

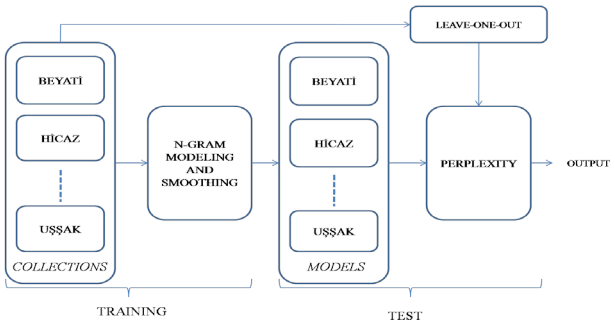

Figure 4. The leave-one-out experimental setup

There are 13 makam classes. Each of them has unequal number of music pieces. Since our approach is statistical, for each class, it is desirable to have a training set and a separate test set. This approach is feasible in case there is enough data. Since machine-readable microtonal notation is very hard to find in makam music, the "leave-one-out" strategy will be used in this experimental setup in order to avoid the negative effect of unequal set sizes. For each iteration of the experiment, one music piece will be selected as the input and the leftover music pieces will be used for training the genuine and the imposter makam classes. Using a probabilistic evaluation metric (perplexity), the system will calculate a similarity measure between the input and already built makam models.

4.2 Evaluation

Given a note sequence, the perplexity will estimate how well this sequence can be statistically generated by the makam models in our search space. Between each of the makam models, and the input sequence, the system calculates a similarity measure, and the makam that produces the maximum similarity measure becomes the output of the system. The performance criteria of the experimental procedure is binary, which is either a success or failure. The matching performance of the entire system, which is the accuracy (Recall), will be given as a proportion of successes over the total test trials in terms of Total Average (Tot-Ave) and the Weighted Average (W-Ave).

byati	hicaz	hczkr	hsyni	huzzm	krdhz	mahur	muhyr	nhvnt	rast	saba	segah	Ussak	Ref	Rcl.
24	0	0	3	0	0	0	1	0	0	0	0	11	**byati**	61.5
0	112	0	0	0	0	0	0	0	0	0	0	0	**hicaz**	100
0	0	48	0	0	0	0	0	0	0	0	0	0	**hczkr**	100
1	0	0	50	0	0	0	13	0	0	0	0	0	**hsyni**	71.4
0	0	0	0	62	0	0	0	0	0	0	1	0	**huzzm**	98.4
0	0	0	0	0	49	0	0	0	0	0	0	0	**krdhz**	100
0	0	0	0	0	0	51	0	0	0	0	0	0	**mahur**	100
2	0	0	11	0	0	0	35	0	3	0	0	0	**muhyr**	68.6
0	0	0	0	0	0	0	0	78	1	0	0	0	**nhvnt**	98.7
0	0	0	1	0	0	0	0	0	76	0	0	4	**rast**	91.6
0	0	0	1	0	0	0	0	0	0	41	0	0	**saba**	97.6
0	0	0	1	2	0	0	0	0	0	0	70	0	**segah**	95.9
16	0	0	11	0	0	0	4	0	3	0	2	49	**ussak**	57.6
55.8	100	100	64.1	96.9	100	100	63.6	100	91.6	100	95.9	70	Prc.	

Table 2. Confusion Matrix for Arel Theory Notation (n=3)

4.3 Tests with the Arel Theory Notation

In information retrieval, the output of such kind of systems are evaluated given 2 different measures. Given a music piece, the Recall (Rcl) suggests, how many of the queries for each of the makams are correctly found. On the other hand, precision is how many of the retrieved makams belong to the correct reference makam class. Precision becomes more meaningful when there is equal number of test trials from each makam classes. As seen from Table 3, 4 makams has perfect recall rate which are *Hicaz, Hicazkar, Kürdilihicazkar* and *Mahur*. The makam which shows the worst performance is *Beyati* as the recall rate is 61.5% and it is confused with *Uşşak*, the most (for n=3).

The confusion matrix also suggests that there are concrete similarities between these makam sets: *Beyati - Uşşak* and *Hüseyni - Muhayyer*. In theory these makam couples use exactly the same microtonal note sets as well as the tonics and the effect of this similarity can be practically seen in our experiments.

Table 3 shows the change in Recall metric when the order of the n-grams increased from 1 to 3. Also, the last column shows which n-gram shows the best performance with respect to Recall. As seen from results, for the makam, *Hicazkar, Hüseyni, Rast* and *Segah*, increasing the order of n-grams from 1 to 2 or 3, improves the makam detection performance of the classifier. For *Hicaz, Hüzzam, Mahur, Nihavent* and *Saba* increasing the order of the n-grams did not have any positive influence. On the other hand, increasing N has negative effect on the performance of the classifier for the makams *Beyati, Muhayyer* and *Uşşak*.

There might be a number of reasons for performance fluctuation within different makams. The one that we believe the most important is the unequal number of notes for training each makam classes. Even though a smoothing technique is used, the frequency of widely seen sequences become more dominant for makams that have few training samples (such as *Beyati, Hüzzam, Segah* and *Uşşak*), which makes these makams harder to be distinguished from the ones that are similar.

	n=1	**n=2**	**n=3**	**Best-N**
Beyati	64.1	56.4	61.5	1
Hicaz	100	99.1	100	1,3
Hicazkar	97.9	100	100	2,3
Hüseyni	50	64.3	71.4	3
Hüzzam	98.4	98.4	98.4	1,2,3
Kürdilihicakar	100	100	100	1,2,3
Mahur	100	100	100	1,2,3
Muhayyer	80.4	68.6	68.6	1
Nihavent	98.7	98.7	98.7	1,2,3
Rast	74.7	92.8	91.6	2
Saba	97.6	97.6	97.6	1,2,3
segah	93.2	97.3	95.9	2
Uşşak	68.2	62.4	57.6	1
Tot-Avg	86.3	87.9	88.2	3
W-Avg	86.4	87.4	87.8	3

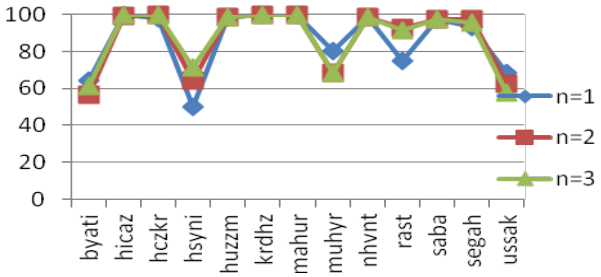

Table 3. Change in Recall w.r.t. n-gram order for data using the Arel Theory notation

4.4 Tests with microtonal intervals

Considering that the real world application of this system will operate with audio inputs, and it is known that a direct transcription of audio to the microtonal sequence used above is not easy, tests on data represented as sequence of microtonal intervals are also applied. The interval between consecutive notes are computed in commas as explained in Section 2.1 This also ensures that the system functionality is independent of the starting note of the music piece, the type or the tuning of the instrument that plays the piece.

	n=1	n=2	n=3	Best-N
Beyati	56.4	64.1	59	2
Hicaz	61.6	81.2	93.8	3
Hicazkar	66.7	85.4	85.4	2,3
Hüseyni	30	50	60	3
Hüzzam	69.8	87.3	85.7	2
Kürdilihicazkar	38.8	71.4	77.6	3
Mahur	80.4	90.2	94.1	3
Muhayyer	51	47.2	43.1	1
Nihavent	70.9	92.4	97.5	3
Rast	69.9	89.2	88	2
Saba	97.6	97.6	97.6	1,2,3
Segah	69.9	94.5	94.5	2,3
Uşşak	21.2	51.8	56.5	3
Tot-Avg	58.9	77.3	80.6	3
W-Avg	60.3	77.1	79.4	3

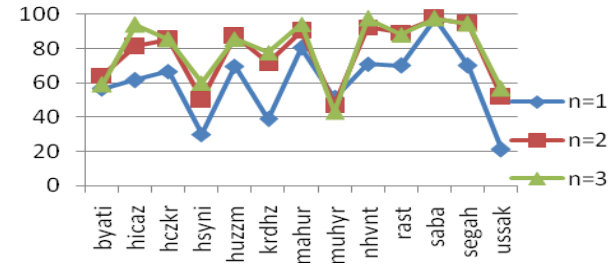

Table 4. Change in Recall w.r.t. n-gram order with data represented as microtonal intervals

	n=1	n=2	n=3	Best-N
Beyati	66.7	56.4	61.5	1
Hicaz	98.2	100	100	2,3
Hicazkar	91.7	97.9	97.9	2,3
Hüseyni	42.9	58.6	72.9	3
Hüzzam	98.4	98.4	98.4	1,2,3
Kürdilihicakar	100	100	98	1,2
Mahur	70.6	82.4	74.5	2
Muhayyer	80.4	70.6	66.7	1
Nihavent	97.5	97.5	97.5	1,2,3
Rast	72.3	75.9	80.7	3
Saba	97.6	97.6	97.6	1,2,3
segah	75.3	84.9	89	3
Uşşak	69.4	54.1	56.5	1
Tot-Avg	81.7	82.8	84.5	3
W-Avg	81.6	82.6	83.9	3

Table 5. Change in Recall w.r.t. n-gram order for 12 TET

The basic goal in this experimental setup is to achieve at least a close performance to the test explained in 4.3, and thus the cost of losing absolute note level information can be tested over the system performance. The makam detection performance for the microtonal representations can be seen in Table 4.

4.5 The Baseline: 12-TET Input Tests

Finally we evaluated the performance of our system on data represented using 12-TET since it is the representation used in the only available system in literature that does makam detection using n-grams [2], [9]. In addition to important differences in the implementation, modeling and evaluation, this study uses and compares different data representations, where in [2], [9] only the 12-TET representation is used. Since the basic strategy is building n-grams for both systems, we ran our experimental setup on the same database using the leave-one-out technique.

Table 5 shows the results with respect to increasing n-grams per each makam in the database. Since neither the evaluation nor the modeling technique is clearly explained in [2], [9], the standard modeling and smoothing techniques in our system was used when implementing the baseline (i.e. the system using the 12-TET representation). As seen from the results, the best performing n-gram order is 3, similar to results gathered from Arel theory tests. However, the system using the Arel Theory notation outperforms the baseline for both the Weighted Average and the Total Average.

The overall comparison of the performance of all the tests can be seen in Table 6. For n=3 where the best performance for all the systems were achieved, we observe that by using the Arel Theory notation as opposed to 12-TET, an improvement of 3.7% is achieved.

Recall	n=1	n=2	n=3
Arel Theory	86.3	87.9	88.2
12-TET	81.7	82.8	84.5
Delta (in commas)	58.9	77.3	80.6

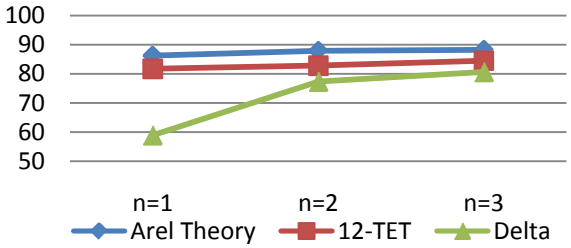

Table 6. Overall Performance Comparison.

5. DISCUSSION AND CONCLUSION

In this work, we implemented a perplexity based makam detection system on symbolic data of TMMT. N-gram based statistical makam models were built using the SRILM toolkit. Necessary smoothing was performed in order to compensate the negative effect of unequal set sizes.

Experimental set up was designed using the leave-one-out approach. For each of the test trials, one song from

the database was chosen as the input. The rest of the pieces were used for modeling the makam classes.

Three different experimental setups were created. The tests with data represented using Arel Theory showed that the overall recall performance of the system is 88.2%. Increasing the order of n-grams boosted the classification performance as expected. However, the effect is different for different makams. We observed that increasing the n-gram order did not help when trying to distinguish makams that use the same scale such as *Beyati-Uşşak* and *Muhayyer-Hüseyni*. The second experimental setup was for a real application case, where there is no direct note level transcription. For this test, the data is represented as intervals (in commas) between consecutive notes. This experiment was designed to provide reference information for research on makam detection directly from audio where exact note level transcription is not available. For audio, due to different instrumentation, and tuning, the only reliable information is the intervallic movement. The result showed that the makam detection accuracy is %80.6 using with n order 3. Note that, higher order n grams did not improve the experimental results beyond n=3 because of data sparsity.

For comparison with a previous study [2], [9], both the data representation in [2], [9] and additional representations (Arel and interval representation) are tested and compared. It is observed from tests using a large dataset, and challenging makam couple sets, that a system using Arel Theory representation outperforms a system using the 12-TET representation on average 3.7% percent. Increasing the n-gram order beyond 3 did not improve the performance of the tests due to lack of data.

Future work includes defining global tonal features that could help distinguishing makams having the same microtonal scale and tonic such as *Uşşak - Beyati* and *Hüseyni - Muhayyer*. Features related to global progression of the melody could give clues that cannot be captured by n-grams that concentrate more on local short length movements.

6. ACKNOWLEDGEMENT

This work was funded in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 267583 (CompMusic) and in part by TÜBİTAK ARDEB grant no:3501-109E196. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect those of the European Union or the TÜBİTAK.

7. REFERENCES

[1] S. Abdoli,. "Iranian Traditional Music Dastgah Classification," *ISMIR, Florida*,2011

[2] A. Alpkoçak and A. C. Gedik. "Classification of Turkish songs according to makams by using n grams," *In Proceedings of the 15. Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN)*, 2006.

[3] T. Aoyagi "Makam Rast: Intervallic Ordering, Pitch Hierarchy, Performance and Perception of a Melodic Mode in Arab Music," *University of California*, 2001.

[4] H. S. Arel, "Türk Musikisi Nazariyatı Dersleri, Hazırlayan Onur Akdoğu," *Kültür Bakanlığı Yayınları /1347, Ankara*, p.70. 1991,

[5] N. Darabi, N. Azimi and H. Nojumi, "Recognition of Dastgah and Makam for Persian Music with Detecting Skeletal Melodic Models," *The second annual IEEE BENELUX/DSP Valley Signal Processing Symposium* 2006

[6] S. Doraisamy, "Polyphonic Music Retrieval: The N - gram Approach," *PhD Thesis, University of London,* 2004.

[7] S. Downie "Evaluating a simple approach to music information retrieval: Conceiving melodic n-grams as text," *PhD thesis, University of Western Ontario, 1999.*

[8] A. C. Gedik, and B. Bozkurt, "Pitch-frequency histogram-based music information retrieval for Turkish music," *Signal Processing*, 90(4), 1049-1063, 2010.

[9] A.C. Gedik, C. Işıkhan, A. Alpkoçak and Y. Özer, "Automatic Classification of 10 Turkish Makams," *International Congress on Representation in Music & Musical Representation*, *İstanbul*, 2005.

[10] H. S. Heaps "Information Retrieval: Computational and Theoretical Aspects," *Academic Press*, 1978.

[11] L. Ioannidis, E. Gómez, and P. Herrera, *"*Tonal-based retrieval of Arabic and Middle-East music by automatic makam description*," CBMI*, 2011.

[12] M. K. Karaosmanoğlu, "A Turkish makam music symbolic database for music information retrieval: SymbTr," *submitted to ISMIR*, 2012.

[13] C.L. Krumhansl, "Cognitive Foundations of Musical Pitch," *Oxford University Press, NewYork*, 1990.

[14] I. H. Özkan, "Türk musikisi nazariyatı ve usulleri: kudüm velveleleri,". *Ötüken Neşriyat*, 2006.

[15] A. Stolcke: "Srilm – an Extensible Language Modeling Toolkit," *Proceedings of the International Conference on Spoken Language Processing,* 2002.

[16] S. Şentürk, and P. Chordia, "Modeling Melodic Improvisation in Turkish Folk Music Using Variable-length Markov Models," *ISMIR*, *Florida*, 2011.

[17] E. Unal, E. Chew, P. Georgiou and S. Narayanan, "Perplexity based cover song identification System for short length queries," *ISMIR, Florida*, 2011