SEARCHING LYRICAL PHRASES IN A-CAPELLA TURKISH MAKAM RECORDINGS

Georgi Dzhambazov, Sertan Şentürk, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona
{georgi.dzhambazov, sertan.senturk, xavier.serra}@upf.edu

ABSTRACT

Search by lyrics, the problem of locating the exact occurrences of a phrase from lyrics in musical audio, is a recently emerging research topic. Unlike key-phrases in speech, lyrical key-phrases have durations that bear important relation to other musical aspects like the structure of a composition. In this work we propose an approach that address the differences of syllable durations, specific for singing. First a phrase is expanded to MFCC-based phoneme models, trained on speech. Then, we apply dynamic time warping between the phrase and audio to estimate candidate audio segments in the given audio recording. Next, the retrieved audio segments are ranked by means of a novel score-informed hidden Markov model, in which durations of the syllables within a phrase are explicitly modeled. The proposed approach is evaluated on 12 a-capella audio recordings of Turkish Makam music. Relying on standard speech phonetic models, we arrive at promising results that outperform a baseline approach unaware of lyrics durations. To the best of our knowledge, this is the first work tackling the problem of search by lyrical key-phrases. We expect that it can serve as a baseline for further research on singing material with similar musical characteristics.

1. INTRODUCTION

Searching by lyrics is the problem of locating the exact occurrences of a key-phrase from textual lyrics in musical signal. It has inherent relation to the equivalent problem of keyword spotting (KWS) in speech. In KWS, a user is interested to find at which time position a relevant keyword (presenting a topic of interest) is spoken [16].

Most of the work on searching for keywords/key-phrases in singing (a.k.a lyrics spotting) has borrowed concepts from KWS. For spoken utterances phonemes have relatively similar duration across speakers. Unlike that, in singing durations of phonemes (especially vowels) have higher variation [8]. When being sung, vowels are prolonged according to musical note values. Therefore, adopting an approach from speech recognition might lack some singing-specific semantics, among which the durations of sung syllables. Furthermore, key-phrase detection has high potential to be integrated with other relevant MIR-applications, because lyrical key-phrases are often correlated to musical structure: For most types of music a section-long lyrical phrase is a feature that represents the corresponding structural section (e.g. chorus) in a unique way. Therefore correctly retrieved audio segments for, for example, the first lyrics line for a chorus can serve as a structure discovery tool.

In this work we investigate searching by lyrics in the case when a query represents an entire section or phrase from the textual lyrics of a particular composition. Unlike most works on lyrics spotting or query-by-humming, where a hit would be a document from an entire collection, in our case a hit is the occurrence of a phrase, being retrieved only from all performances of the given composition. In this respect the problem setting is more similar to linking melodic patterns from score to musical audio (addressed in [15]), rather than to lyrics spotting. We assume that the musical score with lyrics is present for the composition of interest. The proposed approach has been tested on a small dataset of a-cappella performances from a repertoire of Turkish Makam music. For a given performance, the composition is known in advance, but no information about the structure is given. Characteristic for Makam music is that, in a performance there might be reordering or repetitions of score sections.

2. RELATED WORK

2.1 Lyrics spotting

A recent work proved that lyrics spotting is a hard problem even when singing material is a-capella (for pop songs in English) [8]. The authors adopt an approach from KWS, using a compound hidden Markov model (HMM) with keyword and filler model. Keywords are automatically extracted from a textual collection of lyrics. This work's best classifier (multi-layer perceptron) yielded an f-measure of 44%, averaged over top 50% of keywords. Notably, the achieved results on singing material are not very different from results on spoken utterances of same keywords.

One of the few attempts to go beyond keywords is the work of [4]. Their goal was to automatically link phrases that appear in the lyrics of one song to the same phrase in another song. To this end, a keyword-filler model is

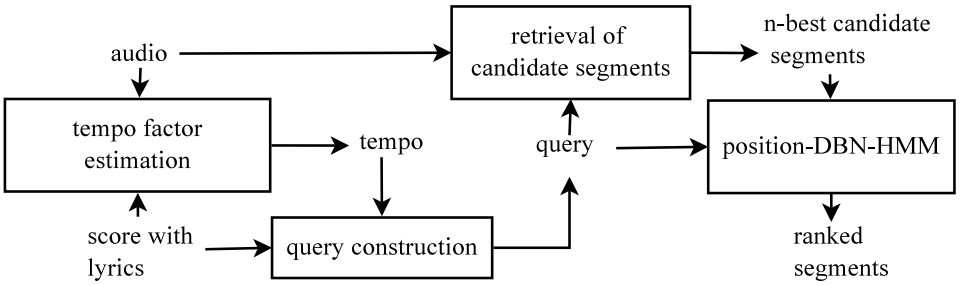

Figure 1. Approach overview: A key-phrase query is constructed in two variants: in the first stage candidate segments from audio are retrieved. In the second stage the query is modeled by a DBN-HMM aware of the position in music score. The DBN-HMM decodes and ranks candidate segments

utilized for detecting characteristic phrases (of 2-3 words) in sung audio. The method has been evaluated on polyphonic audio from Japanese pop, achieving 30% correctly identified links. Another modeling approach has been chosen in [1]. The authors propose subsequence dynamic time warping (SDTW) to find a match to an example utterance of a keyword as a subsequence of features from a target recording.

In summary, performance of the few works on lyrics spotting is not sufficiently good for practical applications. A probable reason for this is that hitherto approaches do not take into account the duration of syllables, which, as stated above, is an important factor that distinguishes speech from singing. In addition to that, syllable durations have been shown to be a strong reinforcing cue for the related task of automatically synchronizing lyrics and singing voice [3].

2.2 Position-aware DBN-HMMs

The modeling in most of the above mentioned approaches relies on HMMs. A drawback of HMMs is that their capability to model exact state durations is restricted, because the wait time in a state becomes implicitly an exponential distribution density [13, 20, IV.D].

One alternative to tackle durations can be seen in dynamic Bayesian networks (DBN) [12], which allow modeling of interdependent musical aspects in terms of probabilistic dependencies. In [18] it was proposed how to apply DBNs to represent jointly tempo and the position in a musical bar as latent variables in a HMM. In a later work this idea was extended by explicitly modeling rhythmic patters to track beats in music signals [7]. Relying on a similar DBN-based scheme, in [5] it has been shown, that the dependence of score position on structural sections makes it possible to link musical performances to score. In this paper for brevity we will refer to HMMs, which use DBNs to describe their hidden states, as DBN-HMMs.

3. APPROACH OVERVIEW

Figure 1 presents an overview of the proposed approach. A user first selects a query phrase from the lyrics of a composition of interest. Input are an audio recording, the queried lyrics and their corresponding excerpt from musical score. Only recordings of performances of the composition of the query are being searched. Output is a ranked list of retrieved hit audio segments and their timestamps.

One of the common approaches to KWS in speech, known as acoustic KWS, is to decompose a keyword into acoustic phoneme models [16]. Similarly, in a first stage of our approach a SDTW retrieves a set of candidate audio segments that are acoustically similar to the phonemes-decomposed query.

In a second stage, durations of the query phonemes are modeled by a novel DBN-HMM (in short position-DBN-HMM). Tracking the position in music score, it augments the phoneme models with score reference durations. Next, we run a Viterbi decoding on each candidate segment separately. This assures that only one (the most optimal) path is detected for each candidate audio segment. Only full matches of the query are considered as hits and all hit results are ranked according to the weights derived from the Viterbi decoded path.

In what follows each of the two stages is described in details, preceded by remarks on tempo estimation and how a query key-phrase is handled.

3.1 Tempo factor estimation

Often a performance is not played at the tempo indicated in the score. To estimate a factor τ, by which the average tempo of the performance differs relative to the score tempo, we use the tonic-independent partial audio-score alignment methodology explained in [15]. The method uses Hough transform, a simple line detection method [2], to locate the initial section from score in the audio recording. We derive the tempo factor τ from the angle θ of the detected line (approximating the alignment path) in the similarity matrix between the score subsequence and the audio recording.

3.2 Query construction

A selected lyrical phrase serves as a query twice: first a *simple query* for retrieval of candidate segments and then a *duration-informed query* for the decoding with position-DBN-HMM.

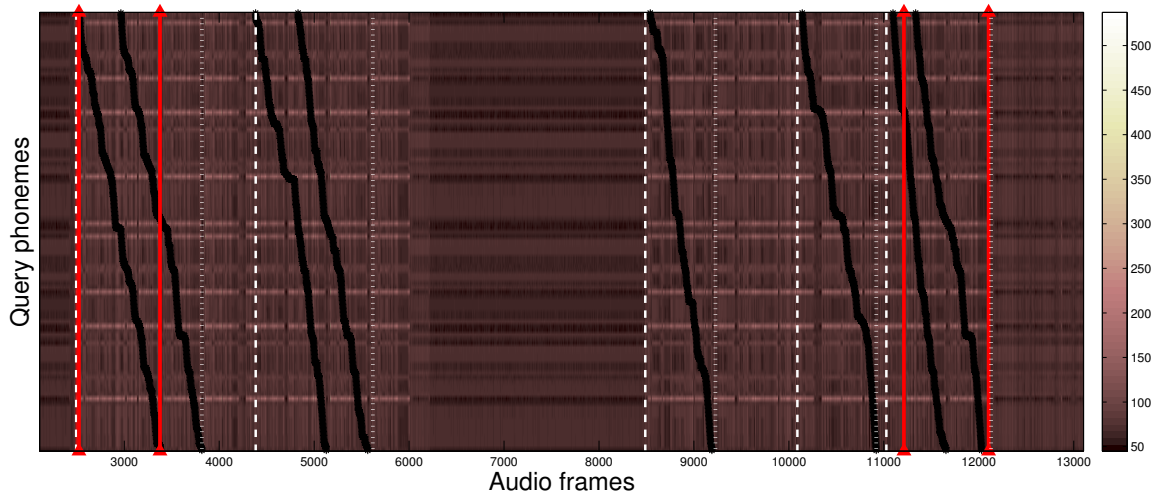

Figure 2. Distance matrix $\mathcal{D}$ for an audio excerpt of around 100 seconds. Retrieved paths are depicted as black contours. White vertical lines indicate beginning (dashed) and ending (dotted) of candidate audio segments, whereas red lines with triangle markers surround the ground truth regions.

3.2.1 Acoustic features

For each of 38 Turkish phonemes (and for a silent pause model) a 3-state HMM is trained from a 5-hours corpus of Turkish speech [14]. The 3 states represent respectively the beginning, middle and ending acoustic state of a phoneme. The transition probabilities of the HMMs are not taken into account. The phoneme set utilized has been developed for Turkish and is described in [14]. The formant frequencies of spoken phonemes can be induced from the spectral envelope of speech. To this end, we utilize the first 12 MFCCs and their delta to the previous time instant, extracted as described in [19]. For each state a 9-mixture Gaussian distribution is fitted on the feature vector.

3.2.2 Simple query

For the first step no score-position information is utilized: lyrics is merely expanded to its constituent phoneme models. Let $\lambda_n \in \Lambda$ be a state of phoneme model at position n in the query, where Λ is a set of all 3×38 states for the 3 phonemes.

3.2.3 Duration-informed query

Unlike the simple query, a duration-informed query exploits the note-to-syllable mappings, present in sheet music. For each syllable a reference duration is derived by aggregating values of its associated musical notes. Then the reference durations are spread among its constituent phonemes in a rule-based manner, resulting in reference durations R_ϕ for each phoneme ϕ [1].

To query a particular performance of a composition, R_ϕ are rescaled by the tempo factor τ (see section 3.1). Now this allows to define a mapping

$$f(p_n, s_n) \rightarrow \lambda_n \tag{1}$$

that determines the true state λ_n from a phoneme network, being sung at position p_n within a section s_n. A position p_n can span the duration of a section $D(s_n) = \sum_{\phi \in s_n} R_\phi$.

4. RETRIEVAL OF CANDIDATE SEGMENTS

SDTW has proven to be an effective way to spot lyrics, in which the feature series of an audio query can be seen as a subsequence of features of a target audio [1]. In our case a query of phoneme models Λ with length M can be seen as subsequence of the series of MFCC features with length N, extracted from the whole recording. To this end we define a distance metric for an audio frame y_m and model state λ_n as a function of the posterior probability.

$$d(m, n) = -\log\, P(y_m|\lambda_n) \tag{2}$$

where for phoneme state model λ_n

$$P(y_m|\lambda_n) = \sum_{c=1}^{9} w_{c,\lambda_n} \mathcal{N}(y_m; \mu_{c,\lambda_n}, \Sigma_{c,\lambda_n}) \tag{3}$$

with $\mathcal{N}$ being the Gaussian distribution from a 9-component mixture with weights w_{c,λ_n}. Based on the distance metric 2 a distance matrix $\mathcal{D}^{N \times M}$ is constructed.

4.1 Path computation

Let a warping path Ω be a sequence of L points $(\omega_1, .., \omega_l)$, $l \in [1, L]$ and $\omega_l = (m, n)$ refers to an entry $d(m, n)$ in $\mathcal{D}$. Following the strategy and notation of [11] to generate Ω we select step sizes $\omega_l - \omega_{l-1} \in \{(1, 1), (1, 0), (1, 2)\}$ corresponding respectively to diagonal, horizontal and skip step. A horizontal step means staying in the same phoneme in next audio frame. The step size $(0, 1)$ is disallowed because each frame has to map to exactly one phoneme model. To counteract the preference for the diagonal and

[1] In this work a simple rule is applied: consonants are assigned a fixed duration (0.1 seconds) and the rest of the syllable is assigned to the vowel.

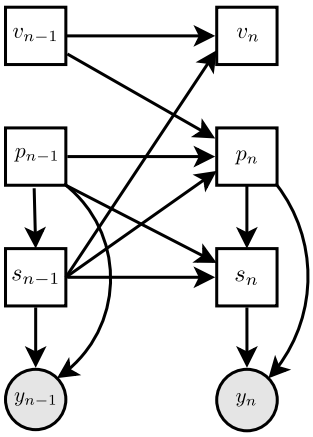

Figure 3. Representation of the hidden layers of the proposed model as a dynamic Bayesian network. Hidden variables (not shaded) are v - velocity, p - score position and s - section. The observed feature vector y is not shaded. Squares and circles denote respectively continuous and discrete variables

the skip step, we set rather high values for the local weights w_d and w_s [11].

A list of r candidate paths $(\Omega_1^*, ...\Omega_r^*)$ is computed by iteratively detecting the current path with maximum score. After having detected a path Ω^* with final position in frame n^* a small region of 5% of M: $(n^* - 5\%M,\ n^* + 5\%M)$ is blacklisted from further iterations, as described in [11]. This assures that the iterative procedure will not get stuck in a set of paths from a vicinity of a local maximum, but instead will retrieve as many relevant audio segments as possible.

4.2 Candidate segment selection

Analysis of the detected query segments revealed that a path often matches only partially the correct section segment. However, usually different parts of a segment have been detected in neighbouring paths. To handle this, we consider candidate segments - segments from the target audio, within which a frame y_m belongs to more than one path Ω. In other words, a candidate segment spans audio from the initial timestamp of the leftmost path to the final timestamp of the rightmost path. An example of retrieved candidate segments is presented in Figure 2. It can be seen that the two ground truth regions lie within candidate segments, which consist of more than one path.

5. POSITION-DBN-HMM

In this section we present the novel position-DBN-HMM for modeling a lyrical phrase. Its main idea is to incorporate the phonetic identities of lyrics and the syllable durations, available from musical score, into a coherent unit. The dependence of the observed MFCC features (that capture the phonetic identity) on musical velocity and score position are presented as DBN in Figure 3.

5.1 Hidden variables

1. Position p_n from musical score for a section ($p_n \in \{1, ..., D(s_n)\}$). $D(s_n = Q)$ is the total duration for a section s_n as defined in section 3.2.3. Note that $D(s_n)$ for a given section is different for two performances with different tempo, because of the tempo factor τ.
2. Velocity $v_n \in \{1, 2, ..., V\}$. Unit is the number of score positions per audio frame. Staying in state $v_n = 2$, for example, means that the current tempo is steady and around 2 times faster than the slowest one.
3. Structural section $s_n \in \{Q, F\}$ where Q is the queried section and F is a filler section. A filler section represents any non-key-phrase audio regions, and practically allows with equal probability being in any phoneme state (see section 5.3)

We compensate for tempo deviations by varying the local step size of the v variable. To allow handling deviations of up to half tempo, the derived $D(s_n = Q)$ is multiplied by 2. This means that $v = 1$ corresponds to half of the detected tempo. For the experiments reported in this paper, we chose $V = 5$. Furthermore we set $D(s_n = F) = V$. This assures that even in fastest tempo there is an option of entering the filler section.

The proposed model is different from the model proposed in [7] in two aspects:

- $D(s_n = Q)$ is not fixed but depends on the section of interest and the detected tempo of performance
- a section s_n (a pattern in the original model) is not fixed, but can vary between a query and filler states $\{Q, F\}$

Since all the hidden variables are discrete, one can reduce this model to a regular HMM by merging all variables into a single 'meta-variable' x_n:

$$x_n = [v_n,\ p_n,\ s_n] \tag{4}$$

Note that the state space becomes the Cartesian product of the individual variables.

5.2 Transition model

Due to the conditional independence relations presented in Figure 3, the transition model reduces to

$$\begin{aligned} P(x_n|x_{n-1}) = \quad & P(v_n|v_{n-1}, s_{n-1}) \times \\ P(p_n|v_{n-1}, p_{n-1}, s_{n-1}) \times \quad & P(s_n|p_{n-1}, s_{n-1}, p_n) \end{aligned} \tag{5}$$

5.2.1 Velocity transition

$$p(v_n|v_{n-1}) = \begin{cases} \phi/2, & v_n = v_{n-1} \pm 1 \\ 1 - \phi, & v_n = v_{n-1} \\ 0, & \text{else} \end{cases} \tag{6}$$

where ϕ is a constant probability of change in velocity and is set to 0.2 in this work.

5.2.2 Position transition

The score position is defined deterministically according to:

$$p_n = (p_{n-1} + v_{n-1} - 1) \mod D(s_{n-1}) + 1 \quad (7)$$

where the modulus operator resets the position to be in a beginning of a new section after it exceeds the duration of previous section $D(s_{n-1})$

5.2.3 Section transition

$$P(s_n|p_{n-1}, s_{n-1}, p_n) = \begin{cases} P(s_n|s_{n-1}), & p_n \leq p_{n-1} \\ 1, \ p_n > p_{n-1} \ \& & s_n = s_{n-1} \end{cases} \quad (8)$$

A lack of increase in the position is an indicator that a new section should be started. $P(s_n|s_{n-1})$ is set according to a transition matrix $A = \{a_{ij}\}$ where $i \in \{Q, F\}$ and self transitions a_{QQ} and a_{FF} for query and filler section respectively can be set to reflect the expected structure of the target audio signal. In this work we set $a_{QQ} = 0$, since we expect that a query might be decoded at most once in a candidate audio segment. The value $a_{FF} = 0.9$ is determined empirically.

5.3 Observation model

For the query section the probability of the observed feature vector in position p_n from section s_n is computed for the model state λ_n by a mapping function $f(p_n, s_n)$, introduced in section 3.2. A similar mapping function has been proposed for the first time in the DBN-HMM in [5].

Then

$$P(y_n|p_n, s_n = Q) = P(y_n|\lambda_n) \quad (9)$$

which reduces to applying the distribution defined in Equation 3.

In case of the filler section the most likely phoneme state is picked.

$$P(y_n|p_n, s_n = F) = \max_{\lambda \in \Lambda} P(y_n|\lambda) \quad (10)$$

Note that position p_n plays a role only in tracking the total section duration $D(s_n = F)$.

5.4 Inference

An exact inference of the 'meta-variable' x can be performed by means of the Viterbi algorithm. A key-phrase is detected whenever a segment of the Viterbi path $\bar{\Omega}$ passes through a section $s_n = Q$. The likelihood of this path segment is used as detection score for ranking all retrieved key-phrases.

6. DATASET

The test dataset consists of 12 a-cappella performances of 11 compositions with total duration of 19 minutes.

statistic	value
#section queries	50
average cardinality $\bar{C_q}$	3.2
maximum cardinality C_{qM}	6
#words per section	5-14
#sections per recording	6-16
#phonemes per section	26-63

Table 1. Statistics about queries (lyrics sections with unique lyrics) in the test dataset. The low value of $\bar{C_q}$ are due to the small number of performances per composition.

The compositions are drawn from the CompMusic corpus of classical Turkish Makam repertoire [17]. The a-capella versions have been sung by professional singers and recorded especially for this study. Scores are provided in the machine-readable symbTr format [6], which contain marks of section divisions. A performance has been recorded in-sync with the original recording, whereby instrumental sections are left as silence. This assures that the order, in which sections are performed, is kept the same [2] .

We consider as a query q each section from the scores, which has unique lyrics: in total 50. Note that the search space is restricted to all recordings of the composition, from which the section is taken. In a given recording we annotated the section boundary timestamps. Let C_q be the total number of relevant occurrences (cardinality) of a query q. Table 1 presents the average cardinality $\bar{C_q}$ and other relevant statistics about sections.

7. EVALUATION

7.1 Evaluation metrics

Having a ranked list of occurrences of each lyrical query, the search-by-lyrics can be interpreted as a ranked retrieval problem, in which the users are interested in checking only the top K relevant results [10]. This allows to reject irrelevant results by considering only top K results in the evaluation metric. We consider this strategy as appropriate since a query has low average cardinality ($\bar{C_q} = 3.2$). Let the relevance of ranked results for a query q be $[r_q(1), ..., r_q(n_q)]$ where n_q is the number of retrieved occurrences. Note that a detected audio segment is either hit or not, making $r_q(k) \in \{0, 1\}$.

For each of the queried score sections an average precision $\bar{P_q}$ at different values of K is computed as:

$$\bar{P_q} = \frac{1}{C_q} \sum_{k=1}^{K} r_q(k) P_q(k) \quad (11)$$

as defined in [10], where $P_q(k)$ is precision at k. The relevance $r_q(k)$ of k^{th} retrieved occurrence is binary and set to 1 only if both retrieved boundary timestamps are within a tolerance window of 3 seconds from ground truth. This window size has been introduced in [9] and is commonly used for evaluating structural segments. The hits are

[2] The dataset is available here: http://compmusic.upf.edu/turkish-makam-acapella-sections-dataset

K	1	2	3	4	5	6
SDTW	8.3	12.1	16.2	19.0	22.0	25.7
DBN-HMM	5.0	7.7	18.75	28.8	35.0	37.9

Table 2. MAPs (in percent) for ranked result segments for two system variants: baseline with SDTW and complete with position-DBN-HMM.

ranked by the likelihoods of the relevant Viterbi path segments. Resultsacc are reported in terms of mean average precision (MAP) as the average over all $\bar{P}_q$.

7.2 Experiments

To assess the benefit of the proposed modeling of positions, we conduct a comparison of the performance of the complete system and a baseline version without the position-DBN-HMM [3] . For the baseline, as result set we consider the audio segments corresponding to the list of candidate paths $(\Omega_1^*, ...\Omega_r^*)$ derived after SDTW (see section 4.1). As a ranking strategy, SDTW-paths are ordered by means of the sum of distance metrics $d(m, n)$, which is derived from the observation probability. We report results at different values for K in Table 2. Results for $K > C_{qM}$ are ommited. Furthermore, we picked empirically $r = 12$ candidate paths in SDTW, which is twice C_{qM}.

The results confirm the expectation that the performance of SDTW alone is inferior. Retrieving relevant candidate paths seemed to be very dependent on the weights w_d and w_s for the diagonal and skip steps. We noted that adapting weights for a recording according to the detected tempo factor τ might be beneficial, but did not conduct related experiments in this work. The optimal values ($w_d = 6.5$ and $w_s = 11$) in fact guaranteed good coverage of relevant segments in the slowest tempo in the dataset.

As K increases, the MAP for both DBN-HMM and SDTW improves, as more hits are being found on lower ranks. However top ranks are relatively low for DBN-HMM. This indicates that the Viterbi weigthing scheme might not be optimal. In general, MAP for DBN-HMM, at higher values at K gets substantially better than the baseline, which suggests that modeling syllable durations is beneficial. A further reason might be that the position-DBN-HMM can model tempo in a more flexible way and is thus not affected by the difference between the tempo indicated in the score and the real performance tempo.

7.3 Comparison to related work

For the sake of comparison to any future work we report in Table 3 the f-measure, derived from the precision $P_q(k)$ and recall $R_q(k)$ as defined in [10]. Unfortunately, no direct comparison to previous work on lyrics spotting [1,4,8] is possible, because these works rely on speech models for languages different from Turkish. Furthermore, the evaluation setting in none of the works is comparable to ours.

[3] To facilitate reproducibility of this research source code is publicly available here: https://github.com/georgid/Position-DBN-HMM-Lyrics

K	1	2	3	4	5	6
DBN-HMM	12.4	15.5	19.2	24.2	31.3	37.8

Table 3. F-measure (in percent) for the position-DBN-HMM for ranked results segments

In [8] a result is considered true positive if a keyword is detected at any position in an expected audio clip. The authors argue that since a clip spans one line of lyrics (only 1 to 10 words) this is sufficiently exact, whereas we are interested in detecting the exact timestamps of a key-phrase. In addition to that, their longest query has 8 phonemes, which is much less than the average in our setting.

In [4] the accuracy of the key-phrase spotting module is not reported, but instead only the percentage of the correctly detected links connecting key-phrases from a song to another song. It can be inferred from it that an upper bound on the performance of the key-phrase spotting lies around an accuracy of 30%. Further, on creating a link for a given key-phrase only the candidate section with highest score for a song has been considered, which might ignore any other true positives.

8. CONCLUSION

In this study we have investigated an important problem that has started to attract attention of researchers only recently. We tackle the linking between audio and structural sections from the perspective of lyrics: we proposed a method for searching in musical audio for the occurrences of a characteristic section-long lyrical phrase. We presented a novel DBN-based HMM for tracking sung phoneme durations. Evaluation on a-cappella material from Turkish Makam music shows that the search with the proposed model brings substantial improvement compared to a baseline system, unaware of syllable durations.

We plan to focus future work on applying the proposed model to the case of polyphonic singing. We expect further, that this work can serve as a baseline for further research on singing material with similar musical characteristics.

We want to point as well that, the proposed score-informed scheme is applicable not necessarily only when musical scores are available. Scores can be replaced by any format, from which duration information can be inferred: for example annotated melodic contour or singer-created indications along the lyrics.

Acknowledgements This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583) and partly by the AGAUR research grant.

9. REFERENCES

[1] Christian Dittmar, Pedro Mercado, Holger Grossmann, and Estefanıa Cano. Towards lyrics spotting in the

syncglobal project. In *Cognitive Information Processing (CIP), 2012 3rd International Workshop on*, pages 1–6. IEEE, 2012.

[2] Richard O Duda and Peter E Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[3] Georgi Dzhambazov and Xavier Serra. Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Sound and Music Computing Conference*, Maynooth, Ireland, 2015.

[4] Hiromasa Fujihara, Masataka Goto, and Jun Ogata. Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 281–286, Philadelphia, USA, September 14-18 2008.

[5] Andre Holzapfel, Umut Şimşekli, Sertan Şentürk, and Ali Taylan Cemgil. Section-level modeling of musical audio for linking performances to scores in turkish makam music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 19/04/2015 2015.

[6] M Kemal Karaosmanoğlu. A Turkish makam music symbolic database for music information retrieval: Symbtr. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, 2012.

[7] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, November 4-8 2013.

[8] Anna M. Kruspe. Keyword spotting in a-capella singing. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 271–276, Taipei, Taiwan, 2014.

[9] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):318–326, 2008.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[11] Meinard Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[12] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[13] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[14] Özgül Salor, Bryan L. Pellom, Tolga Ciloglu, and Mbeccel Demirekler. Turkish speech corpora and recognition tools developed by porting sonic: Towards multilingual speech recognition. *Computer Speech and Language*, 21(4):580 – 593, 2007.

[15] Sertan Şentürk, Sankalp Gulati, and Xavier Serra. Score informed tonic identification for makam music of turkey. In *Proceedings of 14th International Society for Music Information Retrieval Conference*, pages 175–180, Curitiba, Brazil, 2013.

[16] Igor Szöke, Petr Schwarz, Pavel Matejka, Lukás Burget, Martin Karafiát, Michal Fapso, and Jan Cernockỳ. Comparison of keyword spotting approaches for informal continuous speech. In *Interspeech*, pages 633–636, 2005.

[17] Burak Uyar, Hasan Sercan Atlı, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. A corpus for computational research of Turkish makam music. In *1st International Digital Libraries for Musicology Workshop*, pages 57–63, London, United Kingdom, 2014.

[18] Nick Whiteley, A. Taylan Cemgil, and Simon Godsill. Bayesian modelling of temporal structure in musical audio. In *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria (BC), Canada, October 8-12 2006.

[19] Steve J Young. *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer, 1993.

[20] Shun-Zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010.