ON THE USE OF NOTE ONSETS FOR IMPROVED LYRICS-TO-AUDIO ALIGNMENT

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

Lyrics-to-audio alignment aims to automatically match given lyrics and musical audio. In this work we extend a state of the art approach for lyrics-to-audio alignment with information about note onsets. In particular, we consider the fact that proceeding to next lyrics syllable usually implies a change to a new musical note. To this end we formulate rules that guide the transition between consecutive phonemes when a note onset in present. These rules are incorporated into the transition matrix of a variable time hidden Markov model (VTHMM) phonetic recognizer based on mel frequency cepstral coefficients (MFCCs), which are extracted in a way robust to background instrumental sounds. An estimated melodic contour is input to an automatic note transcription algorithm, from which the note onsets are derived. The proposed approach is evaluated on 12 acapella audio recordings of Turkish Makam music using a phrase-level accuracy measure. Evaluation of the alignment is also presented on a polyphonic version of the dataset in order to assess how degradation in the extracted onsets affects performance. Results show that the proposed model outperforms a baseline approach unaware of onset transition rules. To the best of our knowledge, this is the first work tackling lyrics tracking, which combines timbral features with a melodic feature.

1. INTRODUCTION

Lyrics are one of the most important aspects of vocal music. When a performance is heard, most listeners will follow the lyrics of the main vocal melody. The goal of automatic lyrics-to-audio alignment is to generate a temporal relationship between textual lyrics and sung audio. In this particular work, the goal is to detect the start and end times of every phrase (1-4 words) from lyrics.

In recent years there has been substantial amount of work on the extraction of pitch of predominant singing voice from polyphonic music [17]. Some algorithms have been tailored to the music characteristics of a particular singing tradition [13]. This has paved the way to an increased accuracy of note transcription algorithms. One of the reasons for this is that a correctly detected melody contour is a fundamental precondition for note transcription. On the other hand, lyrics-to-audio alignment is a challenging task: to track the timbral characteristics of singing voice might not be straight forward. Additional challenge is posed when accompanying instruments are present: their spectral peaks might overlap and occlude the spectral components of voice. Despite that, most work has focused on tracking change from one to another phoneme only by timbral features [5]. In fact, at the change of phoneme in parallel to timbre other musical aspects change: for example an articulation accent or change of pitch, both of which contribute to the perception of a distinct vocal note onset. The fact that the first vowel onset in a syllable occurs simultaneously to a note onset has been used successfully in rule-based synthesis of singing voice [20].

In this work we present a novel idea of how to extend a standard approach for lyrics-to-audio alignment by utilizing as a complementary cue automatically detected vocal note onsets. We apply a state of the art note transcription method to obtain candidate note onsets. The proposed approach has been evaluated on time boundaries of short lyrics phrases on acapella recordings from Turkish Makam music. An experiment on polyphonic audio reveals the potential of the approach for real-world application.

2. RELATED WORK

2.1 Lyrics-to-audio alignment

The problem of lyrics-to-audio alignment has inherent relation to the problem of text-to-speech alignment. For this reason most of current studies exploit an approach adopted from speech: building a model for each phoneme based on phoneme acoustic features [6, 14]. To model phoneme timbre usually MFCCs are employed. A state of the art work following this approach [6] proposes a technique to adapt a phonetic recognizer trained on speech: the MFCC-based speech phoneme models are adapted to the acoustics of singing voice. Further, automatic segregation of the vocal line is performed, in order to reduce the spectral content of background instruments. In general, in this approach authors consider only models of phonetic timbre and are thus focused on making them more robust as a mean to improve performance.

Few works for tracking lyrics combine timbral features with other characteristics of the main vocal melody. For example in [8] a system for automatic score following of

singing voice combines melodic and lyrics information: observation probabilities of pitch templates and vowel templates are fused to improved alignment. To our knowledge no work, which does not involve musical score, has employed features of the vocal melody contour in general, and note onsets in particular.

2.2 Automatic note segmentation

While the general problem of automatic music transcription has been a long-investigated problem, automatic singing transcription has attracted the attention of MIR researcher only in recent years. A fundamental part of singing transcription is automatic note segmentation. A probabilistic note event model, using a HMM trained on manual transcriptions is presented in [12]. The idea is that a note consists of different states representing its attack, sustain and decay phase. Then an onset is detected when the decoding path goes though an attack state of a new note.

A recent work on singing transcription with particularly good onset accuracy has been developed for singing voice from the flamenco genre [13]. It consists of two stages: predominant vocal extraction and note transcription. As the primary step of note transcription notes are segmented by a set of onset detection functions based on pitch contour and volume characteristics, which take into account the peculiar for flamenco singing high degree of microtonal ornamentation.

3. PROPOSED APPROACH

A general overview of the proposed approach is presented in Figure 1. An acapella audio recording and its lyrics are input. A variable time hidden Markov model (VTHMM), guided by phoneme transition rules, returns start and end timestamps of aligned words. For brevity in the rest of the paper our approach will be referred to as VTHMM.

First an audio recording is manually divided into segments corresponding to structural sections (e.g. verse, chorus) as indicated in a structural annotation, whereby instrumental-only sections are discarded. All further steps are performed on each audio segment. If we had used automatic segmentation instead, potential erroneous lyrics and features could have biased the comparison of a baseline system and VTHMM. As we focus on evaluating the effect of VTHMM, manual segmentation is preferred. In what follows each of the modules is described in details.

3.1 Vocal pitch extraction

To extract the melody contour of singing voice, we utilize a method that performs detection of vocal segments and in the same time pitch extraction for the detected segments [3]. It relies on the basic methodology of [16], but modifies the way in which the final melody contour is selected from a set of candidate contours, in order to reflect the specificities of Turkish Makam music: 1) It chooses a finer bin resolution of only 7.5 cents that approximately corresponds to the smallest noticeable change in Makam

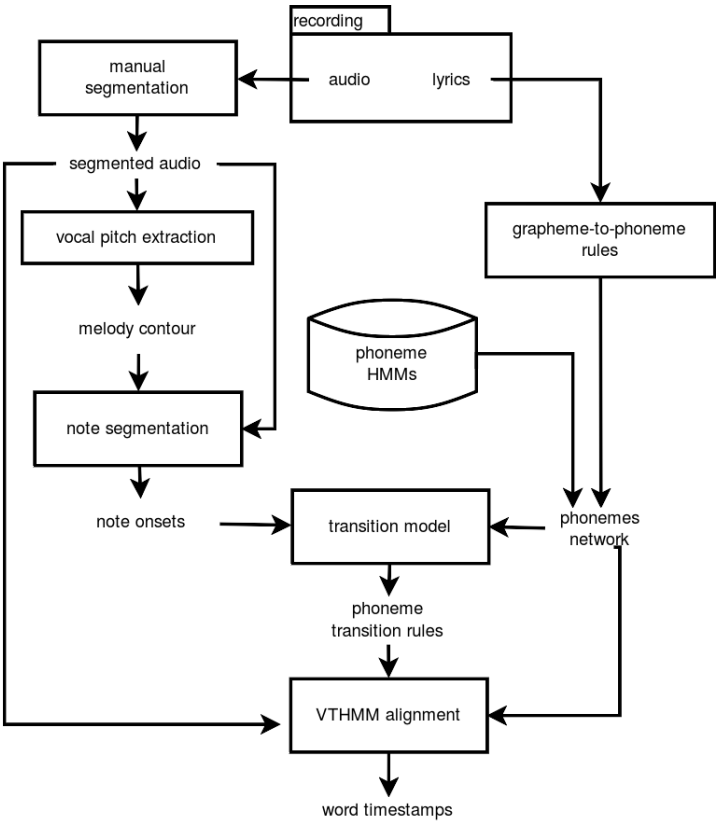

Figure 1. Overview of the modules of the proposed approach. One can see how phoneme transition rules are derived. Then together with the phonemes network and the features extracted from audio segments are input to the VTHMM alignment

melodic scales. 2) Unlike the original methodology, it does not discard time intervals where the peaks of the pitch contours have relatively low magnitude. This accommodates time intervals at the end of the melodic phrases, where Makam singers might sing softer.

3.2 Note segmentation

In a next step, to obtain reliable estimate of singing note onsets, we adapt the automatic singing transcription method, developed for polyphonic flamenco recordings [13]. It has been designed to handle singing with high degree of vocal pitch ornamentation. We expect that this makes it suitable for material from Makam classical singing having as well heavily vibrato and melismas. We replace the original first stage predominant vocal extraction method with the vocal pitch detection method described above.

The algorithm [13] considers two cases of onsets: interval onsets and steady pitch onsets. Gaussian derivative filter detects interval onsets as long-term changes of the pitch contour, whereas steady-pitch onsets are inferred from pitch discontinuities. As in the current work phoneme transitions are modified only when onsets are present, we opt for increasing recall at the cost of losing on precision. This is achieved by reducing the value of the parameters

cF: the minimum output of the Gaussian filter. The extracted note onsets are converted into a binary onset activation at each frame $\Delta n_t = (0, 1)$. Recall rates of extracted note onsets are reported in table 2.

3.3 Phoneme models

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 [21]. For each phoneme a one-state HMM, for which a 9-mixture Gaussian distribution is fitted on the feature vector. The lyrics are expanded to phonemes based on grapheme-to-phoneme rules for Turkish [18, Table 1] and the trained HMMs are concatenated into a phonemes network. The phoneme set utilized has been developed for Turkish and is described in [18]. A HMM for silent pause *sp* is added at the end of each word, which is optional on decoding. This way it will appear in the detected sequence only if there is some non-vocal part or the singer makes a break for breathing.

3.4 Transition model

We utilize a super transition matrix with time-dependent self-transition probabilities which falls in the general category of variable time HMM (VTHMM) [10]. For particular states, transitions are modified depending on the presence of time-adjacent note onset. Let t' be the timestamp of the closest to given time t onset $\Delta n_{t'} = 1$. Now the transition probability can be rewritten as

$$a_{ij}(t) = \begin{cases} a_{ij} - g(t,t')q, & R1 \text{ or } R3 \\ a_{ij} + g(t,t')q, & R2 \text{ or } R4 \end{cases} \quad (1)$$

$R1$ to $R4$ stand for phoneme transition rules similar to these presented in [20]. They are derived from the phonemes network by picking the states i and j for two consecutive phonemes. The term q is a constant whereas $g(t,t')$ is a weighting factor sampled from a normal distribution with mean at t':

$$g(t,t') = \begin{cases} f(t;t',\sigma^2) \sim \mathcal{N}(t',\sigma^2), & |t-t'| \leq \sigma^2 \\ 0 & else \end{cases} \quad (2)$$

Since singing voice onsets are regions in time, they span over multiple consecutive frames.To reflect that fact, $g(t,t')$ serves to smooth in time the influence of the discrete detected Δn_t, where σ^2 has been selected to be 0.075 seconds. In this way an onset influences a region of 0.15 seconds - a threshold suggested for vocal onset detection evaluation in [7]. Furthermore, this allows to handle slight timestamp inaccuracies of the estimated note onsets.

3.4.1 Phoneme transition rules

Let V stands for vowel, C for consonant and L for vowel, liquid (LL, M, NN) or the semivowel Y. Rules $R1$ and $R2$ represent inter-syllable transition, e.g. phoneme i is followed by phoneme j from the following syllable:

$$\begin{array}{ll} R1: & i = V \quad j = \neg L \\ R2: & i = C \quad j = L \end{array} \quad (3)$$

For example in rule R2 if a syllable ends in a consonant, a note onset imposes with high probability that a switch to the following vowel or liquid is done. Rules $R3$ and $R4$ are for intra-syllabic phoneme patterns:

$$\begin{array}{ll} R3: & i = V \quad j = C \\ R4: & i = \neg L \quad j = V \end{array} \quad (4)$$

Essentially, if current phoneme is vocal and next is non-voiced (e.g. $R1$, $R3$), Eq. 1 discourages transition no next phoneme and encourages transition in the opposite cases. An example of $R4$ can be seen for the syllable KK-AA in Figure 2 where the note onset triggers the change to the vowel AA, opposed, for example, to onset at Y for the syllable Y-E-T. . Note that these rules assume that

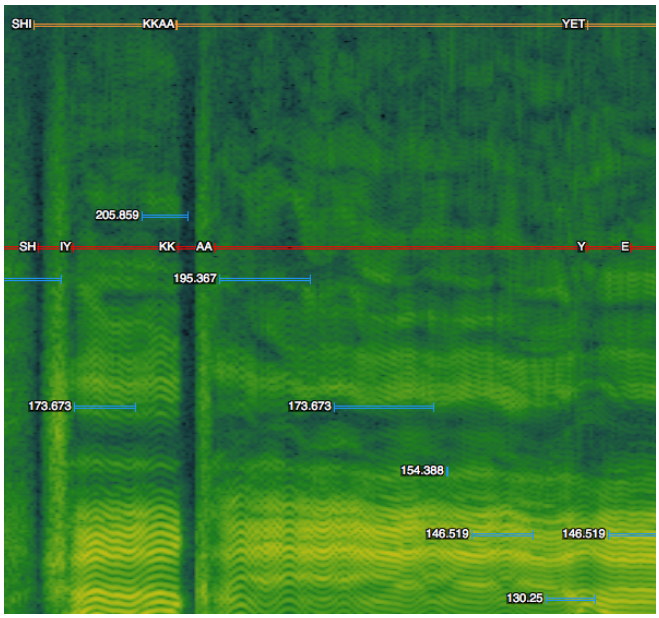

Figure 2. Ground truth annotation of syllables (in orange/top), phonemes (in red/middle) and notes (with blue/changing position). Audio excerpt corresponding to word şikayet with syllables SH-IY, KK-AA and Y-E-T.

a syllable has one vowel, which is the case for Turkish [1]. The optional silent phoneme *sp* is handled as a special case: transition probability from any phoneme to *sp* is derived according to intra-syllable rules, and the one from any phoneme skipping to the phoneme following *sp* according to inter-syllable rules.

3.4.2 Alignment

The most likely state sequence is found by means of a forced alignment Viterbi decoding.

[1] One-vowel syllabic languages include as well Japanese and to some extent Italian

total #sections	#phrases per section	#words per phrase
75	2 to 5	1 to 4

Table 1. phrase statistics about dataset

$$\delta_t(j) = \max_{i \in (j, j-1)} \delta_{t-1}(i)\, a_{ij}(t)\, b_j(O_t) \qquad (5)$$

Here $b_j(O_t)$ is the observation probability for state i for feature vector O_t and $\delta_t(j)$ is the probability for the path with highest probability ending in state j at time t (complying with the notation of [15, III. B]. [2]

4. DATASET

The test dataset consists of 12 a-cappella performances of 11 compositions with total duration of 19 minutes. The performances are drawn from anonymous corpus of classical Turkish Makam repertoire with provided annotations of musical sections [1]. Solo vocal versions of the originals have been sung by professional singers, especially recorded for this study, due to the lack of appropriate acapella material in this music tradition. 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. Another contribution of this work is that we make available the annotated phrase time boundaries [3] . A musical phrase (as proposed by [11]) spans 1 to 4 words depending on the duration of the words. Table 1 presents statistics about phrases, while the total number of words in the dataset is 732.

Additionally, the singing voice for 6 recordings from the dataset has been annotated with MIDI notes following the musical score [4] . On annotation special care is taken to set the note onset on the time instant, at which a steady melodic contour begins, avoiding setting it on a preceding unvoiced phoneme, which is important for rules R3 and R4 to make sense (see Figure 2).

4.1 Evaluation metric

Alignment is evaluated in terms of alignment accuracy as the percentage of duration of correctly aligned regions from total audio duration (see [6, figure 9] for an example). A value of 100 means perfect matching of all phrase boundaries in the evaluated audio. Accuracy can be reported not only for an audio segment, but as well on total for a recording, or on total for all recordings together.

[2] To encourage reproducibility of this research an efficient open-source implementation together with documentaion is available here. link suppresssed for anonymity

[3] Annotations and audio are available under CC license here. link suppressed for anonymity

[4] Creating the annotation is a time-consuming task, but we plan to annotate the whole dataset in the future

	cF	5	4.5	4.0	3.5	3.0
acapella	OR	62.2	65.7	73.8	75.3	78.2
	AA	79.1	82.3	85.5	**85.7**	82.0
polyphonic	OR	58.8	64.2	71.9	72.2	74.4
	AA	65.2	68.4	**72.8**	72.1	71.3

Table 2. VTHMM performance on acapella and polyphonic audio, depending on onset detection recall (OR). Alignment accuracy (AA) is reported on total for all recordings.

5. EXPERIMENTS

5.1 Experiment 1: alignment with oracle onsets

As a precursor to the following experiments, an alignment is run between lyrics and 6 of the recordings with an input of manually annotated notes as an oracle for note onsets. This is done to test the general feasibility of the proposed model on the dataset unbiased from errors in the note segmentation algorithm and to set a glass-ceiling alignment accuracy. We have tested with different values for q from Eq. 1 achieving best accuracy of 91.5% at $q = 0.23$.

5.2 Experiment 2: comparison to a baseline

As a baseline we conduct alignment with unaffected phoneme transition probabilities, e.g. setting all $\Delta n_t = 0$, which resulted in alignment accuracy of 75.2%. Further we measured the impact of the onset detector, varying onset detection recall by changing the minimum output of the Gaussian filter (parameter cF of the note segmentation method introduced in section 3.2). We adopt their threshold: an onset is considered as correctly detected if it is located within 0.15 seconds of a ground truth onset, as has been previously suggested in [7]. In table 2 is summarized alignment accuracy with VTHMM depending on recall. On acapella best improvement over the baseline is achieved at recall of 75% (at $cF = 3.5$). This is somewhat lower than the best recall of 81-84% achieved for flamenco [13]. Setting recall higher than that degraded performance because there are too many false alarms resulting in forcing false transitions.

Figure 3 allows a glance at the level of detected phonemes: the baseline HMM switches to the following phoneme after some similar for all phonemes amount of time. One reason for this might be that the waiting time in a state in HMMs with fixed transition matrix cannot be randomly long [22]. In contrast, for VTHMM the presence of note onsets at vowels activates rules R1 or R3, which allows waiting in the same state more time as there are more onsets (for example AA from the word SH-IY-KK-AA-Y-E-T has associated 5 onsets). We chose to modify cF because setting it to lower values increases the recall of *interval onsets*: Often in our dataset to a vowel, sustained relatively long, correspond several notes with different pitch. In fact characteristic for Turkish classical music is that a single syllable may have a complex melodic progression spanning many (up to 12 in our dataset) notes [4]. However, for cases of vowels held long on same pitch, con-

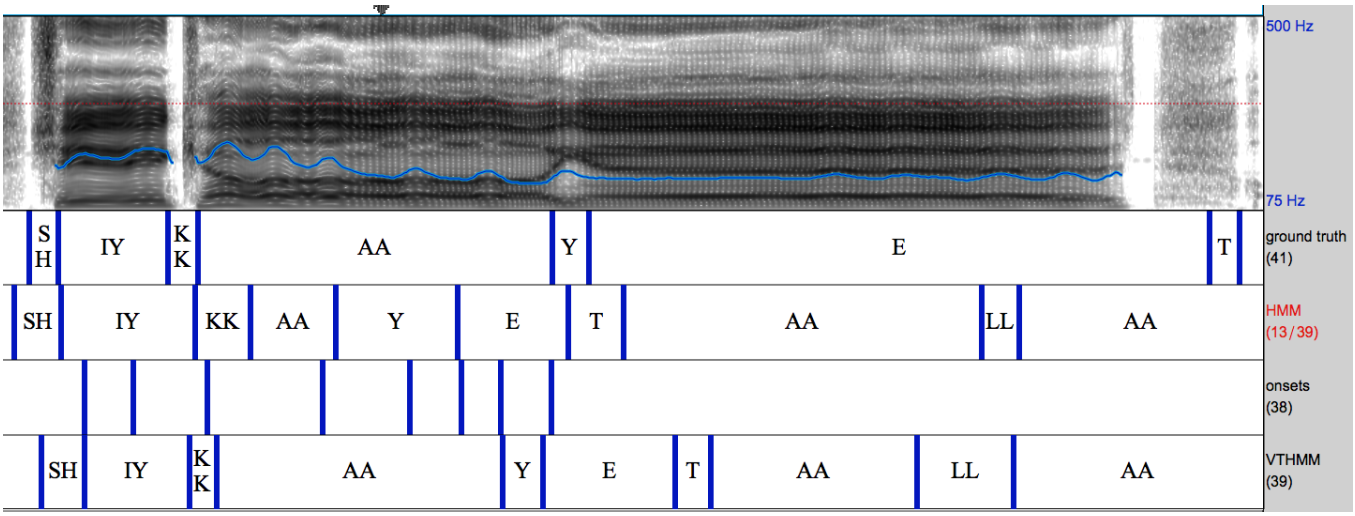

Figure 3. Example of boundaries of phonemes for the word şikayet (SH-IY-KK-AA-Y-E-T): *on top*: spectrum and pitch; *then from top to bottom*: ground truth boundaries, phonemes detected with HMM, detected onsets, phonemes detected with VTHMM; (excerpt from the recording 'Kimseye etmem şikayet' by Bekir Unluater).

ceptually VTHMM is not capable of bringing any benefit. This is illustrated as well in Figure 3 by the prematurely detected end boundary of E from the word SH-IY-KK-AA-Y-E-T.

Further, we examined alignment accuracy per recording (Figure 4). It can be observed that VTHMM performs consistently better than the baseline HMM (with some exceptions of where accuracy is close).

5.3 Experiment 3: recognition of phonemes

In general comparison to other lyrics alignment systems is hurdled because there is no hitherto work developed for Turkish language. However, to have an idea of how adequate are the trained phoneme HMMs we have annotated phoneme boundaries for some excerpts of total length of 6 minutes. In [9] phonemes are recognized in acapella singing with no lyrics given in advance. With phoneme MFCC-based HMMs - the same as our modeling setting - a phoneme recall rate of 44% is reported. Although in our case of forced alignment, recognizing phonemes is relatively easier, given that the phonemes are ordered in a sequence, we measure lower overall phoneme recall of 37%. This indicates that acoustic phoneme models trained only on speech might not be the most optimal choice.

6. EXTENSION TO POLYPHONIC MATERIAL

To test the feasibility of the proposed approach on polyphonic material, the alignment is evaluated on the original versions of the recordings in the dataset. Typical for Turkish Makam is that vocal and accompanying instruments follow a the same melodic contour in their corresponding registers with slight melodic variations. However, the vocal line has usually melodic predominance. This special type of polyphonic musical interaction is termed heterophony [4]. In the test dataset a singer is accompanied by one to several string instruments.

We applied the vocal pitch extraction and note segmentation methods directly, since both are methods developed for singing voice in a setting that has heterophonic characteristics. However, instrumental spectral peaks deteriorate significantly the shape of the vocal spectrum. To attenuate the negative influence of instrumental spectrum, a vocal resynthesis step is necessary.

6.1 Vocal resynthesis

For the regions with predominant vocal, based on the extracted melodic contours and a set of peaks in the original spectrum, the vocal content is resynthesized as separate audio using a harmonic model [19]. MFCCs are extracted from the resynthesized vocal part, because the harmonic partials preserve the overall spectral shape of the original singing voice [5]. More details and examples of the resynthesis step can be found in previous work that showed that the application of the harmonic model is suitable for aligning lyrics in Makam music [2]. A conceptually similar resynthesis step has been as well taken in current methods for alignment of lyrics in polyphonic Western pop music [6, 14].

6.2 Experiment 4: comparison of acapella and polyphonic

The onset recall rates after note segmentation are not much worse than acapella as presented in table 2. Although the degree of degradation in onset detection is slight, degradation in alignment accuracy is reasonable. This can be attributed most probably to the fact that our MFCC-based models are not very discriminative and get confused by artifacts, induced from other instruments on resynthesis. However, applying VTHMM still improves over the baseline (see table 3). Note that the margin in accuracy between

[5] In fact, resynthesis is not an obligatory step, but was performed in order to allow to track the intelligibility of different vocals after the application of the vocal detection

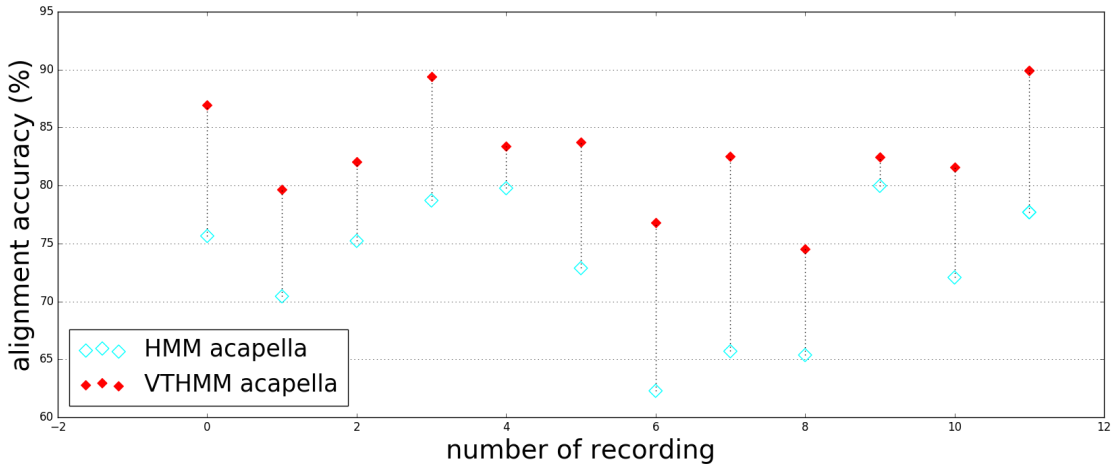

Figure 4. Comparison between results for VTHMM and baseline HMM on acapella. A connected line represents results for one recording.

the baseline and the oracle glass ceiling is only about 8% on polyphonic recordings, which is about twice wider in the case of solo voice.

	HMM	best VTHMM	oracle
acapella	75.2	85.7	91.5
polyphonic	67.2	72.8	74.9

Table 3. Comparison of accuracy of baseline HMM, VTHMM and, VTHMM with oracle onsets. Alignment accuracy is reported on total for all recordings.

7. CONCLUSION

In this work we evaluated the behavior of a HMM-based phonetic recognizer for lyrics-to-audio alignment in two settings: with and without considering singing voice onsets as additional cue. Compared to hitherto work on lyrics alignment, this is, to our knowledge, the first attempt to include information about an aspect of the melodic contour in the inference process. Updating transition probabilities according to onset-aware phoneme transition rules resulted in an improvement of absolute 10 percent for aligning phrases of solo voice from Turkish Makam recordings. In particular, due to rules discouraging premature transition, the decoding is allowed to stay adequate duration in sustained vowels.

Alignment on same data with instrumental accompaniment brought as well some small improvement over a baseline with no onset modeling. Having onset detection performing not substantially worse than acapella indicates that improving the phoneme acoustic models in the future could lead to even more reasonable improvement.

A limitation of the current alignment system is the prerequisite for manually-done structural segmentation, which we plan to automate in the future.

8. REFERENCES

[1] suppressed for anonymity.

[2] suppressed for anonymity.

[3] Hasan Sercan Atlı, Burak Uyar, Sertan Sentürk, Barış Bozkurt, and Xavier Serra. Audio feature extraction for exploring turkish makam music. In *3rd International Conference on Audio Technologies for Music and Media, Ankara, Turkey*, volume 12, page 2014, 2014.

[4] Eric Bernard Ederer. *The Theory and Praxis of Makam in Classical Turkish Music 1910–2010*. University of California, Santa Barbara, 2011.

[5] Hiromasa Fujihara and Masataka Goto. Lyrics-to-audio alignment and its application. *Dagstuhl Follow-Ups*, 3, 2012.

[6] Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1252–1261, 2011.

[7] Emilia Gómez and Jordi Bonada. Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Computer Music Journal*, 37(2):73–90, 2013.

[8] Rong Gong, Philippe Cuvillier, Nicolas Obin, and Arshia Cont. Real-time audio-to-score alignment of singing voice based on melody and lyric information. In *Interspeech*, 2015.

[9] Jens Kofod Hansen and IDMT Fraunhofer. Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. 2012.

[10] Michael T Johnson. Capacity and complexity of hmm duration modeling techniques. *Signal Processing Letters, IEEE*, 12(5):407–410, 2005.

[11] M. Kemal Karaosmanoğlu, Barış Bozkurt, Andre Holzapfel, and Nilgün Doğrusöz Dişiaçık. A symbolic dataset of Turkish makam music phrases. In *Fourth International Workshop on Folk Music Analysis (FMA2014)*, 2014.

[12] Willie Krige, Theo Herbst, and Thomas Niesler. Explicit transition modelling for automatic singing transcription. *Journal of New Music Research*, 37(4):311–324, 2008.

[13] Nadine Kroher and Emilia Gómez. Automatic transcription of flamenco singing from polyphonic music recordings. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(5):901–913, 2016.

[14] Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08*, 2008.

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

[16] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1759–1770, 2012.

[17] Justin Salamon, Emilia Gomez, Daniel PW Ellis, and Gael Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *Signal Processing Magazine, IEEE*, 31(2):118–134, 2014.

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

[19] Xavier Serra. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Technical report, 1989.

[20] Johan Sundberg. The kth synthesis of singing. *Advances in cognitive Psychology*, 2(2-3):131–143, 2006.

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

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