Sung Note Segmentation for a Query-by-Humming System

Pradeep Kumar, Manohar Joshi, Hariharan, S. Dutta-Roy, Preeti Rao

Department Of Electrical Engineering,
Indian Institute of Technology Bombay, India
{pradeepp, prao}@ee.iitb.ac.in

Abstract. Retrieval performance in query-by-humming (QBH) systems depends crucially on the accurate note segmentation and labeling of user queries. To facilitate note segmentation, querying is often restricted to the easily detected syllable “ta”, which is not necessarily the syllable most preferred by users. In this work, new acoustic features based on the signal energy distribution as obtained from the singing perception and production points of view are investigated. Performance evaluations on a manually labeled database of syllabic humming show that a specific mid-band energy combined with a biphasic detection function achieves high correct detection and low false alarm rates on the sonorant consonant syllables /da/, /la/ and /na/. The resulting onset detector is incorporated in the signal-processing front-end of an available QBH system (hitherto constrained to ta-syllable queries only). QBH retrieval performance results are reported on a large dataset of user queries.

Keywords: query-by-humming**,** syllabic humming, music transcription, note onset detection, note segmentation.

1 Introduction

Query-by-humming (QBH) is an important example of audio content retrieval by acoustic query. Users query the system by humming one or more phrases of the desired song, typically in a neutral syllable (e.g. “la”). The use of a neutral syllable facilitates the singing of the tune with typically one syllable per note. Important acoustic features of the query are those related to the tune or melodic pitch contour. The signal processing front-end of the QBH system transcribes the vocal query into a sequence of note pitches and durations (inter-onset intervals) to be matched against previously stored transcriptions of database melodies.

An important sub-task of query transcription is note onset detection. The performance of note onset detection influences the subsequent note labeling processes, directly impacting the retrieval performance of the QBH system. In fact, Prechelt and Typke [1] state that the most difficult problem is segmenting notes, and recommend that the user mark each note with a short break. To facilitate accurate note segmentation, several recent QBH system implementations have restricted the user query syllable to /ta/ or similar unvoiced phone [1], [2], [3], [4]. The articulation of the unvoiced plosive /ta/ involves a complete closure of the oral tract followed by a burst release before the onset of voicing for the succeeding vowel. The silence interval during the closure clearly demarcates notes in the signal amplitude envelope. While this reduces the burden on the signal processing front-end of the QBH system, surveys on user behavior [5] have shown that commonly preferred syllables include /la/, /na/ and /da/, which typically do not exhibit abrupt changes in signal power for continuous singing.

While note onset detection from musical audio signals has attracted much research interest recently, it turns out that no single method can address the entire variety of audio signals due to the intrinsically variable nature of the onset of sound events [6], [7]. In particular, not much work has been done explicitly on vocal onsets. In the present work, we adopt the approach of detecting changes in acoustic features related to syllable articulation. The hummed signal is a concatenation of sung consonants and vowels making up the sequence of syllable-notes with the consonants marking the note transitions and the vowel onsets aligning with the rhythmic beat. There has been much research directed at segmenting continuous speech into isolated syllables, typically as a by-product of statistical speech recognition. However knowledge-based approaches such as the use of acoustic-phonetic features may be expected to be adequate for the segmentation of syllabic humming apart from being far less computationally demanding. They could also be expected to be more robust across singers and styles. In any case, the differences between speech and singing, in terms of variety of articulation and intonation, warrant a separate study on the usefulness of these approaches to sung syllable segmentation. For reasons described earlier, we restrict our study to the syllables /da/, /la/ and /na/. Such sonorant consonants are acoustically most similar to vowels, among the non-vowel phonemes, and therefore among the most difficult to segment. Apart from the above “speech production” oriented viewpoint, we also consider a perception-oriented feature, namely the loudness.

Several recent articles propose tracking of signal energy in separate frequency bands, typically motivated by auditory filter banks. Band-level energy changes (or loudness changes) are then heuristically combined to detected note onsets [7], [8]. In the present work, this approach is adapted to the specific problem of vocal onsets based on the knowledge of syllable articulation features. Further, the note onset detection method is incorporated into the signal processing front-end of an existing (ta-query constrained) QBH system for Indian film music [2], and results on retrieval performance are presented.

It may be remarked that the use of pitch tracks has also been widely considered for note segmentation. Problems, however, include the observation that passing notes often do not stand out in the pitch track but rather only because of syllable articulation. Apart from this, there can be intended pitch modulation during the note (common, for example, in Indian music styles). The robust detection of note boundaries from alternate timbre (spectral) information would actually help to increase the accuracy of note labeling for melody representation by the proper averaging of pitch estimates obtained across the note duration.

The next section presents energy-based methods for syllabic note onset detection. Sec. 3 describes the experimental evaluation of the onset detection methods with respect to manually labeled onsets. The best performing onset detector is then incorporated in the transcription module of TANSEN, an available QBH system, and retrieval results are reported in Section 4. Finally, a discussion of the results is presented in Sec. 5 followed by the conclusions in Sec. 6.

2 Detection of Syllabic Note Onsets

The detection of syllabic note onsets involves finding a measure that reflects the acoustic signal change associated with a vowel onset, and reliably detecting and localizing rapid changes. Acoustic features are computed at fixed time intervals from the input signal, and the change occurring in the feature at a given time instant is measured by a detection function computed from a comparison of feature values in the neighborhood. Finally the locations of peaks in the detection function whose amplitudes cross a threshold are used to signal detected note onset instants.

Acoustic features could range from simple signal energy to the complete short-term signal spectrum. Motivated by auditory perception, changes in perceived loudness have been used to detect note onsets [8]. A different approach is to explicitly exploit the characteristics of syllabic humming and apply acoustic-phonetic knowledge related to speech production to its segmentation. Apart from finding a measure that captures signal characteristics at note onsets, it is important to find methods to reliably detect prominent changes and localize them accurately in time.

2.1 Acoustic Features

Since the syllables do not display obvious changes in the signal power envelope, simple energy based detection is not expected to work for onset detection. Perceived loudness, on the other hand, is influenced by signal energy as well as the signal spectrum. We compare the use of the loudness feature with an acoustic-phonetic knowledge motivated band energy to distinguish consonants from vowels. Both energy and loudness are derived from the short-time Fourier spectrum, $X_n[k]$, for the frame *n*. The feature sequence is extracted at 10 ms intervals from the short-term spectral analysis of the input signal using a 20 ms Hamming window.

Different loudness models have been applied for note onset detection. These are essentially equivalent to obtaining bark-band filter energies from the signal spectrum, and applying to the energies the equal loudness contour correction [9],[10]. Next the band-level specific loudness in sones is derived using the nonlinear conversion of Eq. (1)

$$L_n[i] = \begin{cases} \left(\dfrac{D_n[i]}{40}\right)^{2.642} & \text{if } D_n[i] < 40 \\ 2^{0.1\,(D_n[i]\,-\,40)} & \text{if } D_n[i] \geq 40 \end{cases} . \qquad \textbf{(1)}$$

$L_n[i]$ is the perceived loudness in the critical band i and $D_n[i]$ is the loudness in phons in band i for the frame n . For each band the loudness in phons is limited to be within the range 2-90 dB. The total loudness is obtained by summing the $L_n[i]$, the specific loudness in each band, as in Eq. (2). The total loudness is treated as the acoustic feature for note detection.

$$F[n] = \sum_{i=1}^{22} L_n[i] \ . \qquad (\mathbf{2})$$

An alternate approach is speech production based. Finding acoustic features to distinguish sonorant consonants from vowels has been an important research problem in speech science. Hermes [11] proposed vowel onset detection in speech by measuring increases in “vowel strength” in terms of the amplitude of spectral envelope peaks in the mid-frequency region (500, 2500 Hz) which typically spans the first two formants of the vowel. The abrupt vocal tract motion leading from the consonant into the vowel involves the rapid movement of formants leading to an abrupt increase in amplitude in regions of the short-term spectrum. In the work of Espy-Wilson [12], non-syllabic speech sounds (i.e. consonants) have been distinguished from syllabic (i.e. vowels) by the absence of significant energy in the frequency bands (640,2800 Hz) and (2000, 3000 Hz). The band energies are found to take on low values specifically for semi-vowels, nasals and voiced stops. Of course, although the difference in mid-frequency energy between the consonants and vowels is, on the average, much greater than the energy change within vowels, sometimes there is considerable overlap between their distributions. Motivated by the above acoustic-phonetic studies, the mid-frequency band energy is used as an acoustic feature to distinguish consonants from vowels. The sub-band energy is calculated from the STFT ($X_n[k]$) of the nth frame of data as given by

$$E[n] = \sum_{k=1}^{N/2} \left(\left| X_n[k] \right| W[k] \right)^2 \ . \qquad (\mathbf{3})$$

where W[k] is a band-limiting filter response with unity gain in the frequency region corresponding to (640, 2800 Hz) and falling off linearly to zero gain over a frequency region of 100 Hz on either side. N is the DFT size. The logarithmic energy is considered as the feature as given by

$$F[n] = 10\log_{10}\left(E[n]\right) \ . \qquad (\mathbf{4})$$

The use of the *log* ensures that relative changes are considered rather than the absolute changes in energy [8]. This normalization is consistent with hearing perception, where the perceived change in intensity is proportional to the intensity.

2.2 Detection Functions

Temporal changes in the feature sequence are estimated via a detection function (DF) based on the computed change or difference in the acoustic feature with time. Peaks (local maxima) of the DF would then occur during transitions at the instants of most rapid change. Ideally the DF should capture the consonant-vowel transitions while minimizing spurious peaks that may arise due to intra-vowel and intra-consonantal fluctuations in the features. Another desirable property is the accurate time-localization of the onsets. A DF that incorporates some smoothing before differencing will ensure that localized fluctuations in the feature occurring within phone regions are suppressed to an extent but also lead to the broadening and lowering of valid onset peaks. The functional form of the DF dictates trade-off achieved between the accuracy of detection of valid onsets and the suppression of false alarms. Widely used is the rectified first-difference function given by

$$\mathrm{DF}[n] = \begin{cases} 0 & , \quad \text{if } \left(F[n] - F[n-1]\right) < 0 \\ F[n] - F[n-1] \,, & \quad \text{else} \end{cases} . \qquad (\mathbf{5})$$

While the above detection functions consider only the pair of adjacent frames, multi-frame smoothing and differencing has the potential of being less susceptible to rapid local fluctuations in the feature that are not related to the larger note transition. A suitable function, motivated by the short-term adaptation characteristic of human hearing, is the biphasic filter [11]. Fig.1 shows a plot of the biphasic filter impulse response (comprised of Gaussian shaped components of different widths). In the present work, we also consider a time-scaled (by a factor of 2) version of the biphasic function to obtain the more localized "compressed biphasic" of Fig.1.

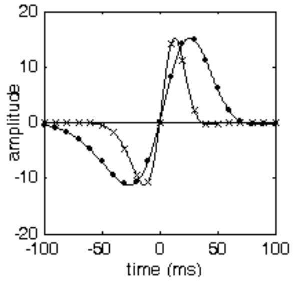

Fig. 1. Biphasic filter function with parameters from Eq. (4) of [11] shown with dots; with modified parameters leading to the compressed form, shown with crosses. The discrete points indicate the 10 ms frame-level values.

The selected discrete-time biphasic filter function is convolved with the feature sequence to obtain the detection function (DF), which is then searched for local maxima. A suitable thresholding of the peak amplitudes is then used to estimate the valid note onsets. Phrase boundaries give rise to invalid detections due to the silence to consonant transition, which are duly suppressed by post-processing based on silence detection.

In the next section, we evaluate the onset detection performance of the loudness and sub-band energy features in combination with each of the proposed detection functions, namely, the first difference and the biphasic filter functions.

3 Evaluation of Note Onset Detection

3.1 Data Collection and Preparation

Humming for each syllable type (/la/, /da/, and /na/) was recorded from six singers (of which only three had had formal musical training). There were 47 song segments in each syllable providing a total of 2513 notes. The humming was recorded with a good quality microphone and PC sound card at 22.05 kHz sampling frequency with 16-bit resolution. The songs were selected from popular Indian movie music [2]. The singers were given no specific instructions except to hum from memory the song phrases, whose lyrics were provided if needed, in each of the syllables /da/, /la/, and /na/.

The songs included a variety of melody and rhythm patterns. Each singer utilized at least one complete octave of his/her pitch range. Manual labeling of the onsets was carried out using the gating technique [11], by listening to segments of duration increasing in 5 ms steps until an onset is just perceived. A vowel onset is then marked at the center of the data frame containing the first major peak in the pitch cycle, immediately after the perceived onset. In total there were over 2500 valid onsets in the recorded data distributed equally across the syllables. Selected audio samples are available at [13].

3.2 Experiment

For a given combination of feature and detection function, the note onset detection performance was obtained by an examination of all the peaks of the DF as returned by the algorithm after classifying these into valid and invalid detections based on comparison with manually marked onset instants. A peak is treated as a valid onset if it is the strongest peak within +/- 20 ms (2 frames) of a manually marked note onset. Statistical distributions of the amplitudes of the valid onsets and invalid onsets (i.e. all remaining DF peaks) were obtained. Performance curves [6] in terms of percentage of true positives (i.e. correct onset detections relative to the total number of actual onsets) versus percentage of false positives (i.e. erroneous detections relative to the number of detected onsets) were traced out by varying the threshold in steps over a wide range.

Fig. 2. shows performance curves in the portion of the operating region of practical importance to a QBH system. A curve that is higher and more towards the left indicates superior performance in terms of the trade-off between hit rate and false alarms. For each curve, the optimal operating point (closest point to the top-left corner) is noted, and the results are displayed in Table 1. The full-band energy/first difference curve lies to the far right in the depicted region. We see that the loudness/first difference does significantly better. However the sub-band energy feature is significantly superior to either indicating that the selected mid-frequency band energy is indeed a good differentiator of vowels and sonorant consonants in sung syllables, consistent with its behavior in speech segmentation [12]. The biphasic filter function and its compressed form both effect a temporal smoothing of the feature which is expected to level out highly localized perturbations in the detection function as might occur in the intra-phone regions. This explains the superior performance of the biphasic filter based detection over the first difference in terms of reduced false positives for a given percentage of true positives.

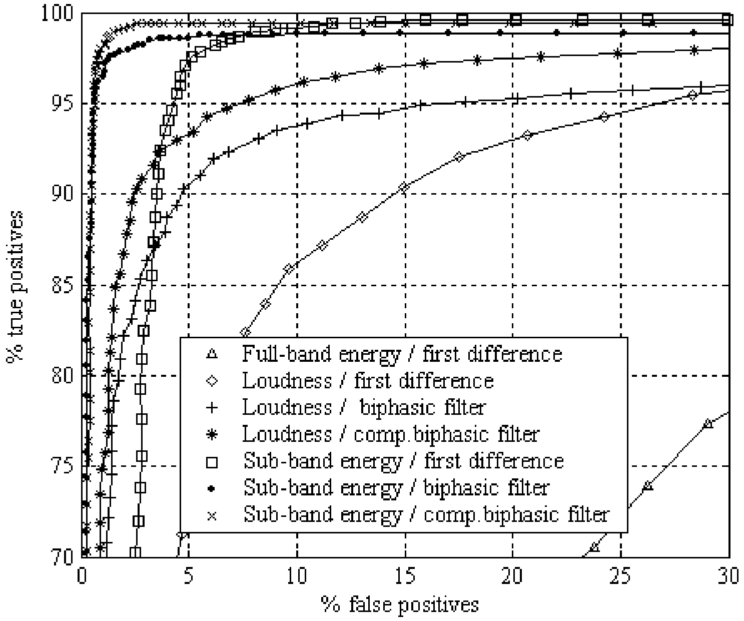

Fig.2. Performance comparison of onset detection methods for the data set comprising of 2513 note onsets.

The biphasic filter based method though falls short of achieving full detection at any threshold due to the flattening out of potential low peaks corresponding to genuine onsets. The better overall performance of the compressed biphasic filter over the unmodified form can be explained by the consonantal durations for /l/ and /n/ which average around 100 ms but can be as low as 20 ms in the fast (i.e. short duration) notes. The shorter negative lobe width of the compressed biphasic function better preserves the consonantal feature evolution in such cases.

Table 1 . Onset detection results. Columns show the percentage of true positives (TP%) and percentage of false positives (FP%) for each method obtained at its optimal fixed threshold (/da/ + /la/ + /na/ mixture: 2513 onsets)

Method	TP%	FP%
Full-band energy first difference	77.4	29.0
Loudness with first difference	87.2	11.6
Sub-band energy first difference	97.1	4.8
Sub-band energy biphasic filter	97.5	1.9
Loudness with biphasic filter	91.9	6.2
Sub-band energy compressed biphasic filter	98.9	1.5
Loudness with compressed biphasic filter	94.3	5.8

4 Retrieval Performance in QBH System

The evaluation results of the last section indicate that the sub-band energy combined with the compressed biphasic detection function achieves the best performance of the methods considered. Accordingly, this method was selected for note onset detection in the signal processing front-end of TANSEN, a QBH system, hitherto constrained to accepting only ta-syllabic queries [2]. In this section, we provide a brief overview of TANSEN followed by the evaluation of the modified system on a large database of user queries.

4.1 TANSEN System Description

TANSEN is a QBH system for Hindi and regional film music, a genre that enjoys a very wide appeal in India. Melody is the most distinctive element of such music, with the tune of the song's signature phrases being the most remembered attribute. TANSEN represents the melody by a note-pitch interval string, and then uses a string matching algorithm to retrieve the best matched database songs. The pitch interval representation is invariant to key transposition, and therefore suitable for melody representation [4].

A signal-processing front end locates note onsets and offsets and computes a pitch value in cents for each note thus extracted. Note onset detection is accomplished by the sub-band energy/biphasic detection function of Sec. 3. Note offsets are located between every two note onsets (and at the end of a phrase) based on the thresholding of negative peaks in the detection function. A note pitch is assigned based on the weighted averaging of pitch estimates computed by a standard correlation-based pitch detection algorithm at 10 ms intervals across the segmented note duration. The note pitch frequency values are then used to derive the sequence of pitch intervals between every two consecutive notes. Pitch intervals are quantised using a non-uniform quantisation scheme designed to obtain a good trade-off between robustness to user errors and retrieval accuracy. A normalized edit distance based string matching algorithm with a suitably defined cost function for insertion, deletion and substitution errors is used to retrieve a ranked list of matched songs from the melody database. Rhythm information in the form of note inter-onset durations is not utilized currently. At the present time, the TANSEN system database contains 300 songs (selected from popular Indian movie music) [13].

4.2 Retrieval Performance

A large data set of recorded user queries was generated. This data set included the hummed songs data described in Sec. 3.1 (47 song segments of 2-3 phrases each) in each of the syllables /la,na,da/ . The songs were selected from among the 300 songs of the reference database. In addition there were 1194 queries hummed using the syllable /ta/ collected from 10 users in total. The /ta/-query set covered all the 300 reference songs such that there were 3 or more queries per reference song. The humming was recorded with a good quality microphone and PC sound card at 22.05 kHz sampling frequency with 16-bit resolution. The singers were given no specific instructions except to hum from memory the song phrases, whose lyrics were provided if needed, in each of the syllables /da/, /la/, and /na/. The songs included a variety of melody and rhythm patterns.

A standard information retrieval performance measure, the Mean Reciprocal Rank (MRR) [14] was used to evaluate the performance of the QBH system with the new note segmentation module incorporated. The MRR of each individual query is the reciprocal of the rank at which the correct response is returned. The mean of the reciprocal ranks across the set of queries is defined as the MRR for the system as given below.

$$\mathrm{MRR} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{r_i}}}{\mathrm{N}} \quad . \tag{6}$$

where r_i denotes the i$^{\text{th}}$ position in the ranked list where the intended query occurs and N is the size of the query set. In case multiple items are retrieved with identical string match distances from the user query, an average rank is assigned to each of the songs. For example, if 3 queries happen to be at minimum edit distance from the reference, all three are assigned rank “2” and the next assigned rank in the list is then “4”. This ensures that the MRR captures the actual precision-recall characteristics of the system. Table 2 shows the MRR results obtained.

Table 2. QBH system MRR results across the different syllabic humming

Syllable	**# queries**	**MRR**	**% in top 10 ranks**	**%rank 1**
/ta/	1194	0.6044	75.5	50.6
/da/	47	0.6229	74.5	53.2
/la/	47	0.6447	80.9	55.3
/na/	47	0.6469	76.6	59.6
Overall	1335	0.6079	75.7	51.2

5 Discussion

We see from Table 2 that the intended query is ranked in top place more often than not. We also note that the performance of the system with the sonorant syllable queries (/la/, /da/ and /na/) compares well with that on the easy-to-segment /ta/ queries indicating that proposed note onset detection algorithm is robust.

Inaccuracies in QBH retrieval performance can arise from a number of possible sources. Imperfect singing on the part of the user can cause notes to be missed or inserted in the query. Note pitches may be off from their intended values due to out-of-tune singing or due to non-steady pitch over the duration of the note. Even with a perfectly rendered query, transcription errors could arise from errors in note segmentation. Missed detections of note onsets and false alarms lead to deletion and insertion errors in the note pitch interval string. Apart from this, note pitch values following a deleted note would be perturbed. From an analysis of the errors in retrieval, it was observed that over 90% of the poorly ranked queries were indeed rendered inaccurately by the user in terms of note deletion/insertion or gross mismatches in note pitch. Of the remaining poorly ranked queries, most were characterized by transcription errors arising from improper note segmentation. We next discuss the note onset detection errors as obtained from the experiments of Sec. 3.

Instances of missed detections as well as the false positives obtained at the optimal operating point (Table 2) were examined for the sub-band energy-compressed biphasic filter method. Missed detections were found to be more likely to occur when the signal intensity was low. Soft singing, with its reduced vocal effort, leads to a disproportionate drop in the mid-band energy of the vowel. Another source of missed detections is the short duration notes arising from tempo/rhythm constraints. The singer stops short of fully articulating the vowel leading to a drop in the difference of mid-band energy at the consonant-vowel boundary. Further, when the consonant (/l/ or /n/) is articulated slowly and clearly, as might happen when singing at a very slow tempo, the feature discontinuity at the vowel onset is less sharp. This is similar to the case of slowly uttered speech in which the tongue release after apical contact is slow [12]. On the other hand, syllables that are stressed, due to coincidence with an accented note in the musical score, show the sharpest spectral discontinuities at the vowel onset. This may be attributed to the rapid formant transitions linked with the firm apical contact and subsequent rapid release of the tongue tip. False positives arise from rapid spectral changes occurring within vowel or consonant segments (intra-phone regions). These are observed to be linked with intonation variations in the singing such as loudness increases or strong pitch modulations, which tend to cause fluctuations in overall intensity, affecting also the mid-frequency region. Across the syllables, false positives were more likely to arise from /da/ due to the larger intra-consonantal variations in sub-band energy for /d/. This is explained by the rapidly changing energy levels associated with the transition through the stop closure of the phone /d/.

6 Conclusion

We have considered the problem of note onset detection in the context of user query transcription for QBH systems where the queries are typically rendered using neutral syllables. The detection of syllabic note onsets involves finding a measure that reflects the acoustic signal change associated with the consonant-vowel transition, and reliably detecting and localizing rapid changes. To facilitate accurate note segmentation, several recent QBH systems restrict the user queries to be sung in the syllable “ta”, the acoustic characteristics of which clearly demarcate note boundaries by means of a steep fall in signal energy. In this work, acoustic features based on the signal energy distribution as obtained from the perception and production points of view were considered. Performance evaluations on a manually labeled database of syllabic humming show that a particular mid-band energy combined with a biphasic detection function achieves high correct detection and low false alarm rates on the sonorant consonant syllables /da/, /la/ and /na/. The resulting onset detector is incorporated in the signal-processing front-end of an available QBH system (hitherto constrained to ta-syllable queries only). QBH retrieval performance results on a large dataset of user queries confirm the efficiency of the note segmentation algorithm.

Acknowledgments. The contribution by our volunteer singers to this work is gratefully acknowledged.

References

1. Prechelt, L., Typke, R.: An Interface for Melody Input: ACM Trans. on Computer-Human Interaction. Vol. 8 (2). 133-149 (2001)
2. Raju, M.A., Sundaram, B., Rao, P. : TANSEN: A Query-by-Humming Based Music Retrieval System. Proc. of the National Conference on Communications (NCC), Chennai, India (2003)
3. Kosugi, N., Sakurai, Y., Morimoto, M. :Sound Compass: A Practical Query-by- Humming System. Proc. SIGMOD Conference ’04. 881-886 (2004)
4. Pauws, S. : CubyHum: A Fully Operational Query- by- Humming System. Proc. Intl. Symp. on Music Information Retrieval (ISMIR 02). (2002)
5. Lessafre, M., Moelants, D., Leman, M., De Baets, B., De Meyer, H., Martens, G., Martens, J. P.: User Behavior in the Spontaneous Reproduction of Musical Pieces by Vocal Query. Proc. of the 5th Triennial ESCOM Conf., Hanover (2003)
6. Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M. B. : A Tutorial on Onset Detection in Music Signals. IEEE Trans. Speech and Audio Processing. Vol. 13(5), 1035-1047 (2005)
7. Alonso, M. Richard, G. David, B. : Extracting note onsets from musical recordings. Proc. IEEE-ICME (2005).
8. Klapuri, A. : Sound Onset Detection by Applying Psychoacoustics Knowledge. Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing. 3089-3092 (1999)
9. Timoney, J., Lysaght, T., Schoenwiesner, M. : Implementing Loudness Models in Matlab. Proc. of the 7th Intl. Conf. on Digital Audio Effects (2004).

10. Collins, N. : A Comparison of Sound Onset Detection Algorithms with Emphasis on Psycho Acoustically Motivated Detection Functions. Proc. of Audio Engineering Society Convention.(2005).
11. Hermes, D. J. :Vowel Onset Detection. J. Acoust. Soc. of Am.87(2), 866-873 (1990)
12. Espy-Wilson, C. Y. : Acoustic Measure for Linguistic Features Distinguishing the Semivowel /wjrl/ in American English. J. Acoust. Soc. of Am.92(2) 736 – 757 (1992)
13. Music Transcription for TANSEN QBH System. Digital Audio Processing Lab, IIT, Bombay. India
    http://www.ee.iitb.ac.in/uma/~daplab/QBHTranscription/index.htm
14. Dannenberg, R.B., Ning Hu. : Understanding Search Performance in Query-by-Humming Systems. Proc. of the Int. Symp. on Music Information Retrieval (2004)