Note Onset Detection in Natural Humming

Pradeep Kumar P
IIT, Bombay, India
pradeepp@ee.iitb.ac.in

Preeti Rao
IIT, Bombay, India
prao@ee.iitb.ac.in

Sumantra Dutta Roy
IIT,Delhi, India
sumantra@cse.iitd.ac.in

Abstract

Many state-of-the-art query-by-humming (QBH) systems restrict the hummed query to be in isolated syllables for easy note segmentation. However, it is observed that users often prefer natural humming. This work addresses note onset detection for natural hum, which is considered difficult to segment. The acoustics characteristics of naturally hummed signals are studied and features useful to note onset detection are proposed. Pitch and energy features are combined to obtain superior note segmentation. Performance results on note onset detection as well as retrieval in an actual QBH system are reported.

1. Introduction

Music information retrieval (MIR), currently an important research area, enables users to search a music database with a small snippet of recorded audio. Query by humming (QBH) is type of MIR system wherein a user hums the query to the system to retrieve melodically similar music [1][2]. The required melodic representation of the query is achieved by transcription based on the accurate detection of note onsets.

The human perception of sound depends on its loudness, pitch and timbre. The perception of a note onset is caused by a noticeable change in the loudness, pitch or timbre of the sound [3]. For note onset detection, features that exhibit significant changes at the onset are preferred. The onsets are differentiated as soft and hard onset, hard onset exhibit abrupt feature change, while soft onset takes time to establish [4][5].

Most commonly used loudness feature is the energy. This is effective when the note onset is characterized by transient or burst followed by steady state sound. But effectiveness decreases with signals having soft onsets. Some of the features related to the timbre are spectral centroid and spectral difference[6]. A sub-band based approach was reported by Duxbury et al.[4] wherein energy is used in the higher sub-bands and spectral difference is used in the lower sub-band to find both hard and soft onsets in instrumental music. Adams [7] used only pitch as a feature for note onset detection in sung query of lyrics. Three classes of segmentation algorithms were explored: predictive filter, LMS detection and curve fitting assuming that ideal pitch contour is piecewise constant. The work is based on the philosophy that segmentation should rely only on pitch information. The main drawback with pitch as a feature was that it fails to capture passing notes, which can be captured by other features. Klapanc and Prefer [5] proposed a hierarchical onset detection approach, which detects the soft onsets efficiently along with hard onsets in sung lyrics.

In vocal rendition of the music, typically each syllable in the lyrics represents a note. The nucleus of the note is the vowel and the consonant prior to it indicates the note boundary. Most of the state-of-the-art QBH systems restrict the hummed query to be in isolated syllables

and prefers /ta/ or /da/ humming for good note segmentation [1][2]. In a study on the most preferred and used type of syllabic humming [8], it was found that /la/, /da/, /na/ and /ta/ are the most preferred humming apart from natural humming, denoted by /hm/. We have reported our work on the syllabic humming of /la/, /na/ and /da/ [11]. Using the acoustic-phonetic approach of speech recognition [9], it was found that the sub-band (640-2800 Hz) gives a sharp energy differences between the sonorant consonants, i.e /l/, /n/ or /d/ and the vowel /a/. This sub-band energy feature performed better than the generally used full-band energy, spectral difference and loudness for the syllabic humming [11]. The note onset detection performance of the sub-band energy for natural humming /hm/, however, was far below that for syllabic humming. In this work, we analyze the pitfalls with the sub-band energy feature for natural humming and try to improve the note onset detection performance with heuristics based on the acoustic characteristics of the natural hum.

2. Note onset detection

Note onset detection comprises of a feature extraction stage followed by a detection function stage based on the time derivative of the feature contour. The detection function exhibits peaks at the abrupt changes or onsets. Any noisy transients in the feature will lead to false peaks. The detection function is compared with a threshold to remove the noisy peaks. A good detection function should have few false peaks and capture all valid onsets. In our earlier work [11] we used a biphasic filter to calculate the detection function, which was motivated by short-term adaptation characteristics of human hearing. Rather than comparing only adjacent frames, the biphasic filter achieves multi-frame smoothing and differencing thus reducing false onsets.

3. Signal characteristics of natural hum /hm/

Perceptually, the presence of two phonemes in the syllabic humming gives clear note onset detection, since both have different timbre (spectra). Natural hum /hm/ is a nasalised humming, with notes sung in phoneme /m/ only. The perception of note onset is due to some kind of discontinuity in the signal, either in fundamental frequency or in the spectrum. It is observed that there are different ways the /hm/ can be articulated.

- During note transition, the singer tries to break the current note leading to a discontinuity in the partials, which reappear with the onset of the next note. But the duration will not be enough for vocal cord to stop vibration completely before the next note begins, leading to a dip in the fundamental frequency between two notes.
- Fundamental frequency glides smoothly towards the next note, without any frequency-dip at the boundary. In this mode all partials are continuous with slight dip in partial energies.
- In another mode, the singer introduces aspiration noise /h/ at the note boundary.

Based on the articulation, different types of discontinuities are observed at the note onset, as shown in Fig 1:

- Clear discontinuity (CD): where there is a clear break in the fundamental frequency contour, this happens when the notes are well separated or aspiration is introduced.
- Tonal discontinuity (TD): where all the harmonics are continuous during the transition between two notes.

- Partial discontinuity (PD): where all the harmonics except for the first one or two are absent during the transition from one note to another. The fundamental frequency will be present throughout the transition.
- Mixed discontinuity (MD): It is similar to partial discontinuity, but not all higher partials are absent during the transition.

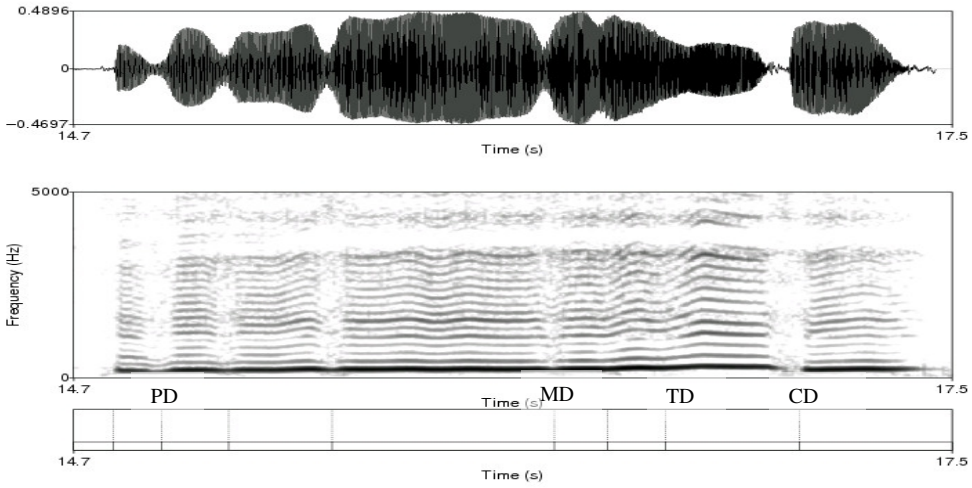

Figure.1. Different types of discontinuities in natural hum

In speech, for the nasal sound /m/ the first formant near 250 Hz dominates the spectrum [10]. A spectral zero occurs in the 750-1250 Hz for /m/. In a well articulated natural hum, at discontinuities, the energy above 700 Hz is very low compared to that below 700 Hz. For low energy hum, during a note, almost all energy is in the first formant region and higher band is highly noisy. As the energy increases, the higher bands are more energized and become less noisy. The anti formant region also have significant energy during the note compared to the discontinuity region. For some singers the frequency region above 3000Hz is noisy, depending on how they energize. Hence the same sub-band frequency region 640-2800 Hz is considered useful for natural hum.

4. Data collection and experimental setup

Natural hums were recorded from 10 singers and split into training and testing databases containing (47 and 50) hums with (831 and 1082) onsets. The recording was done at 22.05 kHz sampling rate, 16 bit resolution, mono. The songs were selected from Bollywood music and have a variety of melody and rhythm patterns. Another small database called modes database was also collected, which have hums with 204 onsets. This was recorded from two trained singers humming in different loudness, tempo and fundamental frequency variation. The ground truth onsets were marked by the first author manually.

Initially, to find which energy feature give the best performance different energy features were experimented, such as full-band energy, sub-band energy, harmonic energy and loudness on one database of 831 onsets. The features were calculated with Hamming window of 20 msec duration, every 10 msec. To evaluate the performance of each feature, the performance evaluation plot (% of true positive versus % of false positives) was obtained [6][11]. The

optimum point on the performance curve is used for comparison. Sub-band energy gave the best performance (87.25%, 14.81%), which was poor compared to that achieved with syllabic humming. At zero thresholds, sub-band energy captured 829 onsets out of the 831 with 2678 false onsets. Due to the different style of singing and how the singer energises this band, the energy variation in this band does not exhibit sharp changes at all note onset. At some onsets the changes are slight and similar to noisy variation.

To improve the performance, the usefulness of other features was considered and wherever it was valid, heuristics were applied. The approach was to validate the onsets detected by sub-band energy using additional features and to remove the false onsets. It was observed that the histograms of the magnitude of valid and invalid peaks exhibited an overlap for threshold values 8 to 200. For the onsets within this range (8-200) we apply the heuristics and validate the onsets and remove possible false onsets.

4.1. Combining features to improve the note onset detection

Usefulness of full-band energy: For some singers it was observed that the sub-band energy region was not energized enough to have large discontinuities during parts of the hum. The full band energy was a found to be a better feature in such cases.

Degree of voicing: Some singers insert aspiration at the note boundary. The aspiration energy was comparable to the energy in sonorant part in both full band and sub-band energy resulting in less discontinuity around the note onset. To capture the aspiration noise, wherever the degree of voicing was less than 0.55, a silence was inserted.

Three point median filtering: The sub-band energy becomes noisy for some singers towards the end of the phrase. To remove these false onsets, a three point median filtering was done on the feature contour.

Dip in fundamental frequency near the note onset: The dip in fundamental frequency that occurs at the note boundary appears to be an uncontrolled action. A dip of more than 8 Hz within 5 frames duration appeared to be a valid dip at any frequency value.

4.2. Algorithm with heuristics for combining features

Three-point median filtering was applied to the sub-band energy. The feature was ANDed with the silence marker, which also mark silence when degree of voicing is less than 0.55. The detected onsets obtained from sub-band energy are used as the candidate onsets. The heuristics are applied to detected onsets within the range 8–200. In case the heuristics applied are valid for a detected onset, its peak value is made 300, i.e. it is a valid onset. The onsets that are found false by heuristics are removed. The other candidate onsets for which heuristics were not valid are retained with same peak values.

Full-band energy: If a candidate onset occurs, look for a valley in the full-band energy within 100 ms duration, before onset. If full-band energy valley occurs, look for full band energy difference between the valley point and two points adjacent to the valley, 50 ms on either side. If it is more than 7 dB, it is a valid onset.

Pitch heuristics near the note onset: Near the candidate onset check for a fundamental frequency valley with a frequency difference of minimum 8 Hz, within the +/-50 ms on either side of the valley. If the condition is satisfied, the candidate onset is valid. In case of tonal onset, the pitch is important feature, with a very slight dip in the sub-band energy contour.

The fundamental frequency difference at points +/-50 ms apart from the onset candidate is considered, if this difference is more than 120 cents, the onset is valid.

Removing multiple onsets: The biphasic filtered detection function signal exhibit positive excursion near the note onset and negative excursion at note offset. Between two note onsets there should be an offset, which means between two positive excursions there should be a negative excursion or at least a region of zero values in the un-rectified detection function for some duration, which is set to 50 ms. The positive excursion also has a minimum duration before it goes to negative region, which is set to 50 ms. If this condition is not satisfied it is a false onset. To remove multiple onsets near the actual onset, look for negative excursion of detection function in between the two onsets. If there is no negative excursion of minimum duration 50 ms, then one of the onset is invalid, retain the higher valued candidate onset.

5. Results

By applying the heuristics, the false onsets were reduced to 425 from 2678, with true onsets being 826 compared to 828. With the modified peak detection, the performance curve optimum point was (93.5%, 6.5%) compared to the previous (87.25%, 14.81%). To validate the algorithm, it was tried on other databases, test and modes database. The modes database performance curve optimum was (90.7%, 9.7%) while with test database, it was (94.5%, 4.5%). Modes database has songs sung in very fast tempo and low loudness, which resulted in slightly poor performance. Finally the 97 natural hums were used as query to the TANSEN QBH system [2][11] with the new transcription module incorporated. The TANSEN system has 300 database songs. The new system gave Mean Reciprocal Rank (MRR) of 0.4060. Out of 97 queries 55 queries came within first 10 ranks with 30 being at first rank and there were 9 outliers (ranks above 100).

6. Conclusion

There is no single feature, which can give good onset detection for all kind of musical signals. A judicious combination of different features needs to be used based on the knowledge of the characteristics of the signal. Based on this philosophy an algorithm for improving the onset detection in natural hum was designed with the knowledge of the variation of different features. The algorithm was validated on different databases and gives good note onset detection performance.

7. References

[1] A. Ghias et al., “Query by humming : musical information retrieval system in an audio database”, ACM Intl. Conf. on multimedia , 1995, pp 231-236.

[2] M. A. Raju, B. Sundaram and P. Rao, “ TANSEN: A query- by-humming based music retrieval system”, Proc. of the National Conference on Communications, 2003, I.I.T. Madras

[3] A. Klapuri, “Sound onset detection by applying psychoacoustics knowledge”, Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Arizona,1999, pp. 3089-3092.

[4] C.Duxbury, M. Sandler and M. Davies, “A hybrid approach to musical note onset detection”, Proc. of 5th Intl. Conf. on Digital Audio Effect, Hamburg, Germany, Sept 2002.

[5] E. Klapanc and Prefer, “Hierarchical note onset detection”, Proc. Int. Computer Music Conf., 2004.

[6] N. Collins, ”A comparison of sound onset detection algorithms with emphasis on psycho acoustically motivated detection functions”, Proc. of Audio Engineering Society Convention, May 2005.

[7] N.Adams, “Automatic segmentation of sung melodies”, http://www.eecs.umich.edu/ systems/ TechReportList.html

[8] M. Lessafre et al., “User behaviour in the spontaneous reproduction of musical pieces by vocal query”, Proc. of the 5th triennial ESCOM Conf., Hanover, 2003, pp. 208-211.

[9] C. Y. E . Wilson, “Acoustic measure for linguistic features distinguishing the semivowel /wjrl/ in American English”, JASA, Aug 1992, pp 736 – 757.

[10] D O’Shaughnessy, “Speech Communications : Human and Machine”, Universities Press (India) Limited

[11] P. Kumar, M. Joshi, S. Hariharan, S. Dutta-Roy and P. Rao, "Sung note segmentation for a query-by-humming system", *Proc. of Music-AI (International Workshop on Artificial Intelligence and Music) in IJCAI*, 2007, Hyderabad, India.