AN ONSET DETECTION ALGORITHM FOR QUERY BY HUMMING (QBH) APPLICATIONS USING PSYCHOACOUSTIC KNOWLEDGE

Balaji Thoshkahna and K.R.Ramakrishnan

Music and Audio Group,Learning systems and Multimedia Labs(LSML)
Department of electrical Engg,Indian Institute of Science(IISc)
Bangalore-560012,India
phone: + (91) 080-22932696,
email: balajitn,krr@ee.iisc.ernet.in

ABSTRACT

We propose a new algorithm for onset detection in hummed queries for QBH applications. The algorithm uses a modified version of a popular loudness model for human hearing to identify onsets in hums. We also propose the use of a local minimum function to identify onsets better. A sub-band based sone scale processing that is advantageous for a simple implementation is used. On an annotated database of syllabic and natural hums, the algorithm identifies onsets correctly on an average 90% of the time with only 6% false positives. The features used in this algorithm can be used in conjunction with other feature / decision fusion based onset detection systems.

1. INTRODUCTION

Onset detection in free singing or humming forms a very important stage in a query by humming (QBH) system[1]. Accurate note onset detection improves the melody transcription stage and hence leads to better retrieval even with simple string matching techniques. Note onset detection in hummed queries is a challenge since human voice has a dynamic nature and methods applied for instrumental audio do not seem to be satisfactory for human hums[2]. Prechelt et al [3] observe that note segmentation forms the toughest part of the transcription algorithm and suggest that notes should be separated with silences / breaks in humming. Since humans untrained in music have tremors in their hums leading to huge variations in their note frequencies, robust note onset detection algorithms for QBH systems have to either consider such tremors (which may, at times, be interpreted as soft onsets) or force users to hum using certain restricted syllables like /ta/ [1]. Lessafre et al [4] in a survey show that users of QBH systems sung using syllables /na/, /la/, /ta/ and /da/ in the decreasing order of preference with natural humming (using the syllable /hm/) being preferred by a very small percentage of users. The above study suggests that researchers must concentrate on solving the note onset problem in sung syllables like the ones suggested above.

Toh et al[2] used multifeature fusion based learning algorithm to model features from onset and non-onset sounds as GMMs. Kumar et al[5] have used a fused detection function based on loudness, sub-band energy, full band energy and their derivatives using biphasic filters on sung syllables /da/, /na/ and /la/. Kumar et al[6] further modified the algorithm in [5] with improved heuristics to include even the natural hum using the syllable /hm/. Both the above mentioned algorithms involve a significant learning portion for choosing thresholds at various levels and also heuristics to optimise their performances.

We present here an algorithm that relies on a psychoacoustic model of loudness and a non-linear smoothing of the sub-band loudness function to enable onset detection in syllabic and natural humming. We also motivate the correlation of our thresholds to perceptual aspects of input audio that enable us to manipulate them easily.

2. LOUDNESS MODEL BASED ONSET DETECTION ALGORITHM

Onset detection for polyphonic audio using psychoacoustics was first proposed by Klapuri[7]. Thoshkahna et al[8] proposed an improved algorithm by using a different representational system of the audio input to the loudness model.

Onset detection in hummed queries poses a challenge due to singer / hum imperfections which lead to spurious onsets [2]. Humans have an innate ability to detect only real onsets and ignore trmeors and modulations while humming. To simulate the same performance, we propose a modification to the sub-band partial loudness function followed by a decision fusion across sub-bands to enable robust onset detection as in [8]. The onset detection algorithm is shown in Fig.1. The algorithm is explained in detail below.

2.1 Normalization of hum audio

This first step takes care of various recording and sampling conditions. All audio are resampled to 8kHz and their RMS(Root mean squared) SPL(sound pressure level) scaled to 70dB to simulate a comfortable hearing level among humans.

2.2 ERB (Equivalent Rectangular Bandwidth) Filterbank

We follow a frame based processing to allow for the dynamic nature of hum signals. The normalized audio is split into frames of 30ms with an overlap of 20ms to ensure a smooth variation in signal characteristics. This signal is passed through an ERB filterbank stretching from 50Hz to 4kHz. There are 51 uniform 0.5 ERB apart filters in the ERB scale[9] in the frequency range of interest. Signal rectification and energy integration within the 30ms window is performed to simulate the workings of the inner ear. Each frame of audio now has 51 excitation energy features that are fed to the range adaptation block that simulates a time localised dynamic range adaptation.

© EURASIP, 2009

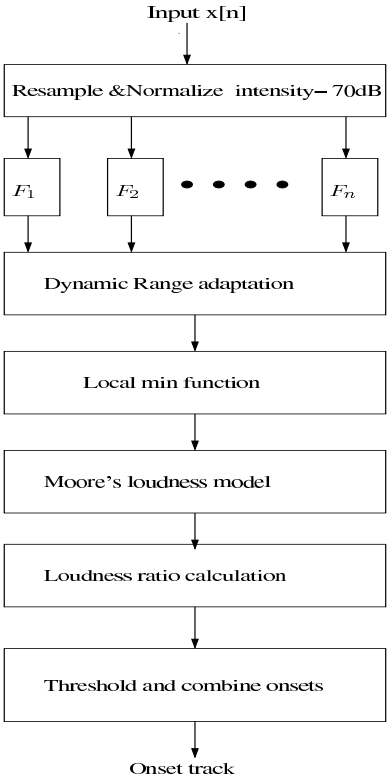

Figure 1: Onset detector

2.3 Dynamic range adaptation

Over a long time window of say 5 secs, we perform a dynamic range adaptation. Within each 5 second window we have 5000 frames of audio. Each frame of audio has 51 features. We call each of these features a T-F (time-frequency) feature. To simulate the dynamic range adaptation, we choose the T-F feature that has the maximum energy over a 5s window. We retain only those T-F features that are within 35dB of this maximum and neglect the rest. This enables us to neglect puffs of breath in the input hum that can be present before actually singing a new syllable (this can create spurious onsets) .

Furthermore, for each frame in this 5s window, we choose a maximum T-F feature and retain only those T-F features that are within 25dB of this maximum and neglect the rest of the T-F features. This step has the effect of neglecting low energy sub-bands from contributing to the actual onset detection process.

This dynamic range adaptation results in around 7% improvement in onset detection for monophonic and polyphonic audio [10] and hence this step is retained for hum audio even though the ear does not display such a phenomenon that we know of.

2.4 Min function processing

This step performs a non-linear smoothing of the audio signal excitation energy to block out potential tremors in singing and background recording noise from influencing the onset decision process. This minimum function smoothing is performed by choosing a small window of 2 frames around the frame of interest and replacing the excitation energy in the

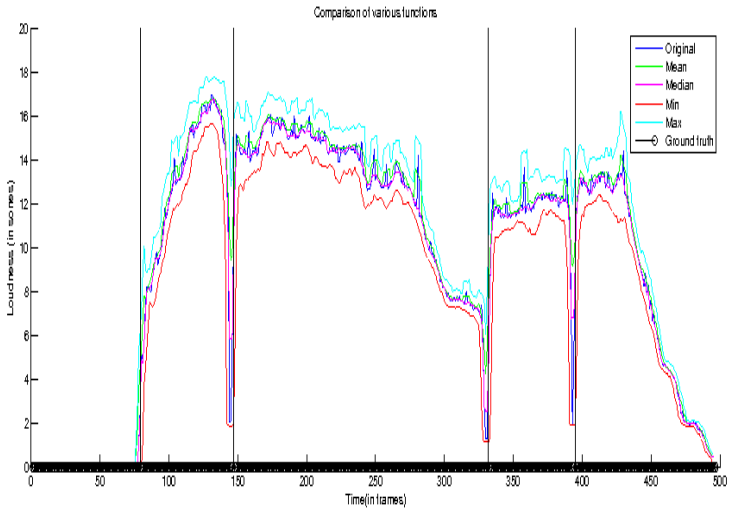

Figure 2: Comparison of various functions of excitation energy (after passing through the loudness model)

frame by the minimum in the whole of the window. The minimum function smoothed excitation energy function $E_{sig_{min}}$ is given by:

$$E_{sig_{min}}(i,j) = \min_{k=j-2}^{k=j+2} E_{sig}(i,k) \quad (1)$$

where $E_{sig}(i,k)$ is the excitation energy for the i^{th} sub-band at the k^{th} frame. Similarly the excitation energy was processed with various other functions such as the local maximum, mean and median. As shown in Fig.2, the minimum function preserves the onsets best while suppressing spurious noise that can be caused either due to recording conditions or tremors in the singer's voice. The min-function processed excitation energy is fed to Moore's loudness model to calculate the loudness in sones.

2.5 Moore's loudness model

Moore's model of loudness has been one of the popular models to explain human perception of loudness[9] and can be easily implemented. We use the same model of implementation as proposed by Timoney et al [11]. For each frame, sub-band excitation energies (T-F features) are fed to the loudness model to be compared to the threshold of hearing at the corresponding sub-band center frequencies. Only sub-bands with the excitation energy greater than the threshold of hearing contribute to the partial loudness (measured in sones). The partial loudness in sub-band i for the k^{th} frame, $L_i(k)$ is given by:

$$L_i(k) = C.(E_{sig_{min}}(i,k)^{\alpha} - E_{th}(i)^{\alpha}) \quad (2)$$

where $E_{sig_{min}}(i,k)$ is the smoothed excitation energy of the k^{th} frame in the i^{th} sub-band and $E_{th}(i)$ is the excitation due to the threshold of hearing at the i^{th} sub-band. We get the $E_{th}(i)$ by passing pure sinusoids (of rms MAF (Minimum Audible Field) values at the filter centers) through the ERB filterbank. The constant α does the audibility range compression that occurs in the human auditory system and has a value of 0.24 and the constant C is used to calibrate the model and

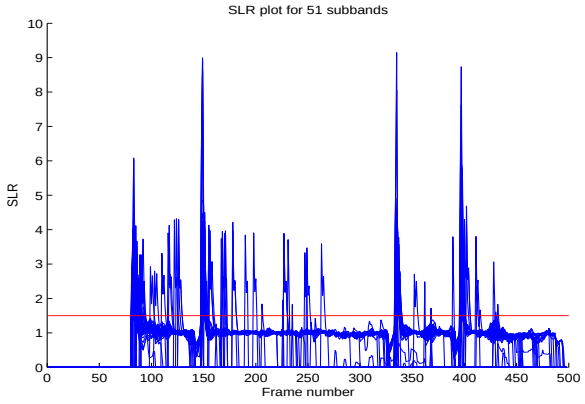

Figure 3: SLR functions for all 51 sub-bands.The peaks show significant subband loudness changes

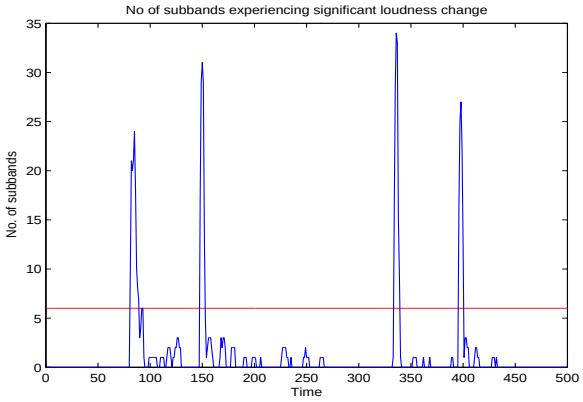

Figure 4: Plot of the Number of sub-bands experiencing a significant SLR vs Time

has a value of 0.04. The total loudness is the weighted sum of sub-band loudness for a frame with the weight being the ERB distance between the sub-band filters (0.5 in our case).

2.6 Loudness change detection function

Since the sone scale is used, loudness change across frames is calculated by taking the ratio of partial loudness in the i^{th} frame to the $(i-1)th$ frame. A ratio of 2 between frame $i+1$ and i means that loudness has doubled from frame i to $i+1$. This can be used as a cue to locate onsets because during onsets some sub-bands shall display significant change in loudness even if the total loudness has not changed. This phenomenon is due to tonal discontinuity [6]. If there has been a significant change in total loudness then due to clear discontinuity or partial discontinuity or mixed discontinuity[6], we can observe a good change in loudness at certain sub-bands. To detect loudness change, we use the loudness ratio between current frame loudness and the average loudness over the previous k frames as the detection function. This averaging has the effect of smoothing out spurious peaks in the loudness ratio function while preserving onsets. Let $L_{i_{mean}}$ be the mean loudness of the i^{th} frame, then

$$L_{i_{mean}}(j) = \frac{\sum_{m=j-k}^{j-1} L_i(m)}{k} \quad (3)$$

$$SLR_i(j) = \frac{L_i(j)}{L_{i_{mean}}(j)} \quad (4)$$

$$SLR_i(j) > Thr_{loud} \quad (5)$$

where SLR_i is the sub-band loudness ratio (SLR) for the i^{th} sub-band. We tried with $k=2$ to $k=7$ in Eqn.3 and found that $k=5$ gave us optimal results . For each sub-band, only those frames where a loudness change greater than $Thr_{loud} = 1.5$ are retained while other frames are made 0. The SLR for all the 51 sub-bands are plotted in Fig.3.

2.7 Binary thresholding and onset grouping

An onset produces spectral / timbral changes across sub-bands [5, 6, 12] and hence the SLR function must display significant loudness changes in several sub-bands for a valid onset. To simulate this we use a second threshold $Thr_{filt} = 6$

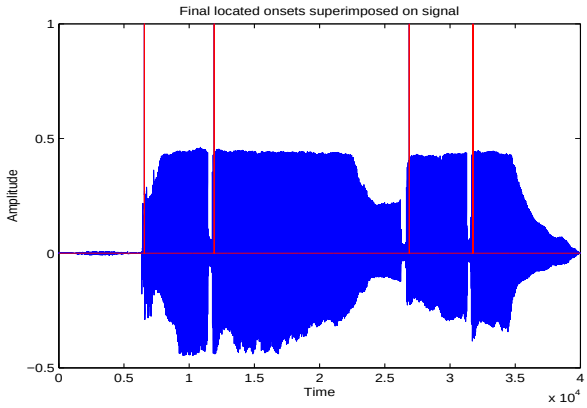

Figure 5: Detected onsets superimposed on hum signal

to choose onsets that are significant in at least 6 sub-bands as shown in Fig.4. The figure shows a plot of number of sub-bands that experience significant loudness change against time. Finally, the group of onsets that are close (less than 60ms apart) are grouped together and represented by the strongest onset in the group[7, 13]. Fig.5 shows the onsets detected superimposed on the original hum signal.

2.8 Experiments and Results

We have tested our algorithm on an annotated database of syllabic and natural hums. The database has hums using the syllable /da/, /la/, /na/ and /hm/. There are 47 hums corresponding to each of the syllables hummed by 5 singers. All the hums have been recorded in a noise free environment at 22kHz sampling rate. Onsets were manually labelled using the gating method[14]. More details about the database and it's annotation can be found in [5, 6]. The parameters used for our simulations were obtained by optimizing the algorithm's performance over a set of 10 hums, disjoint from the test set.

Onsets found using the algorithm are compared with the annotated database and an onset is considered valid (Correct Detection (CD)) if it is closer than $\pm$70ms to the reference onset, otherwise it is marked as a false positive (FP). The database has a total 2994 onsets when we consider only the

Table 1: Accuracy of the onset detection algorithm for each syllabic class of hums

Hum class	Actual	C.D	F.P	%C.D	%F.P
/da/	756	737	18	97.5	2.4
/la/	735	686	46	93.2	6.2
/na/	765	710	32	92.8	4.2
/hm/	738	544	82	73.6	11
/da/+/la/+/na/	**2256**	**2133**	**96**	**94.5**	**4.2**
Total	**2994**	**2677**	**178**	**89.4**	**5.9**

first 10 seconds of each file for our tests. The performance of the algorithm is shown in Table.1. For the 3 classes considered (only /da/,/la/ and /na/) our algorithm matches the system in [5]. The advantage of our algorithm compared to [5, 6] is in the reduced processing, simplicity of implementation and very little post-processing. We achieve comparable results with just one feature with very little use of heuristics while [5, 6] use heuristics based on nature of speech, aspiration etc to achieve optimal results.

As can be seen, compared to the existing algorithms that require heuristics to choose the multiple thresholds, our algorithm gives a good performance and it is simple to choose the thresholds. Both the thresholds Thr_{loud} and Thr_{filt} have a physical correspondence to the audio signal. A higher value of Thr_{loud} would suggest that we consider detecting only hard onsets while a lower value would suggest even soft onsets would be considered. Similarly, a higher value of Thr_{filt} suggests that we detect only sounds that have greater timbral texture while a lower value suggests that we can detect even soft onsets or sounds concentrated in narrow frequency bands (Ex: bass drums).

3. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a simple algorithm using psychoacoustics to detect perceptually relevant onsets in hum/syllabic query audio. Though some optimization needs to be done for natural hums (using /hm/), it's performance for syllabic humming matches that of current state of the art systems. Another advantage is the perceptual perspective to the choices of thresholds which allows us to choose only strong onsets or even weak onsets depending upon their strength. This algorithm is currently acting as a front end for a QBH system currently under development. The algorithm's performance is being optimized to work for real recording conditions in various noisy environments and initial results in this direction are promising.

4. ACKNOWLEDGMENTS

The authors would like to thank Prof.Preeti Rao of IIT,Bombay for providing us the hum database used in our experiments.

REFERENCES

[1] B.Sundaram M.A.Raju and P.Rao, "Tansen:a query by humming based music retrieval system," *Proc of National conference on Communications(NCC)*, 2003.

[2] B.Zhang C.C.Toh and Y.Wang, "Multi-feature fusion based onset detection for solo singing voice," *Intl Society for Music Information Retrieval conference(ISMIR)*, 2008.

[3] L.Prechelt and R.Typke, "An interface for melody input," *ACM-Transactions on Computer-Human Interaction(ACM-TOCHI)*, 2001.

[4] M.Lessafre, D.Moelants, M.Leman, B.D.Baets, H.D.Meyer, G.Martens, and J.P.Martens, "User behaviour in spontaneous reproduction of musical pieces by vocal query," *European Society for the Cognitive Sciences of Music(ESCOM)*, 2003.

[5] P.P.Kumar, M.Joshi, S.Hariharan, S.D.Roy, and P.Rao, "Sung note segmentation for a query by humming system," *Intl Joint Conferences on Artificial Intelligence(IJCAI)*, 2007.

[6] P. Rao P.P.Kumar and S.D.Roy, "Note onset detection in natural humming," *Intl Conference on Computational Intelligence and Multimedia Applications(ICCIMA)*, 2007.

[7] A.Klapuri, "Sound onset detection by applying psychoacoustic knowledge," *IEEE Intl Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 1999.

[8] Balaji Thoshkahna and K.R.Ramakrishnan, "A psychoacoustics based sound onset detection algorithm for polyphonic audio," *Intl Conference on Signal Processing(ICSP)*, 2008.

[9] B.C.J.Moore, B.Glasberg, and T.Baer, "A model for the prediction of thresholds,loudness and partial loudness," *Journal of Audio Engineering Society(JAES),Vol.45,No.4*, 1997.

[10] Balaji Thoshkahna and K.R.Ramakrishnan, "A psychoacoustically motivated sound onset detection algorithm for polyphonic audio," *Intl.Conference on Signal Processing and Multimedia Applications(SIGMAP)*, 2009.

[11] J.Timoney, T.Lysaght, M.Schoenwiesner, and L.McManus, "Implementing loudness models in matlab," *7th Intl. Conference on Digital Audio Effects-DAFx*, 2004.

[12] J.P.Bello, L.Daudet, S.Abdallah, C.Duxbury, M.Davies, and M.B.Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing,*, vol. 13, no. 5, September 2005.

[13] N.Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," *Proc. of Audio Engineering Society Convention*, 2005.

[14] D.J.Hermes, "Vowel onset detection," *Journal of Acoustical. Society. of America.87(2), 866-873*, 1990.