Our query by tapping mechanism takes the user input in the format of tapping on the microphone and the extracted duration of notes is then used to retrieve the intended song in the database. Since there is no singing or humming, no pitch information is used in the retrieval process at all. Most people would think that it is hard to do music retrieval via beat information alone. However, our experiments demonstrate that beat information is also an effective feature in the sense that it can be used to retrieve the intended song from a large collection of music database with a satisfactory recognition rate.

To extract the duration of each note, we need to do frame blocking first and then find the energy of each frame. We find the local maxima of the log energy. The local maxima are legal only when their values are greater than a heuristically determined threshold.

The comparison procedure is based on the concept of dynamic time warping (DTW). Suppose that the (normalized) input timing vector (or test vector) is represented by $t(i), i = 1, \ldots, m,$ and the (normalized) reference timing vector (reference vector) by $r(j), j = 1, \ldots, n$. These two vectors are not necessarily of the same size and we can apply DTW to match each point of the test vector to that of the reference vector in an optimal way.