The Interplay of Information Retrieval and Query by Singing with Words

Iman Amini, Alexandra Uitdenbogerd, Mark Sanderson

School of Computer Science, RMIT

2011
Contents

- Introduction
- Data Collection and Manipulation
- Matching Techniques
- Results
- Conclusion
Speech recognition can be used to exploit sung words from the queries.

Alternative ways, frequencies converted to symbolic notes. [Also Humming]

Many researchers have tried to overcome the challenge of noisy queries by configuring Automatic Speech Recognition System (ASR) system and adapting it to singing.

We focus on choosing and combining best matching algorithms.
Every singing query carries two information:
  - Verbal and Melodic

A complete Music Information Retrieval (MIR) system uses both.

We exploit the **verbal** information and test and combine 5 string and character Matching algorithms:

- Edit Distance
- Syllable Alignment Pattern Searching (SAPS-L)
- Editex
- Okapi
- Word-Count
We chose a light version of speech recognition system *PocketSphinx* which allowed us to experiment with a naturally low starting point of accuracy.

Original lyrics misheard by the Speech Recognition system

<table>
<thead>
<tr>
<th>Recognized Segment</th>
<th>Original Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>me been raining</td>
<td>it’s been raining</td>
</tr>
<tr>
<td>gold who mind do fool around</td>
<td>don’t fool around don’t fool around</td>
</tr>
<tr>
<td>every say the a</td>
<td>every single day</td>
</tr>
<tr>
<td>let there be sang</td>
<td>let there be spring</td>
</tr>
<tr>
<td>morning light and we sang chan</td>
<td>morning light and we sang here</td>
</tr>
</tbody>
</table>
50 audio chunks were manually selected from the sung lyrics and used for test data and queries.

Each query is about 30 seconds long, consisting of 12 to 20 words.

Training data
1 hour speech data with no especial context
2 hrs. recorded singing [43 tracks] First 2 authors of this paper.

Correctness = \frac{N - D - S}{N}

Accuracy = \frac{N - D - S - I}{N}

<table>
<thead>
<tr>
<th>Total words</th>
<th>Percent correct</th>
<th>Error</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1155</td>
<td>40.78%</td>
<td>79.22%</td>
<td>20.78%</td>
</tr>
<tr>
<td>Insertions</td>
<td>Deletions</td>
<td>Substitutions</td>
<td></td>
</tr>
<tr>
<td>231</td>
<td>34</td>
<td>650</td>
<td></td>
</tr>
</tbody>
</table>
Not many singing collection available!

For ASR system we use the default unigram and bigram Language Model

We build our test collection:

Gathered 2359 lyrics collection
2 hrs. song recording sung by two authors of the paper
A professional female singer and a beginner male

We add those 43 songs to our singing collection

No stop words or stemming were used to index the collection
Each query can have only one relevant answer which is the target lyric. Therefore we use **success @ 10** and **mean reciprocal rank** to report the performance of our retrieval systems.
Techniques were grouped into 3 categories

- **Character-Based** (all configured to local alignment on the longer string (lyrics))
  - Edit Distance

- **Word matching**
  - Word Count
  - Okapi

- **Phonetic matching**
  - SAPS-L
  - Editex

- **Combination of Techniques** (Okapi + Phonetic Matching)
  - Okapi- Edit Distance
  - Okapi- Editex
Edit Distance and all character based matching algorithms were configured for local alignment over the longer string (lyric)

\[
\begin{align*}
\text{edit}(0, 0) &= 0 \\
\text{edit}(i, 0) &= i \\
\text{edit}(0, j) &= j \\
\text{edit}(i, j) &= \min[\text{edit}(i - 1, j) + 1, \\
&\quad \text{edit}(i, j - 1) + 1, \\
&\quad \text{edit}(i - 1, j - 1) + r(s_i, t_j)]
\end{align*}
\]
Syllable Alignment Pattern Searching (SAPS)
Important and distinguishing phonemes tend to occur at the start of syllable so configure the scoring based on that.

1. Preprocessing: SAPS-L transforms the plain text into a canonical form in order to obtain an accurate syllable segmentation. For example, it maps “tjV” (where V is any vowel) to “chV” and “ph” to “f”.

2. Segmentation: The strings are segmented into “syllables”. For example, “dancing” is segmented thus: “DanSing”.

3. Alignment and Similarity Calculation: Alignment is similar to edit distance, but with more complex scoring.
Initially we run the experiments with 5 techniques.

<table>
<thead>
<tr>
<th>Word Count</th>
<th>SAPS-L</th>
<th>Editex</th>
<th>Okapi</th>
<th>Edit Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.25%</td>
<td>58%</td>
<td>51.75%</td>
<td>37.5%</td>
<td>56.25%</td>
</tr>
</tbody>
</table>

Then we combine String and character based matching

<table>
<thead>
<tr>
<th>Okapi-Edit Distance</th>
<th>Okapi-Editex</th>
</tr>
</thead>
<tbody>
<tr>
<td>58%</td>
<td>56%</td>
</tr>
</tbody>
</table>

\[
\text{okapi-edit distance} = 0.5(O) - 2(ED) \\
\text{okapi-editex} = O - 2(EX)
\]

\(O\) : score generated by Okapi  \\
\(ED\) : score generated by Edit Distance  \\
\(EX\) : score generated by Editex
Then we experiment with two variables:

1- Query Accuracy
2- Query Length
Experiments with Buckets with varying level of accuracy
(Reiterating that a low starting point of ASR result was a good way of testing the techniques across a larger varying level of accuracy.

50 Queries were sorted by their accuracy scores and 4 different set of buckets were created.

May still contain similarly pronounced words with wrong spellings.
Trend of the retrieval techniques against four different sets of queries

The marginal success of some techniques at 0% level of correctness, “Okapi” in particular, were questionable.
Trend in retrieval effectiveness with varying query length, the longer the queries the better the results.
4 top techniques did not have any statistical difference so we calculate reciprocal ranks for each query to find which technique is more robust and for what point of correctness.
Despite the faulty output produced by the speech recognition, some matching techniques performed reasonably well.

Combination of word based and character based techniques can improve the retrieval performance.

It would also be useful to compare recognition output that is naturally at a higher level of accuracy and to use a larger singing collection.
Thanks!

Questions?