Towards Query by Singing / Humming on Audio Databases

- **Preprocessing**
  - **Karaoke Filter**: removes center pan (information contained in both channels) by inverting one channel and mixing it together with the other into a mono signal:
    
    \[ \text{output} = L - R \]

  - **Requirements**:
    - stereo input signal
    - lead voice (and possibly solo instruments) centered in the stereo mix
    - instruments and backing vocals arranged out of center

- **Local Noise Filter**: derives a local (i.e. continuously updated) power spectrum of frequencies from a noise signal \( N \) which can then be removed from the signal \( S \) (based on versions 1.34, Sep 23, 2006 and 1.39, Jul 27, 2007 of the NoiseRemoval effect by Dominic Mazzoni as part of Audacity)

- **Bandpass Filter (300-1000Hz)**: keeps only frequency range of the input signal \( S \) that is relevant for human voice (lower bound is higher to filter out the bass guitar that might be in the center as well)

#### Modified SAX Approach
- **Piecewise Aggregate Approximation (PAA)**:
  A feature time series \( C=(c_1,...,c_n) \) is aggregated by factor \( n/w \) to \( C=(\tau_1,...,\tau_m) \)

  \[ \tau_i = \frac{w}{n} \sum_{j=1}^{i} c_j, \quad 0 < i \leq n \]

- **Estimation of \( N \) quantiles** of the distribution of the PAA values
- **Discretization** of the values within a quantile to a unique symbol (lookup-table) guarantees an equally distributed alphabet.

- **Symbols are defined to be equidistant**, depending only on the distance in the ordered alphabet:

  \[ d(s_i, s_j) = \frac{2i - j}{N - 1} \]

  (where \( i,j \) are symbol indices and \( N \) is the alphabet size)

- **Extraction of low level audio features (using JAudio)**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Power</td>
<td>0.6207</td>
</tr>
<tr>
<td>Chroma</td>
<td>0.5960</td>
</tr>
<tr>
<td>1st Formant (FF1)</td>
<td>0.5696</td>
</tr>
<tr>
<td>1st derivatives ( d\text{FF1} )</td>
<td>0.5490</td>
</tr>
</tbody>
</table>

- **High Level Patterns**

  - **Manual definition of generic shapes** in the time series of a feature (Audio Power):
    - **Flat patterns** - sections of silence or quiet background with low mean and low variance, or located between elevations
    - **Smooth elevations** - mean above a certain threshold, only one peak, probably describing single syllables
    - **Toothy structures** - elevations with mean above a certain threshold and more than one peak
    - **Undefined or noisy regions** - may result from quiet singing or filtered out instruments

  - **Manual definition of symbol distances**:
    - **round**
    - **toothy**
    - **noise**

  - **Song (“Why” performed by Annie Lennox)**

#### Test database: 200 songs from Rock/Pop/Soul

<table>
<thead>
<tr>
<th>Evaluation Measures:</th>
<th>ground truth</th>
<th>humanized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Accuracy: ( \text{MoA} = \frac{1}{n} \sum_{i=1}^{n} \frac{n - \text{rank}(t_i)}{n} )</td>
<td>0.5965</td>
<td>0.5796</td>
</tr>
<tr>
<td>Mean Reciprocal Rank: ( \text{MRR} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\text{rank}(t_i)} )</td>
<td>0.6262</td>
<td>0.6328</td>
</tr>
</tbody>
</table>

**Best performance**:
- aggregation factor = 4
- alphabet size = 12 (except 3 per bin for chroma)

**Improvement by Boosting** (beginning/chorus):
- \( \text{MoA} = 0.79 \), \( \text{MRR} = 0.3 \)
- top: 23.3%
- top-3: 30%
- top-10: 41.1%

**Features**:
- simple features
- 1st MFCC (MFCC1)
- 2nd MFCC (MFCC2)
- Fundamental Freq.
- 1st Formant (FF1)
- Chroma
- high-level patterns
- HLP/AF
- high-level patterns
- HLP/AF

**Human Queries**

<table>
<thead>
<tr>
<th>Feature combinations</th>
<th>humanized</th>
<th>ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>(MCC1, MRR)</td>
<td>0.5960</td>
<td>0.6262</td>
</tr>
<tr>
<td>(MCC1, FF1)</td>
<td>0.6095</td>
<td>0.6328</td>
</tr>
<tr>
<td>(MCC1, dMCC1)</td>
<td>0.6237</td>
<td>0.6409</td>
</tr>
<tr>
<td>(MCC1, dMCC1, FF1)</td>
<td>0.6237</td>
<td>0.6409</td>
</tr>
</tbody>
</table>

**Symbolic representation capturing the main characteristics of the lead voice allows application of string matching techniques (e.g. Levenshtein distance)**

**Output**

- symbolic representation capturing the main characteristics of the lead voice allows application of string matching techniques (e.g. Levenshtein distance)

**Idea**: exploit spatial arrangement of instruments and voices in the mix

**Karaoke Filter**

- removes center pan (information contained in both channels) by inverting one channel and mixing it together with the other into a mono signal:

  \[ \text{output} = L - R \]

**Requirements**:
- stereo input signal
- lead voice (and possibly solo instruments) centered in the stereo mix
- instruments and backing vocals arranged out of center

**Local Noise Filter**: derives a local (i.e. continuously updated) power spectrum of frequencies from a noise signal \( N \) which can then be removed from the signal \( S \) (based on versions 1.34, Sep 23, 2006 and 1.39, Jul 27, 2007 of the NoiseRemoval effect by Dominic Mazzoni as part of Audacity)

**Bandpass Filter (300-1000Hz)**: keeps only frequency range of the input signal \( S \) that is relevant for human voice (lower bound is higher to filter out the bass guitar that might be in the center as well)