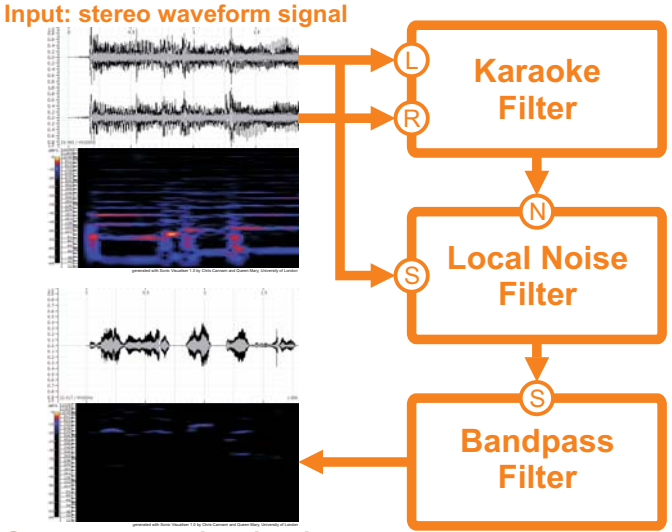
Towards Query by Singing / Humming on Audio Databases

Information Retrieval Group

Preprocessing



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:

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)

Extraction of low level audio features (using JAudio)

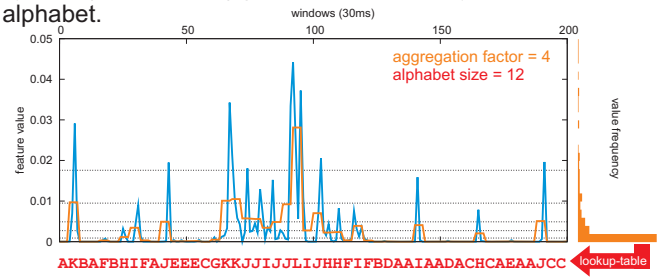
Transformation

Modified SAX Approach

- **Piecewise Aggregate Approximation (PAA) :**
  A feature time series $C=(c_1,..,c_n)$ is aggregated by factor n/w to $\overline{C}=(\overline{c}_1,..,\overline{c}_w)$:

$$\overline{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad , 0 < i \leq n$$

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



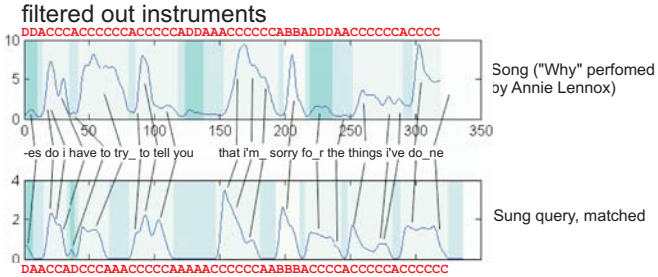
- Symbols are defined to be **equidistant**, depending only on the distance in the ordered alphabet:

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

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

High Level Patterns

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



- **Manual definition of symbol distances:**

	flat	round	toothy	noisy
flat	0	2	1.5	1
round	2	0	0.5	1
toothy	1.5	0.5	0	1
noisy	1	1	1	0

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

Experiments

Test database: 200 songs from Rock/Pop/Soul

Evaluation Measures:
- *Mean of Accuracy:* $MoA = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n-rank(t_i)}{n-1}\right)$
- *Mean Reciprocal Rank:* $MRR = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{rank(t_i)}\right)$

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

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

MIDI Queries	ground truth		humanized	
	MoA	MRR	MoA	MRR
simple features				
1st MFCC (MFCC1)	0.5965	0.0338	0.5678	0.0289
2nd-5th MFCC	0.6096	0.0735	0.5677	0.0252
Audio Power (AP)	0.6262	0.0574	0.5522	0.0495
Fundamental Freq.	0.6098	0.0494	0.5678	0.0289
1st Formant (FF1)	0.5398	0.0344	--	--
Chroma	0.6328	0.1242	0.6380	0.0618
1st derivatives				
dAP	0.6237	0.0924	0.6189	0.0872
dChroma	0.5490	0.0320	--	--
high-level patterns				
HLP(AP)	0.5390	0.0350	--	--
feature combinations				
(AP, dAP)	0.6708	0.0764	0.6085	0.0826
(AP, dAP, Chroma)	0.6979	0.1426	0.6439	0.0820
	(150 MIDI files)		(tempo, pitch and pauses altered)	

Human Queries	hummed		sung	
	MoA	MRR	MoA	MRR
simple features				
1st MFCC (MFCC1)	0.5867	0.0393	0.7325	0.1950
2nd-5th MFCC	0.5361	0.0376	0.6325	0.1307
Audio Power (AP)	0.6127	0.0549	0.6794	0.1558
Fundamental Freq.	0.5113	0.0362	0.5586	0.0585
1st Formant (FF1)	0.5696	0.0251	0.6351	0.0821
Chroma	0.6095	0.0312	0.5879	0.1288
1st derivatives				
dMFCC1	0.5574	0.0641	0.6062	0.1032
high-level patterns				
HLP(AP)	0.5970	0.0588	0.6925	0.1454
feature combinations				
(MFCC1, dMFCC1)	0.5796	0.0456	0.7552	0.2164
(MFCC1, FF1)	--	--	0.7447	0.2457
(MFCC1, dMFCC1, FF1)	--	--	0.7635	0.2329
	(150 queries from a non-professional)		(130 queries from 7 non-professionals)	

Otto-von-Guericke-University Magdeburg, Germany - Faculty of Computer Science - Department for Knowledge and Language Engineering
Information Retrieval Group - http://irgroup.cs.uni-magdeburg.de

Alexander Duda, Andreas Nürnberger, and Sebastian Stober
{stober,nuernb}@iws.cs.uni-magdeburg.de