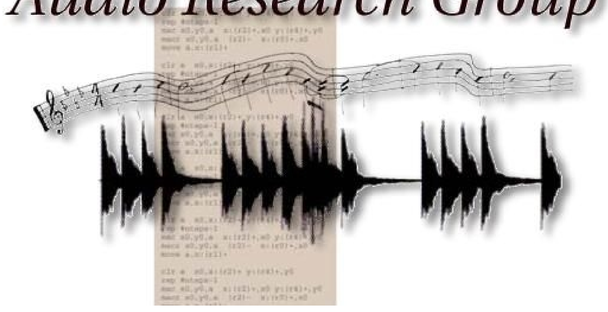
Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods

Annamaria MESAROS, Tuomas VIRTANEN, Anssi KLAPURI

Audio Research Group



Framework

- Polyphonic music – interfering sounds make automatic singer identification difficult
- Two main approaches: computing the acoustic features directly from the polyphonic mixture and classify
  separating the vocal line, computing the acoustic features from the separated vocals and classify
- Acoustic material: 65 songs from 13 singers, 20-30 seconds in length
  each song mixed at 5 different singing-to-accompaniment ratio levels
  accompaniment and mixing procedures are not singer specific

Features and models

- Features: 12 MFCCs computed on 34 ms frames, no delta MFCCs
- Pattern recognition tools:
  - Discriminant functions - a set of linear/quadratic functions of the data;
    the functions are evaluated for each new observation, the observation is assigned to the class having the highest discriminant value
  - Gaussian mixture models trained by expectation maximization;
    find the class that maximizes the likelihood of the test observations (acoustic features of successive time frames, statistically independent)
  - Nearest neighbor classification using symmetric Kullback Leibler divergence

KL divergence approximation

- Kullback Leibler divergence – measure of the difference between two probability distributions: from a "true" probability distribution *P1* to an arbitrary probability distribution *P2*

$$D(p_1(\mathbf{x})||p_2(\mathbf{x})) = \int p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} dx$$ solvable only for single Gaussian

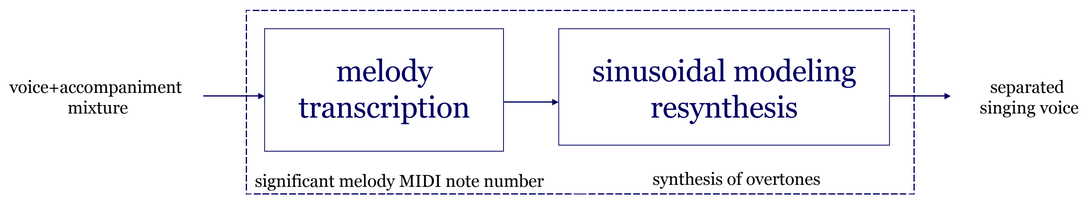
- Symmetrized KL divergence: $S(p_1(\mathbf{x})||p_2(\mathbf{x})) = D(p_1(\mathbf{x})||p_2(\mathbf{x})) + D(p_2(\mathbf{x})||p_1(\mathbf{x}))$

Monte Carlo approximation for multiple Gaussians: $D(p_1(\mathbf{x})||p_2(\mathbf{x})) \approx \sum_{m=1}^{M} \frac{1}{M} \log \frac{p_1(\mathbf{x}_m)}{p_2(\mathbf{x}_m)}$

written using likelihoods $D_{\text{emp}}(p_1(\mathbf{x})||p_2(\mathbf{x})) = \frac{1}{M} \log \frac{L(X_1;\lambda_1)}{L(X_1;\lambda_2)}$ → fixed for each model

$S_{\text{emp}}(p_1(\mathbf{x})||p_2(\mathbf{x})) = \frac{1}{MN} \log \frac{L(X_1;\lambda_1)L(X_2;\lambda_2)}{L(X_1;\lambda_2)L(X_2;\lambda_1)})$ → used as similarity measure for nearest neighbor classification

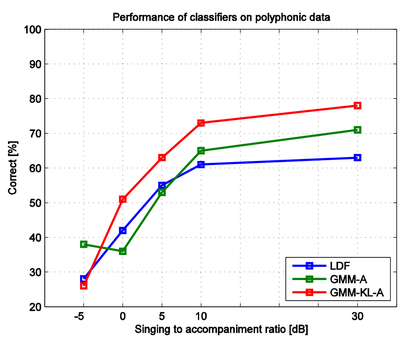
Vocal line separation

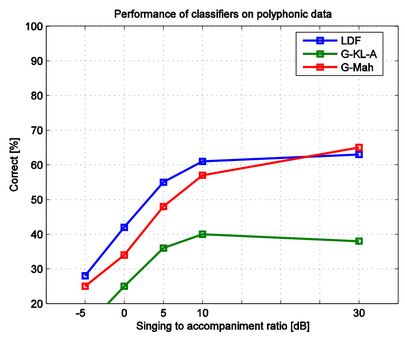


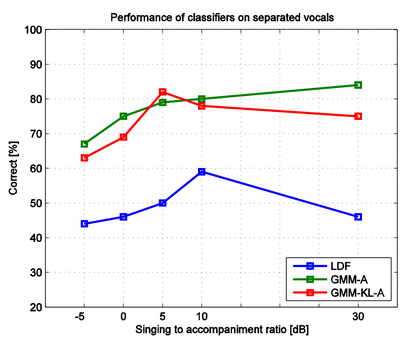
Artist and song level models

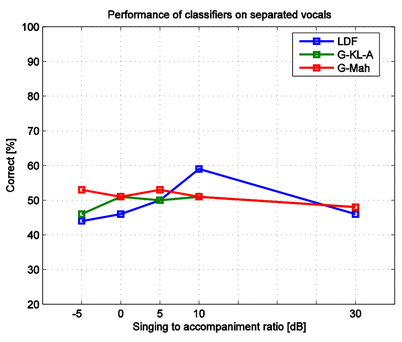
- artist level modeling – one model trained with all the songs from the training set, resulting in 13 models, one for each singer; test song is classified according to closest singer model
- song level modelling – one model for each song from the training set, resulting in several models associated to each singer; test song is classified according to the singer of the song which is closest to the one under analysis

Simulation experiments









Classification:
- 4-fold cross validation, one song for testing, the rest for training the singer model
- run for -5dB, 0dB, 5dB, 10dB and 30dB singing-to-accompaniment ratio mixtures
- using acoustic features computed directly from the polyphonic mixture
- using acoustic features computed on the separated vocal line

In the figures:
- LDF - linear discriminant functions;
- GMM-A - artist-level GMMs, maximum likelihood classification;
- GMM-KL-A - artist-level GMMs, classification using nearest neighbor based on symmetrical KL divergence;
- G-KL-A - artist-level Gaussian, nearest neighbour classification based on symmetrical KL divergence;
- G-Mah - artist-level Gaussian, nearest neighbor classification based on Mahalanobis distance

Conclusions

- The proposed method for KL divergence approximation combined with nearest neighbor classification produces comparable identification results with the best reference methods
- At low singing-to-accompaniment ratio, vocal line separation improves significantly the identification performance

TAMPERE UNIVERSITY OF TECHNOLOGY
Institute of Signal Processing