# Self-learning Acoustic Feature Generation and Selection for the Discrimination of Musical Signals

Björn Schuller, Gerhard Rigoll
*Technische Universität München, Arcisstraße 21, D-80333 München, Deutschland,*
*Email: {schuller|rigoll}@tum.de*

## Introduction

Optimal features for the discrimination of musical signals are largely discussed. Herein we therefore present a self-learning approach to this problem based on time series analysis and evolutionary feature-space optimization. The feature basis is formed by a multiplicity of dynamic acoustic Low-Level-Descriptors as pitch, intensity, and spectral information. These are filtered and pre-processed with special respect to human perception. From hereon a systematic derivation of further contours and functionals by means of descriptive statistics takes place. The resulting high-dimensional space of static features is then optimized by combined sequential floating and genetic search. As learning function we apply Support Vector Machines, known for their high performance within this task. In order to allow more flexibility we integrate alteration and combination of attributes by mathematical operations. Applicability of the proposed approach is demonstrated by extensive test-runs on large public databases of musical signals containing among others segments of drumbeats, a-cappella singing, or multi-instrumental phrases. Outstanding performances can be reported for the discrimination of the signal type out of a stream.

## Databases

Firstly, we use the public Columbia University Speech Music Discrimination (SMD) database introduced and used in [1]. This database contains among other samples segments from a radio broadcast stream. We use the total of 101 samples of music, 80 of speech, and 60 of music overlaid with speech contained in this database herein.

Secondly, we extend our previously introduced SHANGRILA corpus of speech and monophonic singing samples [2]. It comprises of 1,000 samples of speech and 1,114 samples of singing of 58 persons in total. These audio samples have been recorded in 16bit, 11 kHz by use of an AKG MK 1000S-II condenser microphone. They resemble interaction turns with a music retrieval interface. Polyphonic music clips are taken from 200 songs of the MTV-Europe-Top-10 of the years 1981-2000. The clips were cut out at five fixed relative positions of each song resulting in 1,000 clips in total. The genres covered resemble typical mainstream pop-music radio station sound. Additionally, we added 1,000 drum beat clips that consist of various styles as disco, jazz, rock, and techno music. The whole corpus is abbreviated SAB in the ongoing. By this second database we can show results on a higher total of samples and for further audio signal types.

## Systematic Feature Generation

We use systematic generation of functionals out of time-series by means of descriptive statistics: Firstly, selected base-contours, respectively Low-Level-Descriptors (LLD), are calculated well known to carry information about the musical signal type. The original sampling frequency and quantization of the databases is kept, and each 10 ms a 20 ms frame is extracted by weighting with a Hamming window-function. Aiming at broad coverage, estimated feature contours contain log frame energy, pitch based on autocorrelation (ACF) in the time-domain and Dynamic Programming (DP) to minimize deviations on a global level, pitch epochs, harmonics-to-noise ratio based on ACF, formant bandwidth, position and amplitude of the first 5 formants based on LPC, polynomial roots and DP. Further more jitter and shimmer are calculated. For spectral analysis 16 MFCCs, and spectral flux, spectral centroid, as well as spectral roll-off based on linear DFT-spectral coefficients and polynomial dB-correction in accordance to human perception, are extracted. Dominant harmonics in the spectrum are tracked in 47 chromatic semitone intervals within human voice range by summing over three successive partials. Finally, 19 Voc19 coefficients are obtained by JSRU-style 19-channel filter-bank analysis using two second-order section Butterworth band-pass filters. Energy smoothing is done at 50Hz. As a side comment it may be mentioned that within the MPEG-7 standard partly similar LLDs are defined. The contours are subsequently smoothed by symmetrical moving average low-pass filtering with a window size of three. Successively, speed ($\partial$) and acceleration ($\partial^2$) are derived as further LLDs for each basic contour in order to model temporal behavior. Afterwards a total of 19 phrase-wise derived hi-level functionals by means of descriptive statistics per contour is computed. These are linear momentums of the first four orders, namely mean, centroid, standard deviation, Skewness and Kurtosis, as well as quartiles, ranges, extrema, extrema positions, zero-crossing-rates, and roll-off-points. Likewise roughly 7k acoustic features are obtained in total to build a broad feature basis for the subsequent feature space optimization process.

## Classification and Feature Space Optimization

Considering our extensive classifier comparison on the SHANGRILA database in [2] we chose Support Vector Machines (SVM) with couple-wise one-vs.-one multi-class discrimination and polynomial kernel herein. Reduction of less relevant features often leads to higher classification performance, as the classifier is confronted with less complexity. Due to the unusually high initial dimensionality we apply fast Information Gain Ratio based feature selection (IGR-FS), firstly. In this filter-reduction single highly

relevant attributes are found by their entropy. After pre-reduction to the optimal feature set size by IGR-FS we apply the more powerful Sequential-Forward-Floating-Search (SFFS) - a Hill-Climbing search - to further reduce feature set size having SVM as wrapper-function. Apart from mere reduction of the feature space, also a combination with supervised expansion can lead to improved accuracy. We therefore generate novel features based on the so far pre-selected ones: Firstly, alteration of attributes by mathematical operations can be performed to lead to better representations of these. Secondly, by association of attributes we can obtain a further number of new information. As a deterministic and systematic generation comes to its limits applying exhaustive search, we decided for Genetic Algorithm (GA) based search through the possible feature space. As a start-set of effectually different individuals that represent possible solutions to the problem we use partitions of the acoustic feature sets reduced to a reasonable size by now. The partitions are denoted in binary coding, called *chromosomes*. Each chromosome consists of *genes* that correspond to single features within the partition. A feature's gene consists of one bit for its activity status. The partitioning is done randomly throughout initialization and we obtain $N=\dim(\underline{x})/n$ individuals if $\underline{x}$ denotes the feature vector, and $n$ the partition size. By an initialization probability, set to 0.5 in our case, it is randomly decided which original features are chosen for one step of genetic generation. We decided to have a *population* size of 20 individuals at a time. Next a *fitness* function is needed in order to decide which individuals survive. Thereby the aimed at classifier forms a reasonable basis in view of wrapper based set optimization. A cyclic run over multiple *generations* is afterwards executed until an optimal set is found, which resembles a local maximum of a problem: Firstly, a *Selection* takes place, based on the fitness of an *individual*. We use common *Roulette Wheel* selection within this step. Thereby the 360° of a roulette wheel are shared proportional to the fitness of an individual. Afterwards the "wheel" is turned several times, resembling $N$ times selecting out of $N$ individuals. Selected individuals are assembled in a *Mating Pool*. Likewise, fitter individuals are selected more probably. We also ensure mandatory selection of the best one, known as *Elitist Selection*. The oncoming *Crossing* of pairs is fulfilled by picking $N/2$ times individuals with the probability $1/N$. After selection, individuals are put aside. Opposing traditional GA, we use a variable chromosome length from hereon, as we aim at generation of features. First we have to pick to *parents* in order to cross their chromosomes and thereby obtain new *children*. We then choose simple *Single-Point-Crossing* which splits each parent chromosome close to its center and pastes the two halves cross-wise to obtain two children. The fitness thereby also limits the total number of children an individual may produce. Afterwards, *Mutation* takes place: the state of a gene, respectively of a feature within a partition, is randomly changed by a probability of 0.5. Likewise features can be excluded from a set. To generate new feature we insert a random selection of an alteration method out of *reciprocal value*, *addition*, *subtraction*, *multiplication* and *division* [3]. Depending on the mathematical operation the appropriate number of features within an individual is selected for alteration, the operation is performed, and obtained features are appended. Now the evaluation of the population is fulfilled, resembling the fitness-test respectively classification with the feature sub-sets. At this point, one iteration is finished, and the algorithm starts over with selection. We decided for a maximum of 50 generations, and 40 without improvement.

## Experiments and Conclusion

As a general mean of evaluation we use 10-fold stratified cross-validation. The first table shows results for the stepwise optimization of the feature space. For the public database SMD the features were reduced starting from 7k to 197 by IGR-FS, afterwards to 76 by SVM-SFFS. By genetic generation 7 new features could be added basing on these. In a similar way features were reduced and generated for SAB. Within the next table class-wise mean error rates and $F_1$-measures are presented for the SMD database. The final table depicts class-wise mean error-rates for the database SAB. As all samples are evenly distributed among classes, F-Measures are spared.

**Table 1:** Error stepwise feature-space optimization

| Error [%] | SMD | SAB |
|---|---|---|
| **7k features** | 8.39 | 9.15 |
| **+IGR-FS** | 5.71 | 4.97 |
| **+SVM-SFFS** | 3.67 | 3.08 |
| **+GA Generation** | 3.27 | 2.14 |

**Table 2:** Class-wise error and F-Measure database SMD

| SMD | Speech | Music | Mu+Sp |
|---|---|---|---|
| **Error [%]** | 2.50 | 3.96 | 1.67 |
| **$F_1$-Measure [%]** | 98.1 | 97.0 | 95.2 |

**Table 3:** Class-wise error database Shangrila+Beat

| SAB | Speech | Music | Singing | Beat |
|---|---|---|---|---|
| **Error [%]** | 0.25 | 0.10 | 1.44 | 6.78 |

Significant improvements within every step could be demonstrated on both test-sets. Overall achieved error rates are outstandingly low: besides the discrimination of speech, music, monophonic singing, and music overlaid with speech also drum beats could be recognized.

## Literature

[1] E. Scheirer, M. Slaney: "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP 97*, pp. 1331–1334, 1997.

[2] B. Schuller, et al.: "Feature Selection and Stacking for Robust Discrimination of Speech, Monophonic Singing, and Polyphonic Music, " *Proc. ICME 2005*, Netherlands, 2005.

[3] I. Mierswa: "Automatic Feature Extraction from Large Time Series," *Proc. 28. Annual Conference of the GfKl 2004*, Springer, pp. 600-607, 2004.