FEATURE SELECTION AND STACKING FOR ROBUST DISCRIMINATION OF SPEECH, MONOPHONIC SINGING, AND POLYPHONIC MUSIC

Björn Schuller, Bernardo José Brüning Schmitt,
Dejan Arsić, Stephan Reiter, Manfred Lang, Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
Arcisstraße 21, D-80333 München, Germany
(schuller | mueller | arsic | lang | rigoll)@mmk.ei.tum.de

ABSTRACT

In this work we strive to find an optimal set of acoustic features for the discrimination of speech, monophonic singing, and polyphonic music to robustly segment acoustic media streams for annotation and interaction purposes. Furthermore we introduce ensemble-based classification approaches within this task. From a basis of 276 attributes we select the most efficient set by SVM-SFFS. Additionally relevance of single features by calculation of information gain ratio is presented. As a basis of comparison we reduce dimensionality by PCA. We show extensive analysis of different classifiers within the named task. Among these are Kernel Machines, Decision Trees, and Bayesian Classifiers. Moreover we improve single classifier performance by Bagging and Boosting, and finally combine strengths of classifiers by StackingC. The database is formed by 2,114 samples of speech, and singing of 58 persons. 1,000 Music clips have been taken from the MTV-Europe-Top-20 1980-2000. The outstanding discrimination results of a working real-time capable implementation stress the practicability of the proposed novel ideas.

1. INTRODUCTION

Discrimination of polyphonic music and speech grew an important field of research, as audio-stream processing methods started to mature. For example automatic speech recognition applied to soundtracks [1] demands segmentation between music and speech parts prior to speech recognition. Furthermore processing radio broadcasts helps to retrieve only parts containing announcements of the D.J. [2]. However, we also aim to recognize parts of acapella monophonic singing in this work. Such can be applied within certain retrieval scenarios, where one wishes to find e.g. parts of an actor humming or singing in a motion picture or play.

In our case we use this discrimination for a music information retrieval user interface as introduced in [3]. It can be controlled by naturally speaking, singing or playing polyphonic audio clips into a microphone. The system at any time has to recognize in real-time the signal type in order to either forward the input to a speech recognition and natural language interpretation unit, a query by singing or finally a polyphonic music matching engine. The latter two differ substantially in preprocessing and features used. Intended use cases comprise among others music ordering call centers, a living room setup, where audio can be controlled by voice independently of the location by voice, and in-car devices allowing for hands- and eyes-free music selection. An example might be ordering a song heard on the radio by call: *"Hello, I'd like to get a CD with this song [holds the speaker close to the radio]. Please send me the whole new album of the artist."* Another person might utter in a most natural way to her car stereo: *"I'm looking for a song by [Interpret] and it goes like hmmmmm… [hums the melody]. Please choose the album mix and play it loud!"*

So far several works deal with the discrimination of polyphonic music and speech [1, 2, 4], while rather few work on the harder challenge of discrimination between speech and monophonic singing [5] or singing location [6], in our case even of the same person [7]. In these works rather low numbers of features have been considered, and selected by single feature relevance calculation instead of finding an optimal set which is also ideally suited for the target classifier. Classification itself is in general done with single classifiers. We strive to improve on this matter by use of ensembles.

The paper is structured as follows: In section 2 we describe the applied database in detail. Section 3 shows the general segmentation into clips. Sections 4 and 5 deal with extraction and selection of optimal acoustic features. In section 6 we discuss diverse classifiers and introduce

0-7803-9332-5/05/$20.00 ©2005 IEEE

ensemble variants. The final section discusses the results obtained and shows future directions.

2. DATABASE DESCRIPTION

We use our on request obtainable SHANGRILA corpus of speech and monophonic singing samples, as there is no common database available yet for this exact problem. 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 the retrieval interface as described above. 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 songs used in our music retrieval system or typical mainstream pop-music radio station sound.

3. SEGMENTATION

Prior to the discrimination of the signal type we split the continuous audio stream at significant changes in intensity. Likewise we achieve short clips of a few consecutive 20 ms frames based on the reasonable assumption that a change in signal type always includes a variation in intensity. Bi-state energy-threshold activity detection is applied for this task. The dynamically set threshold has to be exceeded, or respectively under-run for a set time interval to indicate the start or end of a new segment. Afterwards an adaptation to the new level takes place. The latter is set rather aggressive, in order to keep the cut sequences short.

Secondly a Support Vector classification is fulfilled applying Mel Frequency Cepstral Coefficients *(MFCC)* and δMFCCs for the discrimination of ambient noise [7]. MFCC have proven highly effective in the field of automatic speech recognition as they model the subjective pitch and frequency content of audio signals [8]. They are computed from the FFT power coefficients. These are filtered by a triangular band pass filter bank, which consists of 12 filters in our case. In Mel-frequency their interval is constant. The total frequency ranges from 0 Hz to 22,050 Hz. An advantage of this first decision is that MFCCs can be computed fast. Next static features as described in the ongoing are derived, if the cut out clip is decided as non-noise.

4. FEATURE EXTRACTION

In order to be able to discriminate the three classes we need to find adequate features characterizing the underlying acoustic signal. Additionally they ideally should not depend on the spoken or musical content itself. Finally they have to fit the chosen modeling by means of classification algorithms. As there are a high number of features generally suited for this task, it seems important to find the optimal feature set in view of maximum performance and generalization capability. As starting basis we extract a large set comprising of 276 features partly introduced in other works [1, 2, 4, 5, 6]. Such comprise Harmonic to Noise Ratio (HNR), spectral centroid, spectral roll-off point, spectral flux, and zero-crossing rate among others.

In a second step we strive to find an optimal set as described in the next chapter. As we cannot provide a detailed introduction of all chosen attributes in this paper we focus on the key ones. They rely mostly on pitch, energy, and spectral characteristics and the time signal itself. The contour of pitch is well-known for its capability to carry a large amount of information considering the perceptual difference of spoken and sung signals. In comparable works [5] the use of pitch information is also propagated. As pitch detection algorithm we use the auto-correlation-based faster AMDF as introduced in [7]. The values of energy are calculated by the logarithmic mean energy within a frame. The spectral features are known to be capable of discrimination between polyphonic music and the human voice.

5. FEATURE SELECTION

As we investigate an initial set of 276 features, dimensionality reduction seems a must considering real-time capability as extraction time can be saved. In general reduction of a feature set is often obtained by means of the well known Principal Component Analysis *(PCA)* and selection of the obtained artificial linear superposition features corresponding to the highest eigen-values. As such reduction still requires calculation of the original features we compare it to a real elimination of original features within the set in order to save computation time. However, we aim at an optimal set as a whole rather than a combination of stand alone high performance attributes. In the latter case redundancies are not recognized, and a higher number of features may be necessary to equal the performance of an optimized set. This may be achieved by so-called wrapper-based feature selection *(FS)* where once a search function and a wrapper, mostly the target classifier, need to be chosen.

We apply a Support Vector Machine *(SVM)* based Sequential Forward Floating Search *(SFFS)* [9], which is well known for its high performance. The search is performed by forward and backward steps eliminating and adding features to an initially empty set. As the relevance of attributes is largely discussed, we non-the-less provide the information gain of the features obtained by filter-

based FS. The following table shows our 10 highest ranked features with their according Information Gain Ratio *(IGR)* for the discrimination of speech/music/ singing. IGR is provided to give an impression of the single feature relevance.

Rank	Gain Ratio	Feature Description
1	0.8155	HNR mean
2	0.6741	Energy below 650 Hz
3	0.6529	MFCC1 std. dev.
4	0.6346	Energy below 250 Hz
5	0.6251	MFCC1 mean
6	0.5307	δRoll-Off-Point mean
7	0.4926	Zero-Crossing-Rate
8	0.4900	Spectral-Flux Mean
9	0.4777	F3 distance to F0
10	0.4769	Roll-Off-Point mean

Figure (1): Discrimination Speech/Music/Singing
IGR FS top ranks including Information Gain Ratio

In the next table the same ranking is shown focusing on the discrimination task between speech and singing.

Rank	Gain Ratio	Feature Description
1	0.8137	δF0 mean
2	0.6856	HNR mean
3	0.6470	Rate of voiced sounds
4	0.5704	Silence durations mean
5	0.5402	δMFCC2 mean
6	0.5378	MFCC1 mean
7	0.5170	SpecFluxStdDev
8	0.5075	δMFCC1 mean
9	0.4923	Spectral-Flux maximum
10	0.4536	Energy below 650 Hz

Figure (2): Discrimination Speech/Singing
IGR FS top ranks including Information Gain Ratio

As can be seen the feature sets for the discrimination between only speech and music differs as pitch and duration information plays a more important role than spectral information herein. The overall Gain Ratio order differed from the SFFS ranking order, which shows that SFFS optimizes a set as a whole. Likewise features with low Gain Ratio may occasionally be preferred in SFFS ranking as they complement a set.

The following figure shows feature selection by PCA FS, SVM SFFS, and IGR based FS in direct comparison. It can be clearly seen that PCA FS achieves lowest error rates at high reduction rates. However, SVM SFFS stays close to PCA FS while at significantly lower feature extraction effort. IGR clearly falls behind the other variants, as it only finds single best features. Only light improvement was observed using larger set sizes, while these cost higher extraction effort which may easily be crucial to real-time processing requirements.

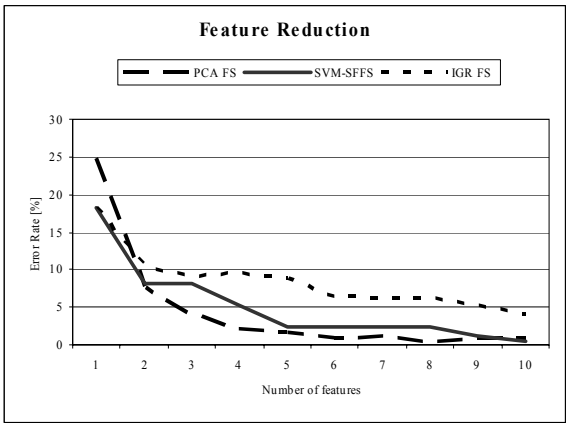

Figure (3): SVM-SFFS vs. PCA and IGR FS

6. ENSEMBLE CLASSIFICATION

With relatively small training sample sizes compared to the dimensionality of the data a high danger of bias due to variances in training material is present. In order to improve instable classifiers as neural nets or decision trees a solution besides regularization or noise injection is construction of many such weak classifiers and combination within so called ensembles. Two of the most popular methods are Bagging and Boosting [10].

Within the first random bootstrap replicates of the training set are built for learning with several instances of the same classifier. A simple majority vote is fulfilled in the final decision process.

In Boosting the classifiers are constructed iteratively on weighted versions of the training set. Thereby erroneously classified objects achieve larger weights to concentrate on hardly separable instances. Also a majority vote, but based on the weights leads to the final result. However, these methods both use only instances of the same classifier.

If we strive to combine advantages of diverse classifiers Stacking is an alternative. Hereby several outputs of diverse instances are combined. In [10] StackingC as improved variant is introduced, which includes classifier confidences e.g. by Maximum Linear Regression. It is further shown that by StackingC most ensemble learning schemes can be simulated, making it the most general and powerful ensemble learning scheme. One major question however is the choice of right base classifiers for the ensembles. In [10] two optimal set built of seven and four classifiers are introduced. However, the performance with the smaller set shows similar results at

less computational effort for training. We use a slightly changed variant of their set, which delivered better results.

In the following table results on the various tasks are presented with StackingC, Bagging, Boosting and selected base-classifiers are shown. However, we can provide only a very brief introduction of the latter in the ongoing. A comprehensive description is available in [10]. The major drawback of the firstly selected well known rather simple Naïve-Bayes *(NB)* classifier is the basing assumption that features are independent given class. Another rather trivial variant is a k-Nearest-Neighbor classifier based on Euclidean distance *(kNN)*. Support Vector Machines *(SVM)* show a high generalization capability due to their structural risk minimization oriented training. In this evaluation we used a couple-wise decision for multi-class discrimination and a polynomial kernel. As Decision Tree we chose an unpruned C4.5. In general these are a simple structure where non-terminal nodes represent tests on one or more features and terminal nodes reflect decision outcomes. The attributes are already weighted by their Gain Ratio.

Classifier	Error [%]
NB	2.35
kNN	1.46
SVM	0.58
C4.5	5.57
Bagging C4.5	4.40
Boosting C4.5	4.11
StackingC SVM NB C4.5 ND	**0.57**

Figure (4): Performances of single classifiers and ensembles for the overall discrimination

All tests have been carried out on the datasets described in section 2 by a three-fold stratified cross-validation [11]. Only mean performance is shown as the standard deviation throughout cycles never exceeded 1.5%. Only results with optimal parameter configuration are shown.

7. CONCLUSION

In this work we presented an approach to the discrimination of speech, monophonic singing, and polyphonic music. By use of feature selection techniques we presented an optimal feature set for this task selected out of 276 original features. Single feature relevance was shown by Gain Ratio computation. The single classifiers were all outperformed by the suggested ensemble classification. Among the latter StackingC was found most robust. A working implementation could be applied in real-time with the reduced feature set and overall error rates could be reduced to 0.57%. The weak classifiers could easily discriminate between speech and polyphonic music. However, for the correct discrimination between speech and monophonic singing more sophisticated algorithms were demanded. StackingC can again be reported as the best alternative hereon. Still this also requires high computational effort. Faster algorithms may therefore be preferred at a slight loss in accuracy. The proposed methods could be successfully integrated into a music retrieval system [3]. In our future work we aim at investigation of genetic feature generation and multi-task learning. Further more we consider combination of features on different timing levels.

8. REFERENCES

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

[2] J. Saunders: "Real time discrimination of broadcast speech/music," *Proceedings of the 1996 ICASSP*, pp. 993-996, 1996.

[3] B. Schuller, M. Zobl, G. Rigoll, M. Lang: "A Hybrid Music Retrieval System using Belief Networks to Integrate Queries and Contextual Knowledge," *Proceedings of the ICME 2003*, Multimedia Human-Machine Interface and Interaction I, Baltimore, MD, USA, Vol. I, pp. 57-60, 2003.

[4] W. Chou, L. Gu: "Robust Singing Detection in SpeechMusic Discriminator Design," *Proceedings of the 2001 ICASSP*, 2001.

[5] D. Gerhard: "Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing," *Journal of the Canadian Acoustical Association* 30:3, pp. 152-153, 2002.

[6] A. L. Berenzweig, D. P. W. Ellis: "Locating Singing Voice Segments Within Music Signals," 2001.

[7] B. Schuller, G. Rigoll, M. Lang: "Discrimination of Speech and Monophonic Singing in Continuous Audio Streams Applying Multi-Layer Support Vector Machines," *Proceedings of the ICME 2004*, Taipei, Taiwan, 2004.

[8] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *International Symposium on Music Information Retrieval*, 2000.

[9] P. Pudil, J. Novovičová, J.Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15/11, pp. 1119–1125, Nov. 1994.

[10] A. Seewald, *Towards understanding stacking – Studies of a general ensemble learning scheme*, PhD-Thesis, TU Wien, 2003.

[11] I. H. Witten, E. Frank, Data Mining, *Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, pp. 133, 2000.