Speech Music Discrimination Using Class-Specific Features

Thomas Beierholm
GN ReSound A/S,
Mårkærvej 2A, 2630 Taastrup, Denmark
tbeierholm@gnresound.dk

Paul M. Baggenstoss
Naval Undersea Warfare Center
1176 Howell St, Newport RI, 02841
p.m.baggenstoss@ieee.org

Abstract

In this paper the application of the class-specific features approach to classification is demonstrated for the problem of discriminating between speech and music. Feature extraction is class-specific and can therefore be tailored to each class meaning that segment size, model orders and the type of features used can be different for the classes. The performance of the discriminator is evaluated and an example of how classification is possible without training is given.

1. Introduction

The general problem of classification of audio signals has reached new levels of interest. Some of the applications of audio signal classification are speech/music classification [1], acoustical environmental classification [2][3], noise type classification [4] and musical genre classification [5]. The topic of audio signal classification is also interesting from a hearing instrument industry point of view. A typical modern digital hearing instrument has a suite of algorithms, each either tuned for a specific listening environment or has the possibility to be tuned for a specific listening environment. A method, e.g. [11], for automatically adapting a hearing instrument for various listening situations (silence, speech, noise, music, wind, etc.) would free users from manually having to change program using a pushbutton located on the hearing instrument, sometimes a task that is problematic for many hearing instrument users.

Typically, an audio signal classification system is build from a discriminative point of view. Features are extracted and cluster analysis, distances measures, entropy analysis or related methods are used to filter out those initially proposed features that are similar in all classes and hence don't provide any useful information for discriminating between classes. Based on the final set of features, probability density functions (PDFs) for the classes are estimated from training data. These class dependent density functions are defined on the same feature space making direct comparisons of the class likelihoods possible. The development of a discriminative audio signal classification system builds on experience, intuition and empirical findings. The advantages are that it is relative straight forward and quick to build a classifier and evaluate its performance. The disadvantage is that for difficult classification problems it might be impossible to find a suitable feature set that paves the way for the required performance.

A set of interesting audio classes from a hearing instrument point of view is speech, music, stationary and non-stationary noise. It has been found from previous work that discriminating between these classes with a very high degree of reliability is difficult. A few of the major problems in developing a high performance classification system for these classes are due to the huge variability within the classes and ambiguity (overlap) between the classes. To deal with these issues many features (much information) must be extracted from the input signal, a requirement that is in contrast to the demand of having a low dimensional feature vector so that class dependent density functions can be accurately estimated.

In this paper we apply a different classification approach to the task of developing an audio signal classification system, the class-specific features (CSF) approach [6]. For a given maximum feature dimension the class-specific approach opens up a possibility for using much more information (more features) when compared to the conventional approach. Each class gets its own separate branch in the classifier, incorporating feature extraction, class dependent likelihood evaluations and the computation of a data-dependent correction term. Furthermore, the class-specific classifier can be shown to be optimal in being equivalent to the Bayesian classifier formulated on the raw data, it provides guidance for selecting features for each class and can provide classification without the need for training or learn density functions from data (called the feature selectivity effect [6]).

We have concentrated on the classification of speech and music and will apply the CSF approach to the classification of these two classes. It may seem strange to start the design of a fully capable audio signal classifier

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

with only two classes but due to the parallel structure of the CSF classifier all classes don't have to be considered initially. In this case, after making branches for the speech and music classes and fine-tuning these branches, adding branches can be done without changes in the already designed and fine tuned branches.

In section 2 the CSF approach is briefly introduced. In section 3 the speech and music branches are described and in section 4 the performance of the 2-branch classifier is evaluated. Finally in section 5 a discussion and directions for further research are given.

2. The CSF approach

Central to the CSF approach is the PDF projection theorem [6]. This theorem can be used for evaluation of a higher dimensional PDF, $p(\mathbf{x}|H_j)$, defined on a N dimensional raw data space under the hypothesis H_j, based on a projection using a lower dimensional PDF, $p(\mathbf{z}_j|H_j)$, defined on a M dimensional feature space. The feature vector $\mathbf{z}_j=T_j(\mathbf{x})$ is class-specific and hence computed from a class dependent feature transformation T_j. Using the PDF projection theorem we can write

$$\begin{aligned} p(\mathbf{x} \mid H_j) &= \frac{p(\mathbf{x} \mid H_{0,j})}{p(\mathbf{z}_j \mid H_{0,j})} p(\mathbf{z}_j \mid H_j) \\ &= J(\mathbf{x}, T_j, H_{0,j})\, p(\mathbf{z}_j \mid H_j) \end{aligned} \quad (1)$$

where $H_{0,j}$ denotes a class dependent reference hypothesis. We will also look at the PDF projection theorem as a factorization of a high dimensional PDF into two terms. The first term is the J function, defined in (1) that is a function of the raw input data and is derived from the choice of reference hypothesis and feature extraction transform. The second term is the PDF of the class-dependent features that in practical use needs to be learned from data. Two things must be fulfilled in order for the projection in (1) to be optimal. The extracted class-specific features, $\mathbf{z}_j$, must be statistical sufficient for the problem of choosing between H_j and $H_{0,j}$, where $H_{0,j}$ is a class-dependent reference hypothesis and the class feature density, $p(\mathbf{z}_j|H_j)$, must be the true density. For practical use we must be satisfied with class-specific features that are approximately sufficient for the problem and class-specific feature density functions that only approximate the true densities and hence the projection in (1) and the classification rule in (2) are only approximately optimal. Using the PDF projection theorem the Neyman-Pearson classification rule can be written as

$$j^* = \arg\max_j \left[\frac{p(\mathbf{x} \mid H_{0,j})}{p(\mathbf{z}_j \mid H_{0.j})} p(\mathbf{z}_j \mid H_j) \right], 1 \le j \le K \quad (2)$$

where K designates the number of classes. We see that we need to determine the distribution of the raw data and the class-specific features under a possibly class dependent reference hypothesis, $H_{0,j}$. This reference hypothesis can theoretically be any reference hypothesis meeting some mild conditions [6], but the best choice is a reference hypothesis that (a) makes evaluation of $p(\mathbf{z}_j|H_{0,j})$ tractable, and (b) admits a simple sufficient statistic when testing against H_j. Because it is tractable and appropriate as well, we choose to use the same reference hypothesis for both classes, denoted simply by H_0. We let H_0 consist of zero mean and unit variance i.i.d. gaussian random variables. In general it's straight forward to evaluate $p(\mathbf{x}|H_{0,j})$, whereas the main difficulty in using the CSF approach relates to evaluating $p(\mathbf{z}_j|H_{0,j})$. As the last point mentioned here, we note that using only the first term in (1) it's actually possible to obtain classification performance without having to estimate the class density functions depending on the application and the required classification performance.

3. Speech and music branches

In order to determine a suitable set of features for each of the classes we take on a descriptive view and look for features that describe examples from each class to the point where it's actually possible to 'reconstruct' the original examples. The idea is that if we can reconstruct the original examples so that by listening it's possible to recognize what is being said or what piece of music is being played then the features convey most of the important information and hence are approximate sufficient.

As described in [6] the CSF classifier is suited for modularization. This is also reflected in Figure 1 where a block diagram of the 2-branch CSF classifier is illustrated. All the modules used in Figure 1 and many more can be found in [9]. Notice that each branch in Figure 1 is broken into separate modules which are arranged in cascade fashion. This is a result of the chain-rule, the mathematical decomposition of the PDF projection theorem into a series of transformations [6]. The computation of the J-function terms for the used features have been worked out in [7][8].

3.1. Speech branch

For the speech class it's well known that an all-pole model excited by either white gaussian noise (unvoiced) or a pulse train (voiced) is a good model, especially for clean speech. It was found that an 8th order LPC model on a segment size of 256 samples with pitch information could reconstruct the signal to the point where there was no doubt that the features represented speech and most of the

words were understandable. The synthesized speech sounded very harsh and contained many artifacts especially due to the completely white residual signal and zero samples overlap.

The speech class model is specified by the auto-correlation function (ACF) values at lags 0 to 8, $\mathbf{r}_8$ and the ACF value and index corresponding to the pitch lag, (r_p,i_p). The circular ACF is used in order not to violate the assumed independence between segments. The first 2 modules in the speech branch in Figure 1 computes these features and the corresponding normalization term. Instead of using the feature set $\mathbf{z}_{speech}=[\mathbf{r}_8,r_p,i_p]$ we will use $\mathbf{z}_{speech}=[\rho,\mathbf{k}_8,r_p,i_p]$ which can be found from a series of invertible transformations which the last modules in the speech branch account for. This feature set, as also noted in [10], is more suitable for gaussian mixture PDF estimation. $\mathbf{k}_8$ denotes Log-Area Ratio (LAR) coefficients and $\rho=\log(r_0)$. We have factorized the PDF of $\mathbf{z}_{speech}$

$$p(\rho,\mathbf{k}_8,r_p,i_p \mid H_s) \approx p(\mathbf{k}_8,r_p \mid H_s,\rho,i_p)p(\rho \mid H_s)p(i_p \mid H_s)$$

where H_s denotes the speech hypothesis. $p(\rho|H_s)$ and $p(i_p|H_s)$ were given uniform distributions. This way the PDF estimation problem becomes more manageable.

3.2. Music branch

Typically, music is characterized by narrow band harmonic components that show up as peaks in the spectrum. For this reason an experiment was conducted where the P largest bin values and indexes were extracted from the power spectrum block-by-block on a segment size of 128 samples. Remarkably for P=4 some music (especially classical music with a few instruments active) can be reconstructed to the point where the music can be recognized. For rock music it's harder to recognize the music, the peaks in the spectrum are not as distinct as for classical music. It was found that P=8 resulted in relative good reconstruction results and was used in the classification experiments.

Thus the feature set for the music class is $\mathbf{z}_{music}=[\mathbf{p}_8,\mathbf{i}_{p8},\check{r}_0]$, where $\mathbf{p}_8$ designates the 8 largest bin values in the power spectrum, $\mathbf{i}_{p8}$ the corresponding indexes and $\check{r}_0$ the residual energy which as noted in [8] ensures that the feature set is a sufficient statistic for inputs with unknown variance. This feature set and the corresponding normalization term are computed by the first 2 modules in the music chain in Figure 1. It was found useful to use $\mathbf{q}_8=\log(\mathbf{p}_8)$ and $\varepsilon=\log(\check{r}_0)$ to obtain a better PDF estimate using a gaussian mixture HMM. As for the speech branch the PDF of $\mathbf{z}_{music}$ is factorized

$$p(\varepsilon,\mathbf{q}_8,\mathbf{i}_{p8} \mid H_m) \approx p(\mathbf{q}_8 \mid H_m,\varepsilon,\mathbf{i}_{p8})p(\varepsilon \mid H_m)p(\mathbf{i}_{p8} \mid H_m)$$

where H_m denotes the music hypothesis. Furthermore, $p(\boldsymbol{i}_{p8}|H_m)$ was approximated by factorizing it into its 8 marginal distributions, each was given a uniform distribution, ε was also given a uniform distribution.

3.3. Implementation issues

Saddlepoint approximation (SPA) is used in 2 of the modules in Figure 1 to compute the normalization term. Some convergence problems were experienced with the SPA when the lag corresponding to the pitch was too high (>120 samples) or when extracting too many features for the music class. In order to prevent convergence problems a pre-whitening filter was designed and used in all experiments.

As noted in section 2 using the J function term of the PDF projection alone can provide a possibility to design a classifier without the need to estimate or train any PDFs. In fact, the chains can be deliberately designed for this purpose. If the last module (module_log) in the music chain in Figure 1 is left out, reasonable classification performance is obtained using only the J function term, hence providing untrained classification. Because the branches in the CSF classifier are not restricted to be the same, decisions or classification must be made in intervals corresponding to the largest segment size (not reflected in Figure 1).

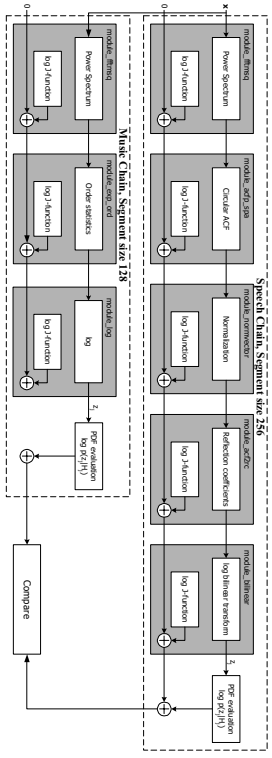

Figure 1, Block diagram of speech and music branches.

4. Evaluation

Before evaluating the approach on some corpus of speech and music, the branches were verified by acid tests[6]. These tests are important as they immediately reveal if there are issues in the branches, they are especially useful for checking the correctness of the computation of the normalization terms.

To evaluate the approach, speech and music samples were collected. From a classical music CD with 12 tracks 1 minute of music was extracted from each track and split into 2 seconds samples leading to a total of 360 classical music samples. From a pop music CD with 20 tracks 1 minute of music was extracted from each track and split into 2 seconds samples leading to a total of 600 pop music samples. From 11 different speakers a total of 255 2 seconds samples were obtained. All sounds were recorded in mono with a 16kHz sample rate. The class-specific density functions were estimated using gaussian mixture HMMs. 130 speech samples and 75 music samples were used in the density estimation. For each 2 seconds sound sample the number of segments (256 samples) classified as speech or music were counted and the sound sample was classified as belonging to the class having most counts. Classification performance is summarized in Table 1.

An experiment was made where only the correction term was used for the classification. In this case all speech samples were classified correct and only 27 samples out of 600 pop music samples were misclassified. However, only half of the classical music samples were classified correct.

	Trained	Untrained
Speech	100%	100%
Music	100%	80%

Table 1, Classification performance.

5. Discussion

Beside further fine-tuning of the chains, future work will be directed into extending the 2-branch classifier with more classes. For instance a stationary noise class (containing e.g. traffic noise) could be represented with a relative large segment size and a low order AR model. The feature selectivity effect of the CSF approach is an effect which can turn out to be very useful in practical use leaving out the burden to learn density functions, saves computations and potentially making calibration provisions unnecessary. The feature selectivity of the speech chain used herein works very well whereas the feature selectivity of the music chain is less distinct. It is also worth noting that when having designed a CSF classifier the step needed to develop an optimal segmentation algorithm is not that big. For instance, a HMM can work on top of the CSF classifier or the learned density functions can be replaced with one big HMM with multiple observation spaces [12] using the Viterbi algorithm to perform the segmentation.

6. References

[1] Eric Scheirer and Malcolm Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", Proc. of ICASSP, Vol.2 , 21-24 April 1997 p.1331-1334

[2] Michael Büchler, "Algorithms for Sound Classification in Hearing Instruments", PhD Thesis at Swiss Federal Institute of Technology, Zurich, no 14498, 2002

[3] Peter Nordquist and Arne Leijon, "Automatic classification of listening environments in a generalized adaptive hearing aid", submitted for publication in JASA.

[4] Paul Gaunard, Corine Ginette Mubikangiey, Christophe Couvreur and Vincent Fontaine, "Automatic Classification of Environmental Noise Events by Hidden Markov Models", Facultè Polytechnique de Mons, Belgium

[5] G. Tzanetakis, G. Essl and P. Cook, "Automatic Musical Genre Classification of Audio Signals", Proc. Int. Symposium on Music Inform. Retriev. (ISMIR), p.205-210, Oct 2001

[6] Paul M. Baggenstoss, "The PDF projection Theorem and the Class-Specific Method" IEEE Trans. on Signal Processing, Vol. 51, No. 3, p.672-685, March 2003.

[7] Steven M. Kay, Albert H. Nuttall, and Paul M. Baggenstoss, "Multidimensional Probability Density Function Approximations for Detection, Classification and Model Order Selection", IEEE Trans. on Signal Processing, Vol. 49, No. 10, p.2240-2252, October 2001.

[8] Albert H. Nuttall and Paul M. Baggenstoss, "Joint Distributions and Two Useful Classes of Statistics, with Applications to Classification and Hypothesis Testing"

[9] http://www.npt.nuwc.navy.mil/Csf/index.html

[10] Paul M. Baggenstoss and Heinrich Niemann, "A theoretically Optimal Probabilistic Classifier Using Class-Specific Features", Proceedings of ICPR, September 2000.

[11] R.A.J de Vries, B. de Vries, "Towards SNR-loss restoration in digital hearing aids", Proc. of ICASSP, Vol.4 , 13-17 May 2002,p.4004-4007

[12] Paul M. Baggenstoss, "A Modified Baum-Welch Algorithm for Hidden Markov Models with Multiple Observation Spaces", IEEE Trans. on Speech and Audio Processing, Vol.9, No.4, May 2001