Investigations into Features for Robust Classification into Broad Acoustic Categories

Joerg Schmalenstroeer[1], Markus Bartek[2], Reinhold Haeb-Umbach[3],

Fachgebiet Nachrichtentechnik Universität Paderborn, 33095 Paderborn, Deutschland

[1] *schmalen@nt.uni-paderborn.de* [2] *bartek@mail.uni-paderborn.de* [3] *haeb@nt.uni-paderborn.de*

Abstract

In this paper we present our experimental results about classifying audio data into broad acoustic categories. The reverberated sound samples from indoor recordings are grouped into four classes, namely speech, music, acoustic events and noise. We investigated a total of 188 acoustic features and achieved for the best configuration a classification accuracy better than 98%. This was achieved by a 42-dimensional feature vector consisting of Mel-Frequency Cepstral Coefficients, an autocorrelation feature and so-called track features that measure the length of "traces" of high energy in the spectrogram. We also found a 4-feature configuration with a classification rate of about 90% allowing for broad acoustic category classification with low computational effort.

Introduction

In many applications it is desirable to classify an incoming audio stream into the broad acoustic classes speech, music, other acoustic events and stationary noise, before further processing. For example, automatic speech recognition can be improved if only those audio data are forwarded to the decoder that indeed contain speech, smart environments can benefit from such a classification to infer the user's activity and hearing aids often incorporate different processing algorithms and settings depending on the type of input signal.

While humans can classify signals into the aforementioned categories usually effortlessly, a classification by machine faces the problem of large intra-class variability: Music being an artistic composition of rhythmic and harmonic acoustic events can be greatly diverse, and the category of other acoustic events, containing in our case non-stationary sounds audible in a home environment, such as door knocking, ringing, the sound of foot steps etc., is similarly vaguely defined.

The paper is organized as follows. At first we will briefly introduce the database and the features we used in the experiments. Subsequently, the experimental results are presented and discussed. Finally, some conclusions are drawn and an outlook is given.

Database

We compiled a database of 12 hours length using material from the CHIL database [1], web radio recordings, and own recordings. To ensure realistic data from inside of a home, non-reverberated recordings were convolved with artificially generated room impulse responses. The speech data was taken from the CHIL speaker recognition task and the events from the CHIL acoustic event detection database. The music category consists of recordings from different music genres ranging from classic to rock. Examples of members in the stationary noise category are the noise produced by vacuum cleaners and fans. The recorded segments have a duration of 2 seconds (e.g. speech) or less (e.g. door slam).

We investigated both time domain and frequency domain features. For this purpose the segments are either processed by a short term analysis (upper index (ST), 40 ms) or a long term analysis (upper index (LT), 160 ms). After the feature extraction step Gaussian mixture models with up to 128 mixtures are trained.

Frequency domain features

In this category belong the 13 Mel-Frequency Cepstral Coefficients of the ETSI Advanced Front End and their first and second order derivatives [2]. We extend this feature vector with a maximum autocorrelation value [3] which is an indicator for the periodicity of the signal. The feature vector is denoted by AFE.

From the spectrogram we calculated the spectral flux (SF), the spectral distribution (SD) features (spectral centroid μ_{SD}, spread σ^2_{SD}, skewness Γ_{SD}, kurtosis Ψ_{SD}) as well as the track features. The latter searches the spectrogram for prominent tracks [4] and the maximum length track ($\max_T$) as proposed in [5].

As reported in [6] the modulation spectrum of speech has a characteristic energy maximum at a frequency of 4 Hz. The feature measuring the 4 Hz modulation energy can be derived from a wavelet transform (index 4Hz). It is intended to support the discrimination between samples containing speech and non-speech samples.

Additionally we calculate the mean, the variance and the higher-order statistics kurtosis and skewness of the frequency domain features, since these central moments delivered promising results in other classification tasks.

Time domain features

In the time domain (TD) we mainly focused on zero crossing rate based features, since these features are not computationally demanding and good results were reported on music/speech discrimination tasks [6]. In addition to the number of zero crossings per block (ZCR) we investigated the use of the distances between zero crossings

(ZCD) and the difference between adjacent local maxima (zero crossing amplitude, ZCA).

Further we calculated the aforementioned features from the first (lower index Δ) and second order (lower index Δ^2) derivatives of the signal. In accordance with the frequency domain features we also investigated the first four central moments of all time domain features.

Experimental results

At first we tested the features independently on their ability to discriminate between the four categories (see table 1). The best frequency domain feature is the spectral spread σ^2_{SD}, followed by the energy E_{SD}, the second central moment of the spectral spread $\mathrm{var}\left[\sigma^2_{\mathrm{SD}}\right]$ and the variance of the 4 Hz modulation energy $\mathrm{var}\left[E_{\mathrm{4Hz}}\right]$.

Feature	classification rate [%]				
	Speech	Music	Events	Noise	avg.
σ^2_{SD}	82.18	88.50	65.61	99.46	82.66
E_{SD}	82.52	88.25	64.59	94.36	81.19
$\mathrm{var}\left[\sigma^2_{\mathrm{SD}}\right]$	80.76	86.68	61.19	99.76	80.66
$\mathrm{var}\left[E_{4Hz}\right]$	79.47	88.42	70.80	82.44	79.56
$\sigma^{2(\mathrm{ST})}_{\mathrm{TD}}$	81.65	87.93	64.70	99.56	82.16
$E^{(\mathrm{ST})}_{\mathrm{TD}}$	83.03	87.03	63.12	99.91	82.08
$E^{(\mathrm{LT})}_{\mathrm{TD}}$	81.49	86.52	60.92	99.83	80.74
$\sigma^{2(\mathrm{LT})}_{\mathrm{TD}}$	80.40	86.56	61.46	99.82	80.64
$\mu^{(\mathrm{ST})}_{\Delta\mathrm{ZCA}}$	80.14	90.04	69.73	85.94	80.58
$\mu^{(\mathrm{LT})}_{\Delta^2\mathrm{ZCA}}$	79.16	91.24	61.98	89.75	79.17
Combined	84.32	90.09	74.40	87.56	84.09

Table 1: Classification rates of the best 10 features

Of the time domain features the energies $E^{(\mathrm{LT})}_{\mathrm{TD}}$, $E^{(\mathrm{ST})}_{\mathrm{TD}}$ and the second central moments $\sigma^{2(\mathrm{LT})}_{\mathrm{TD}}$, $\sigma^{2(\mathrm{ST})}_{\mathrm{TD}}$ delivered the best results. Also the mean of the zero crossing amplitude feature gathered from the first order derivative $\mu^{(\mathrm{ST})}_{\Delta\mathrm{ZCA}}$ and the mean of the ZCA feature retrieved from the second order derivative $\mu^{(\mathrm{LT})}_{\Delta^2\mathrm{ZCA}}$ can be found within the group of the best features.

Since some of the features listed in table 1 describe similar statistical properties it is not surprising that their combination does not result in a significantly better classifier. On the contrary we found other, individually less discriminative features, whose combination clearly outperformed the combination of the 10 best individual features, see table 2. This 4-feature classifier achieves a classification rate of 90.43% (spectral flux of first derivative $SF^{(A)}_{\Delta}$, zero crossing amplitude on second derivative $\mu^{(\mathrm{ST})}_{\Delta^2\mathrm{ZCA}}$, mean of tracks μ_T, mean of statistical moments $\mu^{(\mathrm{LT})}_{\mathrm{TD}}$). We found this set by taking the features with the highest classification results per category.

Despite the fact that the four category discrimination problem was our main research target, we also tested the features on their ability to discriminate only two of the classes (e.g. music/speech). Here, the mean of the *ZCA* features on the first derivative $\mu^{(\mathrm{ST})}_{\Delta\mathrm{ZCA}}$ delievered the best results with a classification rate of 96.39%, followed by the variance of the 4 Hz modulation energy $\mathrm{var}\left[E_{4Hz}\right]$ with 92.57%. The *ZCR* reached a classification rate of only 78.29%.

Feature	classification rate [%]				
	Speech	Music	Events	Noise	avg.
$SF^{(A)}_{\Delta}$	91.50	87.70	26.62	86.95	70.22
$\mu^{(\mathrm{ST})}_{\Delta^2\mathrm{ZCA}}$	77.57	93.13	57.14	90.81	78.00
μ_T	52.77	76.37	89.78	13.27	59.83
$\mu^{(\mathrm{LT})}_{\mathrm{TD}}$	35.07	85.45	81.93	100.00	75.41
Combined	91.14	89.59	91.41	89.56	90.43

Table 2: Classification results of the best 4-feature classifier

Feature	classification rate [%]				
	Speech	Music	Events	Noise	avg.
AFE	99.52	96.22	97.41	99.87	98.26
AFE, σ^2_T	99.44	97.45	98.24	99.80	98.73
AFE, $\max_T$	99.46	97.45	97.90	99.77	98.65
AFE, μ^2_T	99.43	97.23	97.91	99.74	98.58
AFE, $\Psi^{(LT)}_{\Delta^2\mathrm{TD}}$	99.49	96.67	97.51	99.95	98.41

Table 3: Combination of AFE with one single feature

Finally, we tested the combination of the AFE features with one additional feature (table 3). Here the track features (variance of track length σ^2_T, maximum track $\max_T$, mean of tracks μ^2_T) delivered the best results. Of the time domain features the kurtosis of the second order derivative $\Psi^{(LT)}_{\Delta^2\mathrm{TD}}$ reached the highest error reduction.

Conclusions

We investigated the use of time domain and frequency domain features for a classification into broad acoustic categories. The proposed zero crossing amplitude feature showed remarkably good results on this task. Further we showed that the combination of the strongest features does not lead to a superior classifier, whereupon we found a four feature classifier with weak single features which delivered excellent classification results at low computational complexity. The best results were obtained with a combination of Mel-Frequency Cepstral Coefficients and track features. Our future work will focus on the task of selecting and combining features from the large set of features gathered in this work.

References

[1] A. Temko et al.: CHIL - Acoustic Event Detection; `http://chil.server.de/`, 2005

[2] ES 202 212 V1.1.1: Advanced front-end feature extraction algorithm; `http://www.etsi.org/`

[3] B. Wildermoth, K. Paliwal: Use of voicing and pitch information for speaker recognition, SST'00, 2000

[4] T. Quatieri, R. McAulay: Speech transformations based on a sinusoidal representation. IEEE Trans. on Acoustics, Speech and Signal Processing 34, 1986

[5] S. Ghaemmaghami and J. Shirazi: Audio Classification Based on Sinusoidal Model: A New Feature; TENCON, 2008

[6] E. Scheirer, M. Slaney: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. ICASSP '97, 1997