Acoustic Scene Classification

Modelling The Human Auditory System

MUS-15

Marc-Christoph Gerasch

July 11, 2015

1 Introduction

Imagine yourself standing on the platform of a train station, while waiting for your train to arrive, you are listening to the sounds around you. People are hurrying around, talking, trolleys clattering on the floor, announcements from the platform speakers, everything sounds familiar and tells you, you are on the platform of a train station. But what if you were not on the platform, instead you are simply listening to an audio record which sounds like a train station? From your experience, you can allocate the sounds to a train station. Now, what happens if you assign this task to classify sounds of a setting to a computer? This is called Acoustic Scene Classification (ASC), and is a subcategory the field of Computational Auditory Scene Analysis (CASA)[6]. The classification of the scene is often done by analyzing so called events, which are sounds, particular occurring in this kind of scene. The analysis of single events is called Acoustic Event Analysis, and is an own research field but, as it is also part of ASC, the distinction between the two fields is often a bit blurred[1]. The idea to let a computer do scene classification was first stressed by Cherry in 1953. He stated:

> 'One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others ... we may call it 'the cocktail party problem.' No machine has yet been constructed to do just that.' [8]

The problem derived from the field of automatic speech recognition in noisy environments, which is one of the common applications of ASC[8]. Humans are extraordinary good in this task, while for machines there is a lot of room for improvement. The best algorithms only hit the mean human accuracy[1]. Other applications are digital hearing aids, which adapt their amplification according to the environment[4], automatic music transcription[8], as well as context aware applications like robots. In the course of time, different approaches were employed to overcome the problem of CASA. Most of these approaches use statistic based methods to classify the scene. Since humans are very good in ASC, some approaches try to imitate the human auditory system to a certain extend. One of the current examples of each approach is presented and qualitatively compared in this paper.

2 From history to the state of the art

Although speech recognition already started way back in 1932 at Bell Labs, it took 20 years until Cherry stressed the cocktail party problem, and another 40 years, until Bregman in 1990 published the book 'Auditory Scene Analysis' and laid the foundation for the research in the field to the present day[7][8]. During the 90s the development of digital hearing aids pushed the research in ASC, until in 1997 Sawhney and Maes from the Massachusetts Institute of Technology (MIT) Media Lab, implemented the first approach exclusively for ASC, employing neural networks and nearest neighbor classifiers[1]. Only a year later, the MIT Media Lab recorded

samples for evaluation and brought up another approach, using Hidden Markov Models[1]. Not long after that, Mel Frequency Cepstral Coefficients, employed by Eronen et. al. to describe the local spectral envelope of audio signals, known for speech analysis, were applied to ASC and performed very well. As the number of approaches steadily rose, a more general, larger evaluation database was necessary. Due to the lack of specific databases for ASC, in 2003 the Text Retrieval Conference Video Retrieval Evaluation (TRECVID), originally created for video analysis, was chosen and remained the standard in evaluation to the present day[3]. Simultaniously the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics built up a new database for their Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE challenge), which sets the state of the art evaluation system for approaches in Acoustic Scene Classification[3].

3 Technical Methods

In this section the main technical methods, used by the approaches, are explained in a non mathematical way. The two different approaches, explained later in this paper, try to solve the problem of ASC in completely different ways. Still some of the physical methods or algorithms are used by both approaches. The methods can be divided into two groups, which are feature extraction and classification.

3.1 Feature Extraction

Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are probably the most popular features used in ASC. They were originally developed for speech recognition, which explains the usage of the Mel-Scale, the perceptual scale of pitches. Cepstral is an anagram for spectral, which implies the inverse nature of the cepstrum. MFCC are a combination of discrete cosine-, logarithmic- and Fourier-transformations, which provide the possibility to separate the vocal excitation (pitch), from the vocal tract (formants). This separation makes it possible to identify the same sound, e.g. a word, even though it is produced at different pitches. Particularly in terms of source independent sound classification, MFCC are a valuable tool. [5][1].

Fundamental Frequency F_0

Every sound, if not created by a sinus wave generator, has certain harmonics, which are specific to the sound source. These specific harmonics are the reason why every voice or instrument has an individual sound, although producing the tone on the same frequency. If these harmonics are connected, we can estimate fundamental frequences from which we can then derive features. In this way different sound sources or events may be distinguished, as certain harmonics correspond to them[6][1].

Filters

The audio clips are recorded using a high quality wave form audio format which provides 44100 Hz bandwidth. If we want to imitate the human ear, which can only perceive sounds from 20Hz to 20kHz, it would be reasonable to dismiss frequencies above 20kHz. For this purpose, band-pass filters, which only block or let through certain frequencies, can be a valuable tool. Although, the frequencies do not necessarily need to be blocked or let through completely, they may be suppressed or highlighted to a certain extend.

Energy-Related Features

Depending on the recorded scene, some frequency bands are more prominent than others. An analysis of the energy or amplitude for certain sub-bands, in comparison to the overall energy, points out those prominent frequency bands[1]. A high intensity for the frequency bands of 362 - 483 Hz or 410 - 547 Hz could e.g. be a hint to the German 'Martinshorn', the siren used in ambulance, fire brigade or police

vehicles[9].

3.2 Classification

Latent Perceptual Indexing

Latent Perceptual Indexing (LPI) is used to figure out the underlying key events classifying the audio clip. It is derived from Latent Semantic Indexing used for text analysis, where it is used to extract the conceptual content of a text, by establishing associations between terms that occur in similar contexts. Despite being employed to analyze huge amounts of data, it needs a lot of training to be accurate, too[2][6].

Support Vector Machine

Support Vector Machines (SVM) are binary classifiers. In explanation, they are supervised learning models, which sort feature vectors in two classes by comparing them. The two classes are divided by the support vectors, forming a line, which divides all vectors with the largest gap possible (see fig. 1). Following the training, new events are classified by falling on either side of the line. In ASC often more than two classes should be divided, leading to the need of multiclass SVMs. In this case, the dividing line is a higher dimensional hyperplane[1].

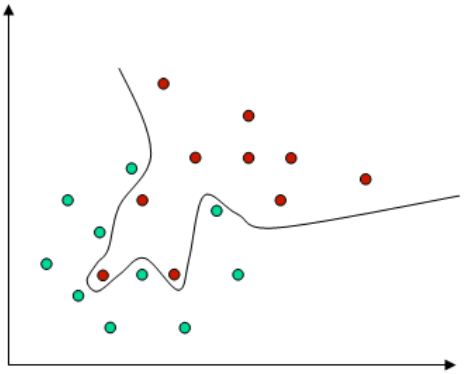

Figure 1: 'Divided vectors by a SVM'; Ennepetaler86 from www.wikipedia.org

Nearest Neighbor Classifier

The idea of a Nearest Neighbor Classifier (NN) is to build a set of feature vectors representing a class, and then compare undefined vectors to this set. To classify an undefined vector, the 'nearest neighbor' to the undefined vector, representing a distinctive class, is chosen to identify the undefined vector. The undefined vector is then assorted to this class of vectors. Usually a k-NN Classifier is used to reduce the error rate. In this case the k closest vectors are found, the most common class is defined by majority vote and assigned to the undefined vector[1].

Neural Networks

Neural Networks are directed graphs containing nodes (neurons) and edges (synapses), as well as a set of adaptive weights. The graph is build using a learning algorithm by adding nodes, edges or changing the weight of the edges. A network may contain a number of layers depending on the complexity of the problem. A 3 layer neural network for example would have a layer of input neurons, connected via synapses to a layer of hidden neurons, again connected via synapses to a layer of output neurons. The functionality of a Neural Network is dependent on the implementation of the learning algorithm[2].

Majority Vote

The audio clips are often windowed to ease the feature extraction and classification. Consequently, an overarching classification is derived from the classified windows by majority vote. Majority Vote simply sums up the number of frames for each class, and the most common class is assigned to the whole audio clip. Additionally, there is the possibility to weight certain frames and achieve a weighted Majority Vote[1].

4 Comparing human and statistic based approach

The main difference between statistic and human based approaches, rests in the link

between the features and the classification of the belonging scene[1]. Statistic approaches use statistical methods such as LPI or SVM, which match the events to the scene. A human approach instead would use high-level grouping based on experience, which links events together to event sets, typical for the scene[7]. The brain provides humans with many features like attention, filtering, highly linked storage networks, as well as the possibility to locate sound sources via our two ears. This is called binaural hearing, and is important for our ability to focus on one sound source amongst others[8]. Statistic based approaches make no use of this feature. For example the movement of a sound source is not traced, since this information is not used for low-level grouping, which links cues e.g. trough common fate. Waves starting at the same time, or frequencies that are harmonics of their base frequency, are probably from the same sound source, and therefore share a common fate[7]. To classify a scene correctly, every possible event has to be evaluated, as in the end a majority vote decides what kind of scene it is. Thus all data available is analyzed in a brute force manner, which results in a high computational effort. A main human ability is to focus on a single sound source as stated by Cherry. Implementations try to imitate this ability and the frequency-dependent perception of the human ear to reduce the amount of data computed[2].

The two approaches presented in this section follow completely different fundamental concepts even though, the main technical methods are similar. The first approach derives from a statistical background, was one of the most accurate approaches developed during the DCASE challenge, and is therefore an example for the state of the art in Acoustic Scene Classification. The second approach is an example for a human mimicking system, and was presented in the IEEE Workshop on Multimedia Signal Processing in 2009. Since it used a different Database for evaluation, a qualitative, instead of a quantitative comparison of the two approaches, is applied.

4.1 Statistic Based Approach by Geiger et. al [6]

The idea of the statistic based approach by Geiger et. al. was a large-scale audio feature extraction of nine different feature types, using the 'The Munich Open-Source Large-scale Multimedia Feature Extractor'. The extracted features are **MFCC 0-25** (the number refers to the frequency bands used), zero crossing rate, spectral flux, centroid, relative position of spectral maximum and minimum, **logarithmic energy**, **F0 (subharmonic summation (SHS)** and F0 envelope (probability of voicing). The main (written in bold letters) features have been explained in more detail in section 3. For the extraction, the binaural source files were mixed down to mono and cut into overlapping windows of a few seconds length. The windows are used to overcome the problem of non stationary scenes. Ideally, each window contains one event in the scene, allowing a distinct classification. After the feature extraction, different methods for classification were applied. An implementation of LPI did not supply satisfying results, since the training data were not satisfying. In the implementation, each record is represented by one single vector, resulting in a total amount of 100 vectors for the comparison. In contrast, the SVM approach windows the record. Each window is represented as a vector, resulting in a larger amount of comparable vectors, although depending on the window size. This method ignores the composition of events in the scene, so that each event is classified on its own. Instead a non weighted majority vote gives the overall result for each record.

The Database contains 10 recordings of 30 seconds length, for each of the 10 classes giving a 100 recordings in total. The audio clips were made with binaural microphones on the ears of a person. Systems were trained on an openly available training data set and evaluated on a secret evaluation data set. The contained scene classes are namely bus, busystreet, office, openairmarket, park, quietstreet, restaurant,

supermarket, tube and tubestation.

During the experimental phase, different set ups were used to find the best combination of methods. It was preliminary known, that MFCC delivered a good result in previous works, therefore MFCC was tried as single feature group, versus MFCC and all the other features together. The results for the LPI method reached a 46% accuracy using all features except for windowing. We would achieve similar results using SVM in combination with all features except for windowing, too. Further, including windowing allows for a higher accuracy of SVM. LPI would need a higher amount of training data to achieve the same result. Using the windowed audio clips, SVM gave a solid 68% accuracy using only MFCC features and 73% using all features. Therefore MFCC features indeed perform very well. On the evaluation data set the performance somehow dropped to 69% just a few percent below the actual winner of the DCASE challenge.

4.2 Human Based Approach by Kalinli et al. [2]

The development of Acoustic Scene Analysis derived from psychological research in the 1980s[7]. David Marr, psychologist, neuroscientist and founder of computational neuroscience, published his Book *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information* in 1982 and laid the foundation for all human based approaches[8].

Kalinli et. al. tried to develope an ASC system focusing on the reduction of computational effort. The core of their model is a method called LISA namely **L**atent **I**ndexing using **Sa**liency. LISA is based on LPI and mainly implements the human attention, but also covers other parts of the human auditory system.

The human auditory system can be divided into the physical perception apparatus, namely the ear, and the auditory cortex in the brain. As explained above, humans usually have two functioning ears, which provide us with the ability to use binaural hearing.

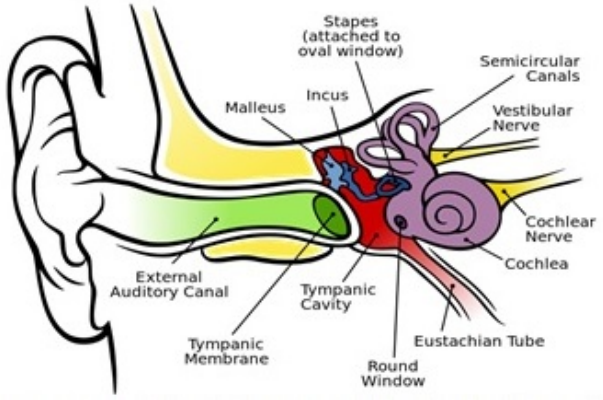

Figure 2: 'Anatomy of the Human Ear'; A. Brockmann

The sound waves reach our ear influenced by the shape of our head and auricles. Through the tympanic membrane and the auditory ossicles, the sound gets normalized and reaches the cochlea (see figure 2). Within the cochlea the sound is perceived in different areas by inner hair cells, which only react to pressure changes while constant noises are suppressed. Each area is built to perceive a unique frequency band. Not every frequency is perceived at the same amplitude, as the human auditory system is focused on speech perception. These functions are implemented using an early auditory system and 128 overlapping constant-Q asymmetric bandpass filters. Over the Cochlear and Vestibular nerves, the stimulation is transported to the auditory cortex, where the actual brain work starts. Since the human receptivity, as well as the computational bandwidth, is limited, the pre-filtered sound is selectively processed depending on the importance of the event. Humans use attention to focus on salient events, such as a siren or an announcement. This is comparable to a spotlight. The focused sound source is highlighted, while others are suppressed. The direction of the focus does not matter, which provides the possibility to follow a source through movement. Attention unfortunately only works for binaural hearing. The approach presented here implements a special salient event detector. It uses

salient changes in different features to create a saliency mapping of the sound clip over time. The features used are intensity, frequency contrast, temporal contrast and orientation, which are extracted using 2D spectrotemporal receptive filters, mimicking the analytical stages (see fig. 3) in the primary auditory cortex.

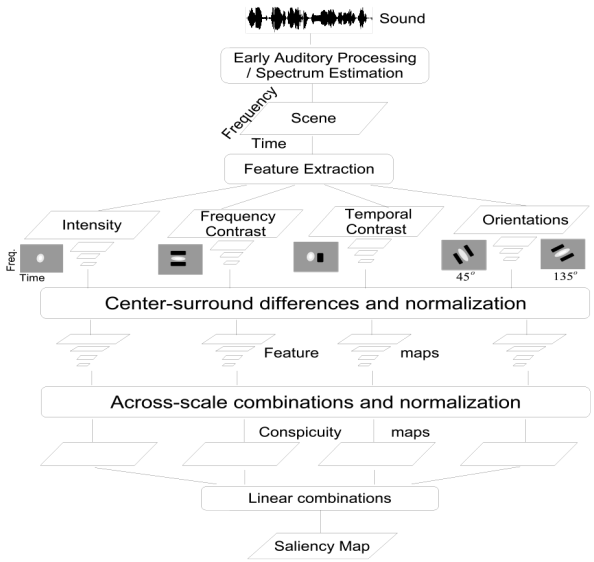

Figure 3: 'Schematic procedure for LISA'; [2]

In the next step, the freatue maps are combined and normalized into one saliency map. This map contains positive values in a range of [0-1], stating the saliency of events in the sound clip over time. Using these information, the most salient events are analyzed in detail, extracting MFCC 0-12 and F_0 features to characterize the events using vectors. The feature extraction is performed on overlapping windows of 20 ms. During the training phase, the classes of the events are learned, and high-level groupings of events for scenes are made. This creates a certain level of experience. All data is stored into a neural network, mimicking the neural storage of the brain. The neural network consists of 3-layers where the input neurons carry the features and the output neurons carry the scene classes. For the testing phase the whole process is repeated but in the end, the event vectors and combinations are run trough the stored data in the neural network to classify the new audio clip (see fig. 4).

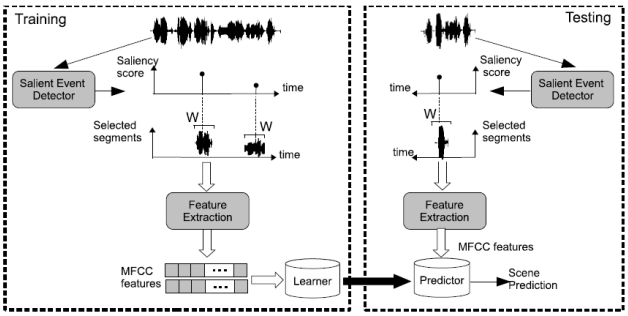

Figure 4: 'Training and Testing'; [2]

The Database used for this approach is the BBC sound effects library, which contains audio clips of varying length from 1 second to 9.5 minutes. Overall there were 2.491 audio clips unequally spread over 21 high-level semantic categories. 10% of the database was used for the evaluation set, while the remaining 90% were used for the training set. The amount of clips per category are presented in the following table (see table 1).

Category	No. of files	Category	No. of files
IMPACT	16	NATURE	85
OPEN	8	SPORTS	151
TRANSPORTATION	295	HUMAN	357
AMBIENCES	311	EXPLOSIONS	18
MILITARY	102	MACHINERY	117
ANIMALS	359	SCI-FI	121
OFFICE	144	POLICE	96
HORROR	98	PUBLIC	44
AUTOMOBILES	53	DOORS	4
MUSIC	25	HOUSEHOLD	38
ELECTRONICS	49		

Table 1: 'Distribution of the clips under each category'; [2]

The results revealed some interesting findings. LISA was capable to reach a 50% accuracy using the top rated 35 salient events, which implies a data reduction of 74%. Saving up to 98% amount of data, LISA achieved a 40% accuracy, using less than the top 5 salient events. The following figures show the amount of data reduced over the number of salient events (see figure 5) and the results for LISA dependent on the number of salient events (see figure 6). The number of clusters, which represents the number of features extracted, varies from 200-2000.

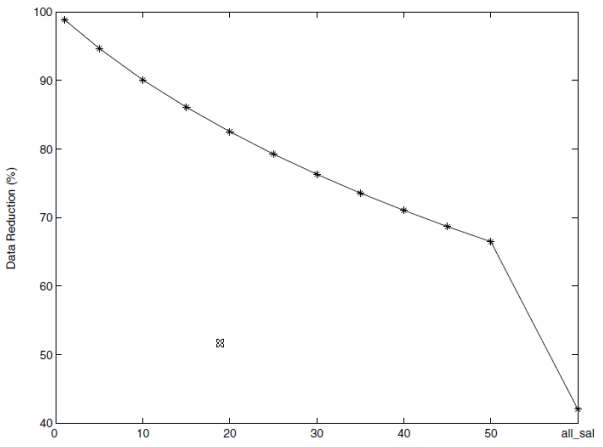

Figure 5: 'Data reduction over the number of salient events'; [2]

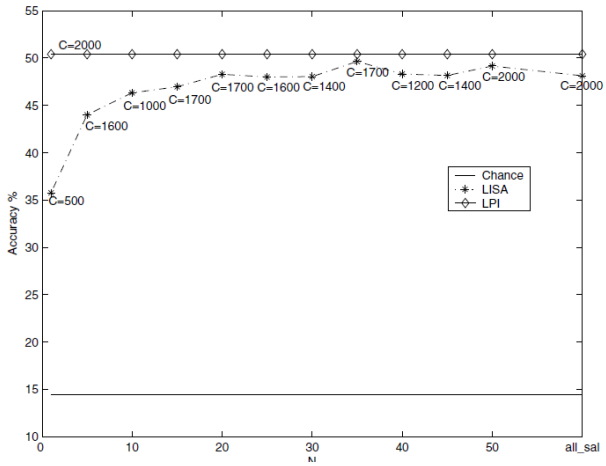

Figure 6: 'Clip accuracy results obtained with LISA and LPI'; [2]

4.3 Comparison

A closer look at the databases shows that the preconditions for the evaluation of each approach are different. The DCASE database contained only 10 categories while the BBC library contained 21 categories. In return, the DCASE database provided less training material than the BBC library. Overall, the distinction of 21 categories has more impact on the classification accuracy then the amount of used training data. Keeping this fact in mind, the juxtaposition of 69% for Geiger et. al. versus 50% for Kalinli et. al. gets less significant. This is supported by a comparison of the results for a baseline algorithm. Both databases were also tested using a baseline algorithm, which is a simple statistic approach using MFCCs. The results for the baseline algorithm should be independently comparable, therefore diverging results for the baseline are a hint for different complexities in scene classification for the databases. For the DCASE database the baseline algorithm showed a 55% accuracy, while for the BBC library it achieved only 40%. Although, to the fact that it can not for sure be said that the baseline algorithms perform equal, the results lead to the assumption that the BBC library scenes were harder to classify, giving an offset of 15%. Taking this fact into account, the difference between the statistical approach and LISA only accumulates to a 4% advantage. Depending on the application, a reduction of 74% computational effort in exchange for a relative improved accuracy of 4% sounds reasonable. This is though highly dependent on the intended goal.

5 Conclusion

In this paper an overview on the topic Acoustic Scene Classification is given stating two state of the art approaches. Although, the approaches follow completely different ideas, it figured, that they use similar methods for the feature extraction and classification. The main difference identified, eventually is focused on the way the audio signal is processed before the feature extraction and classification is performed. Here, the authors of the human based approach put high emphasis on mimicking the human auditory system. They apply several filters and a special salient event detector to reduce the amount of data processed. In contrast, the statistical approach processes all data available without applying intervening variables.

The results of both approaches could not be compared directly due to the different underlying databases for evaluation. A relative comparison taking into account an offset of 15% though suggests, that the performance of the statistic

based approach is slightly more accurate. The goal of the human based approach was not to achieve the best results but, reduce the amount of data. This task was fulfilled with huge success reaching a data reduction of 74% at a loss of relative 4% compared to the statistic approach. Lowering the expectations in accuracy, a data reduction of 98% would be possible while at the same time reaching a relative 55% accuracy. The reduced amount of data necessary for classification, makes this approach applicable for low-power devices such as smart-phones or hearing aids.

By varying the feature extraction and classification system, trimming the human based approach to achieve best results on a reduced amount of data, there probably is potential to go beyond the possibilities of the statistic approaches as stated in [5]. Further research could focus on the inclusion of external information such as geo-locations to further increase the accuracy and enable the classification of unknown scenes in an unsupervised manner.

References

[1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell and Mark D. Plumbley, Senior Member, IEEE. *Acoustic Scene Classification*, School of Electronic Engineering and Computer Science, (2014).

[2] Ozlem Kalinli, Shiva Sundaram, Shrikanth Narayanan. *Saliency-Driven Unstructured Acoustic Scene Classification Using Latent Perceptual Indexing*. MMSP , October 5-7, (2009).

[3] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange and Mark D. Plumbley. *Detection And Classification Of Acoustic Scenes And Events: An IEEE AASP Challenge*. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 20-23, (2013).

[4] EURASIP Journal on Applied Signal Processing, 18 (2005), pp. 2911–2914.

[5] Ben Milner and Dan Smith. *Acoustic Environment Classification*. ACM Transactions on Speech and Language Processing, Vol. 3, No. 2 (2006), pp. 1–22.

[6] Juergen T. Geiger, Bjoern Schuller, Gerhard Rigoll. *Large-Scale Audio Feature Extraction And Svm For Acoustic Scene Classification*. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 20-23, (2013).

[7] Malcolm Slaney. *The History and Future of CASA*. In Perspectives on Speech Separation, Editor: P. Divenyi, Kluwer, (2006).

[8] Deliang Wang and Guy J. Brown. *Fundamentals of Computational Auditory Scene Analysis*. (2006).

[9] http://de.wikipedia.org/wiki/Folgetonhorn, viewed on the 10th of July.