Acoustic Scene Classification with Hilbert-Huang Transform Features

Jürgen Tchorz

Technische Hochschule Lübeck, Institut für Akustik, D-23562 Lübeck, Email: juergen.tchorz@th-luebeck.de

Introduction

Acoustic Scene Classification (ASC) is the task of classifying environments from the sound they produce into one of the provided predefined classes that characterizes the environment in which it was recorded. A typical application of computational ASC is a hearing aid which automatically selects a set of appropriate amplification and feature parameters (such as microphone directivity) according to the acoustic environment (e.g., speech in noise, car or music) [1].

The common approach for ASC is to divide a training set of recordings into frames of fixed duration, to compute a sequence of features from these frames, and to train a model which summarizes the properties of predefined acoustic scenes, which is then used to classify features from recordings not included in the training set.

A range of different audio features have been employed in ASC systems. Simple low-level features such as spectral centroid, spectral slope or statistical distribution of signal amplitudes are used in hearing aids [2]. Frequency-band energy features which sometimes also mimic properties of human auditory bands such as Mel-scale spectra (which are most widely used in ASC), or Gammatone filters [3] represent the spectral shape of an audio frame. Cepstral features, especially MFCCs which summarize the spectral envelope are also used in ASC and other fields of sound processing such as automatic speech recognition. Spatial features such as interaural time- or level differences can be exploited if multi-microphone recordings are available. Estimates of the instantaneous frequency based on the Hilbert-Huang Transform were used as features for the classification of underwater sound [3] and biosignals such as lung sounds [4, 5]. A detailed review on ASC features and statistical models such as Gaussian mixture models or support vector machines to model the properties of sound classes can be found in [6]. In recent years, deep neural networks have increasingly been applied to ASC.

Most audio features used for ASC represent current properties of the input signal in the respective analysis frame (e.g., the frequency spectrum), but not *changes* of these properties over time. So-called delta-features which are sometimes also used in ASC represent differences between subsequent frames, but do not involve further analysis of amplitude and frequency fluctuations of the input signal. However, psychoacoustic studies on human ASC suggest that temporal properties of the signals such as envelope modulations contribute to predict the identification of different brief everyday sounds [7]. Frequency and temporal fluctuations are fundamental attributes of sound. The tonotopical representation of frequency has been found in the cochlea and in different areas in the ascending auditory pathway including the auditory cortex. In neurophysiological experiments, several researchers found neurons in the inferior colliculus and auditory cortex of mammals which were tuned to certain modulation frequencies, i.e., temporal fluctuations. The "periodotopical" organization of these neurons with respect to different best modulation frequencies was found to be almost orthogonal to the tonotopical organization of neurons with respect to center frequencies. Thus, a two-dimensional map represents both spectral and temporal properties of the acoustical signal (see [8] for a review).

Frequency modulations play an important role in e.g. speech perception. Manipulating the FM characteristics of formants in speech changes its perceived phonemic characteristics [9]. Electrode recordings identified neurons in the primary auditory cortex of cats that are systematically distributed according to either the rate and direction of frequency modulation sweeps [10].

Conventional spectral or cepstral features are not well suited to capture these fast modulations. An example is illustrated in Fig.1, which shows spectrograms of a starting light aircraft and of noise which was generated by overlapping random segments of that aircraft sound. The spectrograms look essentially the same when using typical analysis parameters (20 ms window length, 10 ms shift). However, the signals sound clearly different: while it is easy to identify an aircraft by listening to the first signal, the second signal just sounds like spectrally shaped unmodulated noise.

This study focuses on audio features which emphasize on representing spectro-temporal fluctuations of the input signal, namely amplitude modulation spectrograms (AMS), and statistical distributions of instantaneous frequencies based on the Hilbert-Huang Transform which are novel to ASC. For comparison, Mel spectra are computed, as they are used in most current ASC approaches. The classification system comprises a deep recurrent neural network and is evaluated using the DCASE 2018 challenge ASC dataset [11].

Features

(a) Instantaneous frequency distributions (IFD) - The Hilbert-Huang transform [12] breaks down nonstationary and nonlinear signals into a finite number of components to which the Hilbert spectral analysis can be applied. These components are called intrinsic mode functions (IMF) which form a complete and nearly orthogonal basis for the original signal. The signal is decomposed

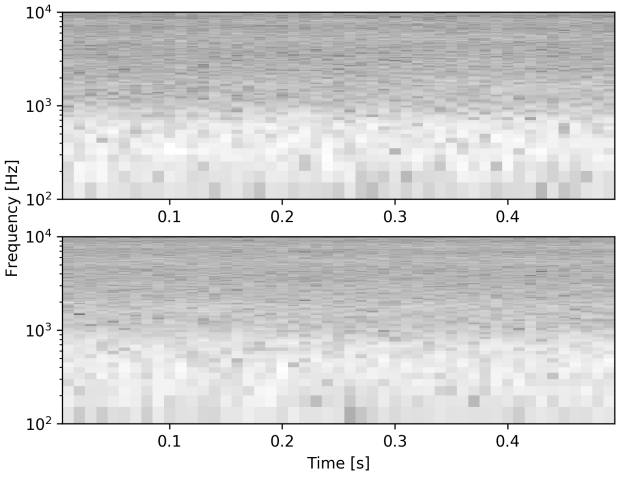

Figure 1: Spectrograms of a starting light aircraft (top) and of noise which was generated by overlapping random segments of that aircraft recording (bottom). Both signals sound clearly different.

in the time domain, and the length of the IMFs is the same as the original signal. Subsequent Hilbert transforms yield the instantaneous frequencies of the IMFs. Fig. 2 shows the instantaneous frequencies of the 2nd

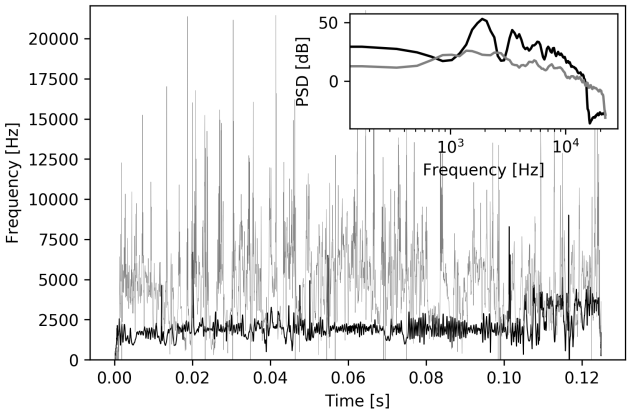

Figure 2: Instantaneous frequency of the 2nd IMF for a short segment of a Kiwi bird call (black line), and for the sound of a small river (thinner gray line). The embedded plot shows the overall spectra of both segments (Kiwi: black; river: gray).

IMFs of 128 ms recordings of a Kiwi bird call and a small river. The overall spectra of the two signals are similar, but the distribution of the instantaneous frequencies differs a lot: while it is relatively constant over time for the bird call, the instantaneous frequency of the river IMF fluctuates much more. Thus, the fluctuations of instantaneous frequencies can be used to distinguish between signals. To further examine the discriminative power of instantaneous frequency distributions (IFD), they have been computed from recordings provided for the DCASE 2018 ASC challenge [11]. For all 10 given acoustic scenes, 6 recordings taken in different European cities were processed. In addition, 6 recordings from different speakers (m/f) taken from the Buckeye Corpus [13] were processed in the same way. "Speech" is not an acoustic scene in the DCASE ASC challenge, but it is an important sound class for e.g. hearing aids, why this class has been included in this example. Each of the 66 recordings had a duration of 10 s. Fig. 3 shows the IFD for the 3rd IMF. The patterns of the IFD show similarities across acoustic

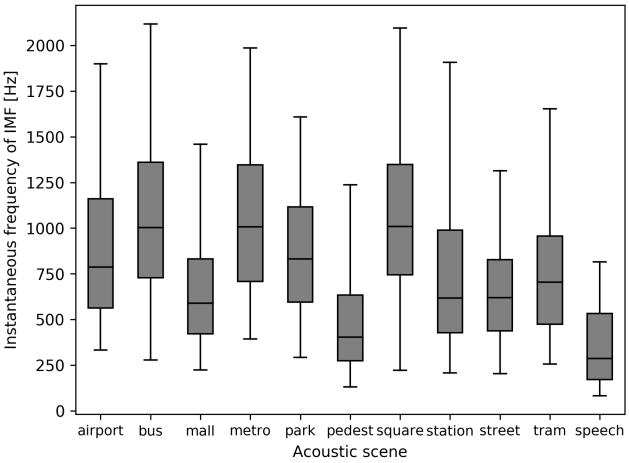

Figure 3: Instantaneous frequency distributions calculated from the 3rd IMF in 11 different acoustic scenes (0.05th, 0.25th, 0.5th, 0.75th and 0.95th percentiles).

scenes, but also disparities which can potentially be used for scene classification.

For feature extraction, the instantaneous frequencies of the first 10 IMFs are calculated from 128 ms segments of the input signal, with an overlap of 64 ms. The instantaneous frequencies are functions of time with the same sample rate as the input signal. For each 128 ms segment, 5 percentiles (the 0.05th, 0.25th, 0.5th, 0.75th and 0.95th) of the instantaneous frequency distribution are calculated for each of the first 10 IMFs. Thus, each 128 ms segment is represented by a feature vector with 50 numbers in total.

(b) Amplitude Modulation Spectrograms - AMS reflect both spectral and temporal aspects of the input signal. In the field of speech processing, AMS features have originally been used in a binaural speech enhancement approach utilizing spatial separation of the speech and noise [14]. For single-channel SNR estimation and speech enhancement, AMS features have been used in combination with simple neural networks with one hidden layer [15]. More recently, AMS features were utilized for a noise suppression with a Bayesian classifier and ideal binary masks. In normal hearing subjects and noise types which were also used for training, substantial improvements in speech intelligibility could be shown [16]. A complementary feature set consisting of AMS features, relative spectral transform and perceptual linear prediction (RASTA-PLP), and mel-frequency cepstral coefficients (MFCCs) was combined with a deep neural network to train binary masks for noisy speech [17]. The authors report substantially increased speech intelligibility in hearing-impaired listeners. While the speech segments for testing intelligibility were not included in the training data, the background noise was also used for training. Thus,

generalization to unknown noise was not investigated in this study.

In the field of acoustic scene classification, AMS features have been used with a neural network classifier in an early approach which just distinguished between speech and noise [18] and in combination with MFCCs [19]. AMS features which have been further reduced to just 9 features using the Covariance Matrix Adaptation Evolutionary Strategy [20]. Using a Linear Discriminant Analysis (LDA) classifier, the authors report an improvement of 10 percentage points for the IEEE AASP Challenge 2013 public dataset, compared to the best previously available approaches.

For AMS generation, fast Fourier transforms (FFT) are computed for overlapping 4 ms segments of the signal with 0.25 ms hop size. Appropriate summation of neighboring FFT bins yield 40 frequency channels with a mel-frequency mapping and spanning from 0 to 22 kHz. The resulting amplitudes in each frequency channel are regarded as envelope signal. The modulation spectra are obtained by computing FFTs in each frequency channel across a Hanning-windowed time segment of 128 ms with an overlap of 64 ms. The modulation frequency resolution is 15.6 Hz. The FFT magnitudes are multiplied by 15 triangular-shaped windows spaced uniformly across the 15.6 - 400 Hz range in each mel frequency channel and summed up to produce 15 modulation spectrum amplitudes. Thus, each AMS pattern representing 128 ms of the input signal consists of 40 x 15 = 600 numbers.

(c) Mel spectra - For frequency analysis, short-term Fourier transformations are calculated with a window size of 2048 samples (46 ms) and a hop size of 1024 samples. Next, the calculated power spectrum is converted to 128 bands Mel scaled features. Finally, the spectra are transformed to the logarithmic scale, normalized by dividing by the standard deviation and subtracting the mean value. These parameter settings have been chosen by the winner team of the DCASE 2018 challenge (Task 1A).

Classifier

For classifying the acoustic scenes, a recurrent neural network (long short-term memory network, LSTM) with three hidden recurrent layers (1000, 1000 and 500 neurons) was implemented. A softmax function and cross entropy loss were used. The minibatch size was 4096, and the learning rate was set to 0.001 per sample with 800 epochs. The network was implemented with the CNTK toolkit running on a GeForce GTX 970 GPU.

Training and testing

The development data set of the DCASE 2018 challenge (Task 1B) consists of 10,080 mono sound files recorded with three different devices using a sample rate of 44.1 kHz in 10 different acoustic scenes. Each sound file has a length of 10 s. After feature extraction, each sound file is represented by 156 feature sets (instantaneous frequency distributions, AMS patterns or Mel spectra) which each represent 128 ms of the input signal (64 ms hop size). The development data set is split into a subset with 7202 sound files for supervised training and 2878 sound files for testing. For testing, each feature set generated from 128 ms segments of the test subset was classified independently. The acoustic scene which has been detected most often within a given sound file was the overall classification result for this sound file.

Table 1: Acoustic scene classification rates (%) for the test subset of the DCASE 2018 development data using different feature sets: instantaneous frequency distributions (IFD), amplitude modulation spectrograms (AMS), Mel spectra (Mel), and concatenation of these three feature sets (All). DCASE 2018 baseline system for comparison.

Scene label	IFD	AMS	Mel	All	baseline
1 Airport	71.1	68.1	61.5	66.1	73.3
2 Mall	51.8	68.9	45.1	51.1	48.2
3 Station	27.5	45.4	55.6	56.3	50.4
4 Pedestrian	45.6	48.4	47.4	47.4	51.2
5 Public squ	32.1	47.2	47.6	51.2	36.2
6 Traffic	81.9	79.8	84.8	85.8	80.5
7 Tram	66.0	56.6	63.6	64.7	51.9
8 Bus	60.8	42.8	69.1	64.4	59.4
9 Metro	38.4	63.0	52.2	52.2	43.3
10 Park	74.1	75.5	81.7	78.4	78.1
Average	**55.0**	**59.8**	**60.7**	**61.7**	**57.2**

Results

The results for the proposed partitioning of the development data set into train and test data are given in Table 1. For comparison, the results of the DCASE 2018 baseline system [11] are also shown. In the baseline system, log mel-band energies were first extracted in 40 bands using an analysis frame of 40 ms with a 50% hop size. The neural network consists of two convolutional neural network layers and one fully connected layer, and uses an input of size 40x500, equivalent to the full length of the segment to be classified.

The classification accuracy varies across acoustic scenes and feature sets. On average, AMS and Mel spectra yield similar results and outperform the baseline system, while classification accuracy with instantaneous frequency distribution features is about 5 percentage points lower. The combination of all three feature sets yields the best results.

Discussion

The proposed IFD feature set which is novel to ASC is discriminative enough to allow for classification rates which are reasonable, but clearly lower than with standard Mel spectra features. However, calculating percentiles from instantaneous frequencies is just a rough descriptive statistics, and other analysis methods might be more discriminative. In addition, the size of the IFD feature set is quite small (781 numbers per second, com-

pared to 9360 numbers for AMS features, and 5565 numbers for Mel-spectral features), which might be important for applications with tight memory resources. The combination of different feature sets appears to be beneficial, which has also been shown in other areas of sound processing such as speech enhancement using deep neural networks[17].

References

[1] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 2991–3002, 2005.

[2] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 2915 – 2929, 2005.

[3] X. Zeng and S. Wang, "Underwater sound classification based on gammatone filter bank and hilbert-huang transform," in *2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Aug 2014, pp. 707–710.

[4] Z. Li and M. Du, "HHT based lung sound crackle detection and classification," in *2005 International Symposium on Intelligent Signal Processing and Communication Systems*, Dec 2005, pp. 385–388.

[5] X. Chen, J. Shao, Y. Long, C. Que, J. Zhang, and J. Fang, "Identification of velcro rales based on hilbert huang transform," *Physica A: Statistical Mechanics and its Applications*, vol. 401, pp. 34–44, 2014.

[6] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[7] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *J Exp Psychol Hum Percept Perform*, vol. 19(2), pp. 250–267, 1993.

[8] S. Baumann, O. Joly, A. Rees, C. I. Petkov, L. Sun, A. Thiele, and T. D. Griffiths, "The topography of frequency and time representation in primate auditory cortices," *eLife*, vol. 4, 2015.

[9] A. M. Liberman, D. P. C., L. J. Gerstman, and F. S. Cooper, "Tempo of frequency change as a cue for distinguishing classes of speech sounds," *Journal of Experimental Psychology*, vol. 52(2), pp. 127–137, 1956.

[10] J. R. Mendelson, C. Schreiner, M. L. Sutter, and K. L. Grasse, "Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps," *Experimental brain research.*, vol. 94, pp. 65–87, 02 1993.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09840

[12] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998. [Online]. Available: http://rspa.royalsocietypublishing.org/content/454/1971/903

[13] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu]*, Department of Psychology, Ohio State University, Columbus, OH, 2007.

[14] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of the Acoustical Society of America*, vol. 95(3), pp. 1593 – 1602, 1994.

[15] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 11(3), pp. 184 – 192, 2003.

[16] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126(3), pp. 1486 – 1494, 2009.

[17] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am*, vol. 138(3), pp. 1660 – 1669, 2003.

[18] J. Tchorz and B. Kollmeier, "Using amplitude modulation information for sound classification," in *Psychophysics, Physiology and Models of Hearing*. World Scientific, Singapore, 1998, pp. 275 – 278.

[19] J. Tchorz, "Combination of amplitude modulation spectrogram features and MFCCs for acoustic scene classification," DCASE2018 Challenge, Tech. Rep., September 2018.

[20] S. Agcaer, A. Schlesinger, F.-M. Hoffmann, and R. Martin, "Optimization of amplitude modulation features for low-resource acoustic scene classification," in *23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2601 – 2605.