COMBINATION OF AMPLITUDE MODULATION SPECTROGRAM FEATURES AND MFCCS FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Jürgen Tchorz

University of Applied Sciences Lübeck
Mönkhofer Weg 239
23562 Lübeck, Germany
tchorz@th-luebeck.de

ABSTRACT

This report describes an approach for acoustic scene classification and its results for the development data set of the DCASE 2018 challenge. Amplitude modulation spectrograms (AMS), which mimic important aspects of the auditory system are used as features, in combination with mel-scale cepstral coefficients which have shown to be complementary to AMS features. For classification, a long short-term memory deep neural network is used. The proposed system outperforms the baseline system by 6.3-9.3 % for the development data test subset, depending on the recording device.

Index Terms— Amplitude modulation spectrograms, MFCCs, acoustic scene classification, deep neural networks, LSTM

1. INTRODUCTION

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 [1] for a review).

The features used in this study (Amplitude Modulation Spectrograms, AMS) mimic these maps and reflect both spectral and temporal aspects of the input signal. AMS features have been used in several areas of psychoacoustics and audio processing. Dau et al. [2, 3] proposed a psychoacoustical model in which temporal information is extracted by amplitude modulation filter banks that are physiologically located in the midbrain. 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 [4]. For single-channel SNR estimation and speech enhancement, AMS features have been used in combination with simple neural networks with one hidden layer [5]. 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 [6]. 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 [7]. 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.

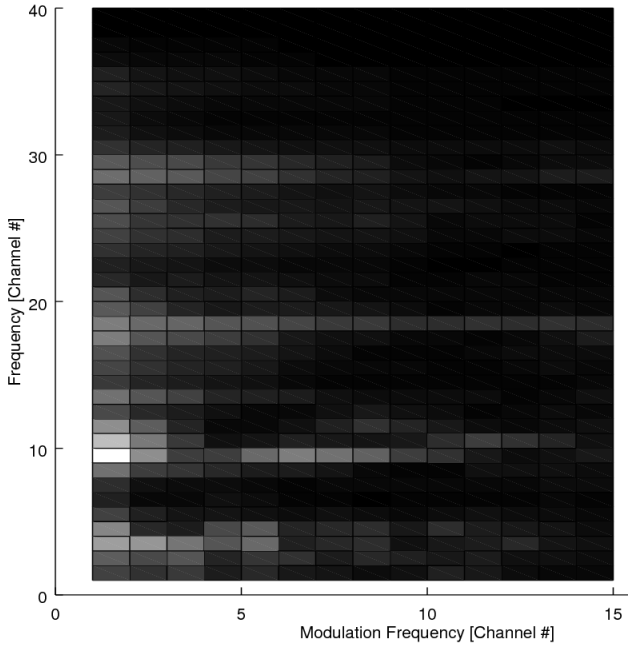

Figure 1: Example of an AMS pattern generated from a 128 ms segment of speech. Bright shading indicates high energy.

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 [8]. More recently, AMS features which have been further reduced to just 9 features using the Covariance Matrix Adaptation Evolutionary Strategy

Table 1: Acoustic scene classification rates for the test subset of the development data, for recording devices A, B and C (percentage).

Scene label	B	C	$\overline{\text{B, C}}$	A	$\overline{\text{A, B, C}}$
1 Airport	44.4	50.0	47.2	68.7	66.1
2 Shopping mall	72.2	66.7	69.5	52.7	54.6
3 Metro station	50.0	27.8	38.9	56.8	54.6
4 Street, pedestrian	33.3	50.0	41.7	56.7	54.8
5 Public square	33.3	44.4	38.9	52.3	50.4
6 Street, traffic	66.7	72.2	69.5	84.6	82.6
7 Tram	33.3	44.4	38.9	74.0	69.7
8 Bus	77.8	72.2	75.0	59.1	61.2
9 Metro	55.6	33.3	44.5	63.6	61.3
10 Park	77.8	72.2	75.5	83.4	82.4
Average	54.4	53.3	53.9	65.2	63.8
Baseline	45.1	46.2	45.6	58.9	

were applied for acoustic scene classification [9]. 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.

In this work, a combination of AMS and MFCC features which have shown to be complementary [10] and a long short-term memory deep neural network are used for acoustic scene classification.

2. FEATURES

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. An example of an AMS pattern generated from a speech portion is shown in Fig. 1.

In addition to AMS patterns, mel-scale cepstral coefficients (MFCCs) are used to represent the input signal in a complementary way [10]. Each AMS pattern representing 128 ms is augmented with 4 MFCC vectors calculated from 32 ms of the input signal and containing 40 mel-frequency cepstral coefficients. Thus, each 128 ms segment of the signal is represented by 600 + (4 x 40) = 760 numbers.

3. CLASSIFIER

For classifying the AMS/MFCC patterns, 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 network was implemented with the CNTK toolkit running on a GeForce GTX 970 GPU.

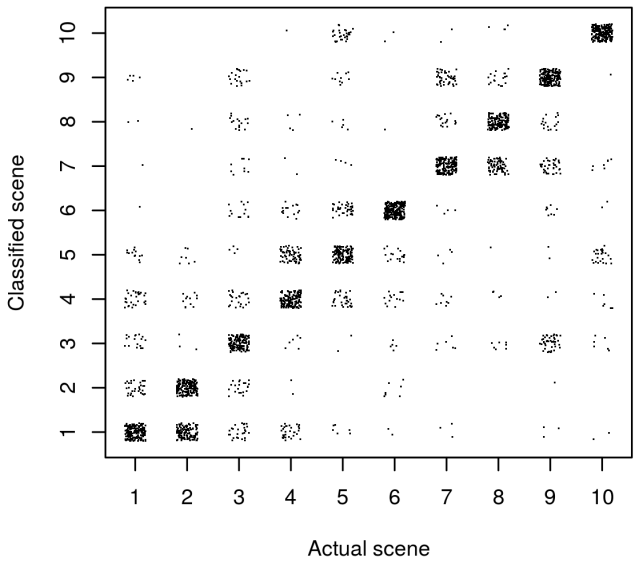

Figure 2: Confusion scatter plot for the development test subset. Each dot represents one soundfile. Scene labels as given in Table 1.

4. TRAINING AND TESTING

The development dataset consists of 10,080 soundfiles recorded with three different devices in ten different acoustic scenes. Each soundfile has a length of 10 s and is represented by 156 AMS/MFCC patterns (64 ms hop size). For training, each AMS/MFCC pattern was fed into the network with its corresponding label. 150 iterations were computed for training. For testing, each pattern generated from the test subset was classified independently. The acoustic scene which has been detected most often within a given soundfile was the overall classification result for this soundfile.

5. RESULTS

The results for the proposed partitioning of the development data set into train and test data are given in Table 1.

The presented classification system allows for increased scores, compared to the baseline system. Classification results for recording devices B and C are worse than for device A, which was the main source for the training data. Thus, the proposed system does not seem to generalize to mismatched recording devices, although the improvement compared to the baseline system is higher for devices B and C than for device A. However, the amount of soundfiles from devices B and C in the test set of the development data is too limited (18 per acoustic scene and device) to draw statistically firm conclusions on this. In addition, overfitting might be an issue. After training the neural network with the full development data set (including the development test data subset), there was 100 % classification accuracy for the test data subset.

The acoustic scenes "bus" and "shopping mall" have better classification rates with devices B and C, compared to device A. This is

also true for the baseline system.

Figure 2 shows the confusions between acoustic scenes. It can be seen that "airport" and "shopping mall" are often confused with each other, as well as "metro", "metro station" and "tram". These confusions are not too surprising, as there are also perceptual similarities between these scenes. On the other hand, for example, "shopping mall" has never been mistakenly classified as "park", and vice versa.

The classification rates based on isolated 128 ms AMS/MFCC patterns (i.e., without considering which scene has been classified most often in a 10 s soundfile) were 39.6 %, 36.2 %, and 47,4 % for devices B, C and A, respectively.

6. DISCUSSION

For the development data, the proposed classification approach with AMS/MFCC features outperforms the baseline system which is based on MFCC features alone. Thus, the parameter size is much larger for the proposed system. However, a reduction of the AMS patterns could be possible without compromising the results [9]. The LSTM neural network has more or less been used "out of the box", but more carefully chosen parameters might have helped to improve performance, for example with respect to avoiding overfitting and enhancing generalization.

7. REFERENCES

[1] 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.

[2] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers." *The Journal of the Acoustical Society of America*, vol. 102, pp. 2892 – 2905, 1997.

[3] ——, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration." *The Journal of the Acoustical Society of America*, vol. 102, pp. 2906 – 2919, 1997.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Chen, Y. Wang, and D. Wang, "Evaluation of sound classification algorithms for hearing aid applications," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, vol. 22(12), 2014, pp. 1993 – 2002.