AUTOMATIC RECOGNITION OF ENVIRONMENTAL SOUND EVENTS USING ALL-POLE GROUP DELAY FEATURES

Aleksandr Diment, Emre Cakir, Toni Heittola, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

ABSTRACT

A feature based on the group delay function from all-pole models (APGD) is proposed for environmental sound event recognition. The commonly used spectral features take into account merely the magnitude information, whereas the phase is overlooked due to the complications related to its interpretation. Additional information concealed in the phase is hypothesised to be beneficial for sound event recognition. The APGD is an approach to inferring phase information, which has shown applicability for speech and music analysis and is now studied in environmental audio. The evaluation is performed within a multi-label deep neural network (DNN) framework on a diverse real-life dataset of environmental sounds. It shows performance improvement compared to the baseline log mel-band energy case. Combined with the magnitude-based features, APGD demonstrates further improvement.

Index Terms— Phase spectrum, sound event recognition, audio classification, neural networks

1. INTRODUCTION

The goal of sound event recognition is to represent an audio signal as a stream of sound events present in the auditory scene. It has applications in indexing and retrieval in multimedia databases, personal health and surveillance. Sound events e.g. within an office environment include keyboard typing, mouse clicking, and various human sounds (speaking, coughing).

Recently, a spectrogram image feature has been incorporated into a robust sound event classifier based on the generalized Gaussian distribution Kullback-Leibler kernel SVM [1]. Detection of overlapping sound events has been done with unsupervised non-negative matrix factorization -based sound source separation [2], as well as by incorporating non-negative dictionaries data based on the spectrogram of the mixture signal and its annotation without source separation [3]. Multi-label deep neural networks (DNN) have been used to detect simultaneous sound events in real-life recordings [4].

A number of standard features are used in the state-of-the-art. Their choice is dictated by the nature of the sound events, characterised both spectrally and temporally. Spectrally, the events are discriminated by being harmonic (whistles, bird songs) or non-harmonic (ventilation noise). From the temporal perspective, the events exhibit transient (knocking, ball hitting floor), stationary (wind and traffic noises) or combined (rain drop within a rain drop texture) properties. This division dictates the motivation for employing corresponding features. One of the standard spectral approaches is MFCCs, and the temporal viewpoint is incorporated e.g. by detecting transient points in spectral energy [5]. Combined spectro-temporal approaches include convolutive non-negative matrix factorisation of mel-frequency spectrograms [6], as well as amplitude modulation spectrogram and Gabor filterbank features [7].

The common spectral features focus on the magnitude part, while spectral information is complete only if phase spectrum is specified as well. Such difficulties, as wrapping of the phase and its dependency on the window position, make direct processing of the phase spectra challenging. One way of overcoming this is the *group delay function from all-pole models* (referred to as APGD, all-pole group delay). Its main aspect is to calculate the group delay function from all-pole models of a signal, formed by linear prediction. The method has been used in formant extraction [8], speaker recognition [9] and musical instrument recognition [10]. Prior to APGD, other ways of incorporating phase had been studied for speech analysis [11] and onset detection of musical instruments [12].

This work proposes using phase information in a form of APGD feature in a multi-label DNN-based environmental sound event recognition system. The calculation of the feature is proposed either primarily or as a complement to the log mel-band energy, motivated by the fact that additional relevant information is concealed in phase. The performance is evaluated on a large diverse real-life dataset. Previously, decorrelation by means of DCT and dimensionality reduction has been performed on the feature. However, within the DNN framework and given sufficient amount of training data, we propose to exclude this step. The novelties of our method include incorporation of the phase into sound event recognition, as well as applying and adjusting the APGD feature without the DCT step for a DNN classifier.

Next, we present the motivation for APGD and its calculation procedure. Section 3 describes the sound event recognition system, which incorporates the feature. Its performance is evaluated in Section 4, and conclusions are drawn in Section 5.

This research has been supported by the Academy of Finland, project number 258708.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. ALL-POLE GROUP DELAY FEATURE

In this section, firstly, the use of phase information for sound event recognition is introduced with the motivation for computing APGD in particular. Secondly, the details of calculation of the group delay function and APGD are presented.

2.1. Motivation for sound event recognition

Phase is overlooked in many audio processing solutions due to the complications related to the unwrapping of the phase spectrum. Still, it can be informative due to its high resolution and ability of indicating peaks in the magnitude spectrum envelope. In speech signals, these correspond to formants.

In the case of environmental audio analysis, the diversity of sound event classes and their natural polyphony renders such argumentation not necessarily universally applicable. Still, it is worthwhile to investigate the contribution of phase information to sound event recognition due to the importance of phase in human perception [13], as well as because of the harmonic structure of the sounds of some acoustic events.

2.2. Group Delay Function

The *group delay function* of a signal $x[n]$ can be obtained as [14]

$$\tau_g(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \quad (1)$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x[n]$ and $y[n]$, $y[n] = nx[n]$, and R and I denote respectively the real and imaginary parts of the Fourier transform. The group delay function is well-behaved only if the zeros of the transfer function of the modelled filter are not close to the unit circle. When zeros of the transfer function are close to the unit circle, the magnitude spectrum exhibits dips at the corresponding frequency bins. Thus, the denominator term in (1) tends to a small value, resulting in a large value of $\tau_g(\omega)$. This leads to spurious high amplitude spikes at these frequencies, masking out the resonance structure in the group delay function.

One way of addressing this issue is by introducing a modification [11] of the group delay function (MODGDF), which suppresses the zeros of the transfer function by adding cepstral smoothing. However, it introduces three parameters to be adjusted to an application scenario, which is computationally expensive. Another approach, which lacks of this complication, is the group delay function of all-pole models.

2.3. Group Delay Function of All-Pole Models

By modelling analysed environmental event sound with a source-filter model and assuming the filter all-pole, the spectrum of such filter with the frequency response $H(\omega)$ may be approximated with aid of linear prediction. Linear prediction is formulated as [15]

$$H(\omega) = \frac{G}{1 - \sum_{m=1}^{p} a(m)e^{-j\omega m}}. \quad (2)$$

Here, G is the signal-dependent gain and p is the model order. The coefficients $a(m)$ are determined by the method of least squares in such a way that the power spectrum of $H(\omega)$ matches the power spectrum of the signal $|X(\omega)|^2$. The all-pole group delay function is computed from the phase response of this filter formed by $H(\omega)$.

An optional discrete cosine transform (DCT) can be applied for decorrelation, and a number of coefficients are excluded, similarly to MFCCs. The feature is calculated in short frames under the assumption of spectral stationariness, and the Fourier analysis is done with DFT. The overall calculation procedure of APGD, illustrated in Fig. 1, is the following.

1. Perform all-pole modelling on the frame. Obtain the filter coefficients $a(m)$.
2. From the $a(m)$, form the frequency response $H(\omega)$ using (2) with $G = 1$ (for normalisation purposes).
3. Compute the group delay function by taking the negative derivative of the phase response of $H(\omega)$. In practice, the derivative is computed using the sample-wise difference.
4. Optionally perform DCT and keep a certain number of coefficients, excluding the zeroth.

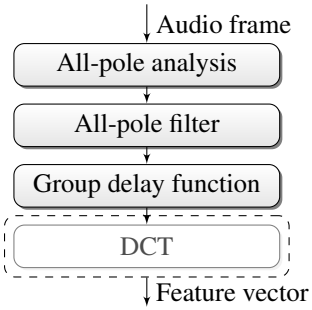

Fig. 1. A block diagram of the calculation of APGD.

3. SYSTEM DESCRIPTION

This section presents the sound event recognition system which incorporates APGD into a multi-label DNN classifier. Details on preprocessing and feature extraction are followed by the description of the classifier and the applied post-processing.

3.1. Preprocessing and Feature Extraction

The signals, from which sound events will be detected, are amplitude normalised, followed by frame-blocking with Hamming windowing. Frames of 50 ms and 50% overlap are used. Thereupon, the features are extracted.

Using both the proposed APGD features and the established log mel-band energies, as well as their combination is foreseen. In the case of APGD, the following parameters are set: LPC order 40, as used previously [10], feature vector lengths 512 (no DCT) and 40 (with DCT and coefficient removal). For the log mel-band energies, the feature vector length is set to 40, based on the past studies [4]. A combined feature case is considered by a frame-wise concatenation.

3.2. Classifier

Both the proposed and the baseline features are used as inputs to a multi-label deep neural network (DNN) to perform multi-label sound event classification. For each time frame t, a feature vector $\mathbf{x}_t$ is used as a learning instance for the DNN. The target output vector for each frame $\mathbf{y}_t$ is a binary vector with elements determined from the manual annotations as

$$y_t(l) = \begin{cases} 1, & l^{th} \text{ event is active in frame } t, \\ 0, & l^{th} \text{ event is not active in frame } t. \end{cases} \quad (3)$$

With the help of the hierarchical topology of the DNN, high-level features are implicitly learned in the higher-level hidden layers. This makes it possible to model the vastly non-linear relationships between the input and the output.

The multi-label DNN architecture is composed of an input layer with $||\mathbf{x}||$ units, two or more hidden layers and an output layer with N units, where N is the total number of sound events. For each layer k, the outputs $\mathbf{h}^k$ are calculated from the weighted sum of the outputs from the previous layer $k-1$, starting from $\mathbf{h}^0 = \mathbf{x}$ (time index t omitted for simplicity) and

$$\mathbf{g}^k = \mathbf{W}^k \mathbf{h}^{k-1} + \mathbf{b}^k, 1 \leq k < M, \quad (4)$$

$$\mathbf{h}^k = f(\mathbf{g}^k), \quad (5)$$

where $\mathbf{W} \in \mathbb{R}^{D \times S}$ is the weight matrix between layers $k-1$ and k, D and S are the number of units for layers $k-1$ and k respectively, $\mathbf{b} \in \mathbb{R}^S$ is the bias vector for layer k, $f(\cdot)$ is the nonlinear activation function for layer k and M is the total number of layers. For the hidden layer activation functions, *maxout* is used [16] and for the output layer activation function, logistic sigmoid is used to get a detection in the range $[0, 1]$. Cross-entropy cost function is chosen during the training of the DNN. Unlike the conventional quadratic cost function, cross-entropy does not suffer from slow learning due to small gradients of the sigmoid function when the sigmoid output is inaccurate by a large margin. Regularisation techniques, such as weight norm regularisation and dropout, were preliminarily tried out without noticeable improvement. In the testing stage, the source-presence prediction vector is binarised with a threshold 0.5 to obtain a binary detection estimation vector.

3.3. Post-processing

The intermittent nature of the real-life sound events and coarse time resolution of the annotations result in noise in the DNN outputs: short periods of the audio with low activity may be erroneously annotated with events. This causes abrupt changes in the DNN outputs between consecutive frames.

To filter out this noise, a median filtering based post-processing is applied. For each binary detection estimation $z_t(l)$, the post-processed estimation $\tilde{z}_t(l)$ is obtained by taking a median of the previous ten frames. Therefore, the previous ten frames should contain at least five consecutive detections to pass through.This effectively filters out the short bursts of detections. Ten frames are used, spanning a 250 ms window, which has been previously found reasonable [4]. A smaller frame size and/or larger overlap could be considered helpful, however, it was not the case in our experiments.

4. EVALUATION

We evaluated the performance of the proposed method in a polyphonic sound event recognition scenario. As a baseline, log mel-band energy was used, as well as MFCCs. Comparison with the previous phase-based method (i.e. MODGDF) was omitted due to its extensive prior parameter adjustment requirement.

4.1. Acoustic Material

An existing database [17] of real-life diverse environmental sound events, collected with binaural in-ear microphones, was used for the evaluation. A total of 103 recordings, each of 10 to 30 minutes long, were collected in ten contexts: basketball, beach, bus, car, hallway, office, restaurant, grocery shop, street and stadium. The recordings were averaged into mono signals in this work, with sampling rate 44.1 kHz and bit depth 24.

The annotations include the start and end times of all clearly audible sound events in the auditory scene. A total of 61 event classes is presented, with examples of speech, laughter, applause, car door, road, dishes, door, music, and footsteps. Events of 9–16 classes are present in each context.

4.2. Parameters

To benefit from the temporal characteristics of sound events, optional context windowing was foreseen by concatenating a feature vector with its neighbouring frames. Here, we studied the effect of context windowing with two frames in each direction.

The optimal hyper-parameters for DNN were estimated over a grid search of possible values and kept constant through different evaluation scenarios. Two hidden layers with 800 hidden units each, 0.02 learning rate and maxout pooling with size 2 were used. The dataset was divided into 70% training, 20% test and 10% validation sets, and five-fold stratified cross-validation was performed.Due to the previously observed [4] clear superiority of DNN over the other state-of-the-art, evaluation of the proposed feature was performed only with DNN.

Table 1. The results of the evaluation of the effect of LPC order on the performance of the APGD feature.

LPC order	10	15	20	30	40	50	60
F_1 score	56.4	57.9	58.6	59.0	59.7	59.1	59.0

Table 2. Feature-wise evaluation results.

Features	F_1 score	
MFCCs	51.1	
log mel-band energy	58.6	
	with DCT	**no DCT**
APGD	59.6	61.3
APGD + log mel-band energy	62.4	62.9

4.3. Evaluation Procedure

Block-wise F_1 score was used as the multi-label evaluation metric. It is calculated inside non-overlapping one-second blocks in the following manner. If an event is detected in one of the instances inside a block and is also present in the same block of the annotated data, that event is regarded as correctly detected. If it is *not* detected in any of the instances inside a block but is present in the same block of the annotated data, that event is regarded as missed. If it is detected in one of the instances inside a block but is *not* present in the same block of the annotated data, that event is regarded as false alarm.

For each one-second block, the numbers of correct, missed and false-alarm events are accumulated. Precision and recall are block-wise calculated and combined as their harmonic mean, the F_1 score.

4.4. Results

The value of LPC order in APGD calculation was set to 40 based on previous studies. To confirm its validity, a brief parameter search was performed. The results (Table 1) show that the value 40 is, indeed, a sound choice for the given problem, while a small variation of this parameter does not drastically deteriorate the performance.

Thereupon, the feature-wise evaluation was performed.The overall F_1 scores of the features, as well as their combination, without context windowing, are presented in Table 2. The proposed APGD feature demonstrates a moderate F_1 score improvement of 2.7 percentage points, and the combined features introduce further improvement of a total 4.3 percentage points over the baseline log mel-band energy features.

Compared to performing DCT on APGD features and discarding zeroth and higher (> 40) coefficients, seen in the previous works, an improvement of the performance was observed with the proposed "no DCT" method. The performance decrease with DCT is more visible when DCT is applied on log mel-band energies, which corresponds to MFCC features. Neural networks, indeed, do not require decorrelation of the features, and dimensionality reduction is not needed due to the sufficient amount of training data. To our knowledge, computing APGD without DCT has not been performed before, and the successfulness of this approach given the favourable aforementioned conditions is considered novel.

Table 3. F1 scores for different polyphony levels and MFCC, log mel-band energy and APGD features with DCT.

Polyphony	MFCC	log mel	APGD	APGD + log mel
1	51.9	60.5	63.2	65.5
2	47.4	57.9	59.8	62.0
3	48.4	59.0	60.8	63.1

Context windowing, not applicable to large feature vectors (APGD yields vectors of length 512), is omitted from the results in Table 2 for the sake of comparability. In the case of smaller feature vectors, however, context windowing introduced F_1 score gains of 1.0 (APGD with DCT), 2.2 (log mel-band energy) and 0.7 (combination) percentage points.

A further investigation is possible considering class-wise performance. Particularly, the proposed methodology introduced most noticeable improvement (F_1 score gains within the range 9–31 percentage points) with the classes: dish washer, motor noise, pressure release noise, refrigerator, road, wheel noise. The noticeable degradation of roughly nine percentage points was observed with classes "tick tock noise" and "keyboard". A possible explanation for such selectivity can be sought for by examining the physics of production of these sounds, however, within the given highly complex polyphonic evaluation scenario, such investigation appears hardly feasible.

The performance of the features on varying polyphony levels is presented in Table 3. The recordings have polyphony levels as high as nine, however, the amount of data for higher polyphony levels is not reliably high, therefore the first three polyphony level results are presented. The accuracy decreases expectedly with increasing polyphony, but the performance drop is not huge. This is a promising sign to use DNN classifiers with APGD or log mel-band energy features on polyphonic environmental sound event detection.

5. CONCLUSIONS

A phase-based APGD feature has been proposed for environmental sound event recognition. A DNN-based classifier incorporating the feature has been implemented and evaluated on a diverse realistic dataset of polyphonic sound events. The evaluation has demonstrated the improvement of the classifier's performance with the proposed feature over the baseline log mel-band energy features. By combining both features, a further F_1 score improvement has been observed.

The findings indicate the importance of phase for detecting a number of sound event classes. Studying the behaviour of the feature computed from clean monophonic recordings of these sounds in relation to the physics of their production is suggested as future work. Convolutional neural networks, seen as a natural solution for the phase invariance problem, are planned to be further investigated. Also, the possible correlation between neighbouring time frames of the sound events suggests studying the use of a recurrent neural network.

REFERENCES

[1] Tran Huy Dat, Ng Wen Zheng Terence, Jonathan William Dennis, and Leng Yi Ren, "Generalized Gaussian distribution Kullback-Leibler kernel for robust sound event recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5949–5953.

[2] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Moncef Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation.," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8677–8681.

[3] Onur Dikmen and Annamaria Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[4] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Polyphonic sound event detection using multilabel deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015.

[5] Courtenay V Cotton, Daniel PW Ellis, and Alexander C Loui, "Soundtrack classification by transient events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 473–476.

[6] Courtenay V Cotton and Daniel PW Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 69–72.

[7] Jens Schröder, Niko Moritz, Marc René Schädler, Benjamin Cauchi, Kamil Adiloglu, Jörn Anemuller, Simon Doclo, Birger Kollmeier, and Stefan Goetze, "On the use of spectro-temporal features for the IEEE AASP challenge 'Detection and classification of acoustic scenes and events'," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[8] Bayya Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[9] Padmanabhan Rajan, Tomi Kinnunen, Cemal Hanili, Jouni Pohjalainen, and Paavo Alku, "Using group delay functions from all-pole models for speaker recognition," in *Proc. Interspeech 2013*, 2013, pp. 2489–2493.

[10] Aleksandr Diment, Padmanabhan Rajan, Toni Heittola, and Tuomas Virtanen, "Group delay function from all-pole models for musical instrument recognition," in *Sound, Music, and Motion*, Mitsuko Aramaki, Olivier Derrien, Richard Kronland-Martinet, and Slvi Ystad, Eds., Lecture Notes in Computer Science, pp. 606–618. Springer International Publishing, 2014.

[11] Hema A. Murthy and Venkata Gadde, "The modified group delay function and its application to phoneme recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, vol. 1, pp. I–68–71 vol.1.

[12] André Holzapfel, Yannis Stylianou, Ali C. Gedik, and Barış Bozkurt, "Three dimensions of pitched instrument onset detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1517–1527, Aug 2010.

[13] Leigh D Alsteris and Kuldip K Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[14] Hideki Banno, Jinlin Lu, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol. 2, pp. 861–864.

[15] John Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[16] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio, "Maxout networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 1319–1327.

[17] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Audio context recognition using audio event histograms," in *Proc. of the 18th European Signal Processing Conference (EUSIPCO 2010)*, 2010, pp. 1272–1276.