SOUND EVENT DETECTION VIA DILATED CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Yanxiong Li[*], *Mingle Liu*[*], *Konstantinos Drossos*[†], *Tuomas Virtanen*[†]

[*] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

[†] Audio Research Group, Tampere University, Tampere, Finland

Introduction

The goal of SED is to detect the activity of target sound events in audio recordings. SED can be applied to many areas related to machine listening, such as traffic monitoring, smart meeting room, automatic assistance driving, and multimedia analysis. In this paper, we propose to use a dilated CRNN, namely a CRNN with a dilated convolutional kernel, as the classifier for the task of SED.

The RNNs have been proven quite promising for SED since they are able to effectively model the temporal context of sound events. The RNN has been combined with convolutional layers, resulting in a CRNN which achieved state-of-the-art results in detecting sound events. The most important is that the dilated convolutional kernel is able to increase the size of the receptive field without introducing extra parameters. Hence, to obtain the same size of receptive field, the CRNN with dilated convolutional kernels (called dilated CRNN) uses much less layers than the CRNN with conventional convolutional kernels (called baseline CRNN), which is able to avoid the overfitting problem caused by deeper networks with a lot of parameters. In addition, with the same number of parameters, the networks with the dilated convolutions are able to capture longer temporal context than that with the conventional convolutions. Modeling the long temporal context by the dilated convolutions has proved to be helpful for improving the performance of some tasks.

In this paper, we propose a sound event detection system which uses dilated convolutions in order to model the long temporal context. The contributions of this study are as follows. First, we propose to use a dilated CRNN as classifier fed by the feature of log mel-band energies for SED. Second, we conduct experiments on three public audio corpora to compare the performance of dilated CRNN with that of baseline CRNN. We find that the dilated CRNNs constantly outperform the equivalent baseline CRNNs and dilation operation in convolutional layers is helpful for improving the performance of the classifier of the CRNN.

Method

The features used in this work are log mel-band energies whose extraction is depicted in Fig. 1, whereas dilated CRNN is proposed to be used as the classifier.

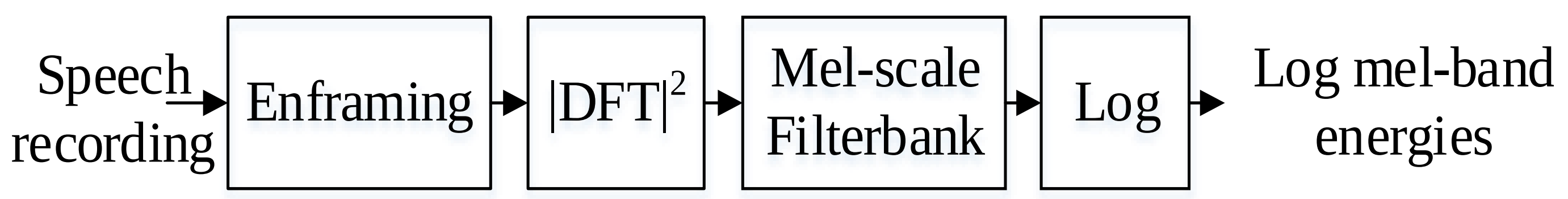
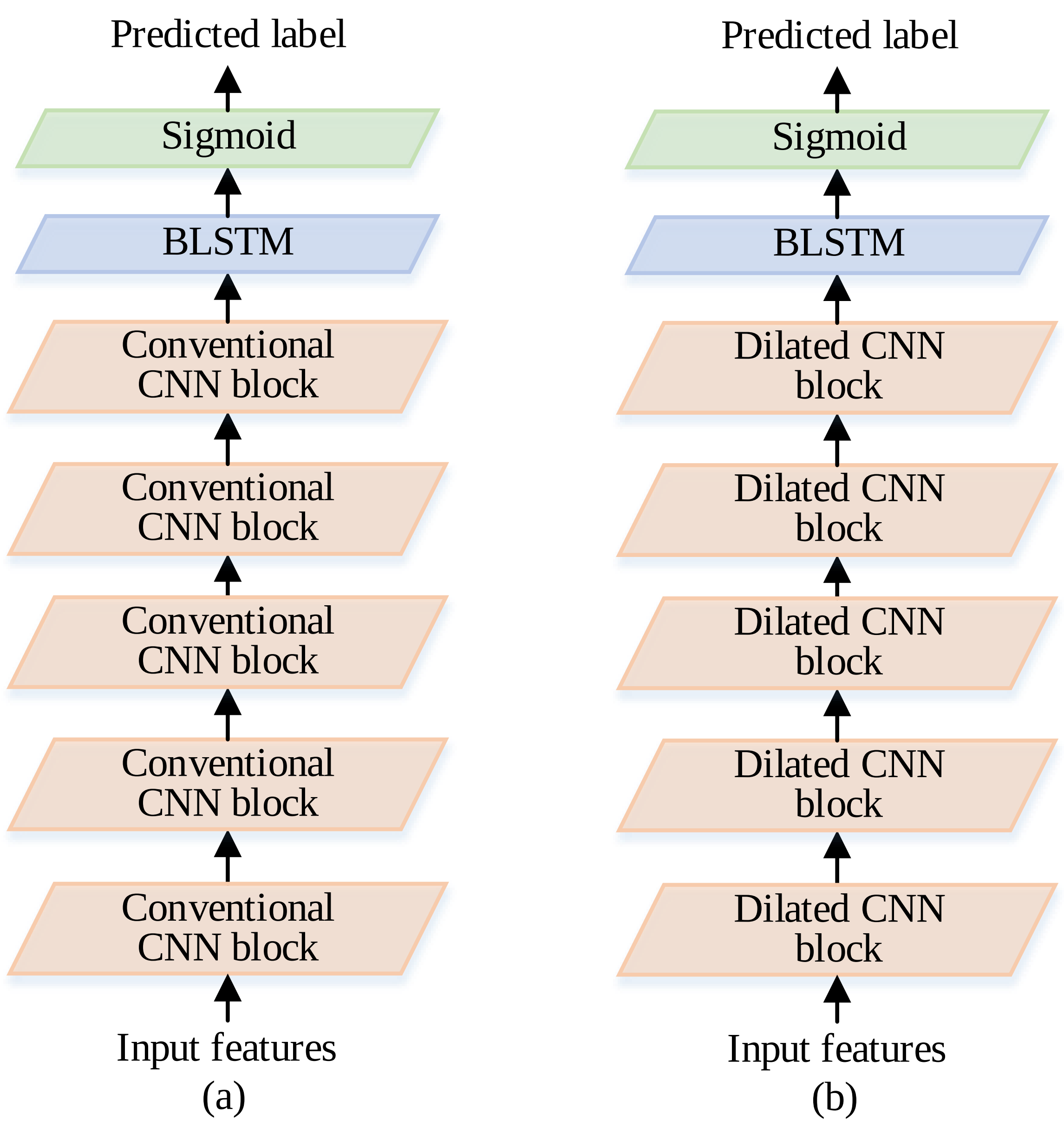

Speech recording → Enframing → $|\mathrm{DFT}|^2$ → Mel-scale Filterbank → Log → Log mel-band energies

Fig. 1 The diagram for extracting features.

Baseline CRNN

In a baseline CRNN, the size of the receptive fields is increased along layers via conventional convolutions. The baseline CRNN used in this work consists of conventional CNN blocks, bidirectional long short term memory (BLSTM) block, and sigmoid layer, as shown in Fig. 2 (a). The conventional CNN block used in this work is composed of a conventional convolutional operation, a rectification linear activation function, a pooling function, an operation of batch normalization, and a dropout operation.


(a): Input features → Conventional CNN block → Conventional CNN block → Conventional CNN block → Conventional CNN block → Conventional CNN block → BLSTM → Sigmoid → Predicted label

(b): Input features → Dilated CNN block → Dilated CNN block → Dilated CNN block → Dilated CNN block → Dilated CNN block → BLSTM → Sigmoid → Predicted label

Fig. 2 The diagrams of: (a) baseline CRNN, and (b) dilated CRNN.

Dilated CRNN

The dilated CRNN used in this work is demonstrated in Fig. 2 (b), in which the dilated CNN block consists of a dilated convolutional operation, a rectification linear activation function, a pooling function, an operation of batch normalization, and a dropout operation. Dilated convolution was originally designed for wavelet decomposition. A two-dimensional discrete convolution operator $*$, which convolves the log mel-band energies F with kernel K of size $(2m+1)*(2m+1)$, is defined by

$$(F*k)(p)=\sum_{s+t=p}F(s)K(t) \qquad (1)$$

where $p,s\in\mathbb{Z}^2, t\in[-m,m]^2\cap\mathbb{Z}^2$, and $\mathbb{Z}$ stands for the set of integers. The dilated version of the convolutional operator $*$, marked $*_r$, is defined by

$$(F*_r k)(p)=\sum_{s+rt=p}F(s)K(t) \qquad (2)$$

where $p,s\in\mathbb{Z}^2, t\in[-m,m]^2\cap\mathbb{Z}^2$, r denotes a dilation rate, and $*_r$ is called a r-dilated convolution. Hence, conventional convolution can be regarded as a one-dilated convolution. Similarly, the expression of a one-dimension r-dilated convolution is the same to Eq. (2), except the domain of definition of the variables, i.e., $p,s\in\mathbb{Z}^2, t\in[-m,m]^2\cap\mathbb{Z}^2$ in the one-dimension r-dilated convolution. The stride of convolutions is set to 1 for keeping the time resolution the same as in the input.

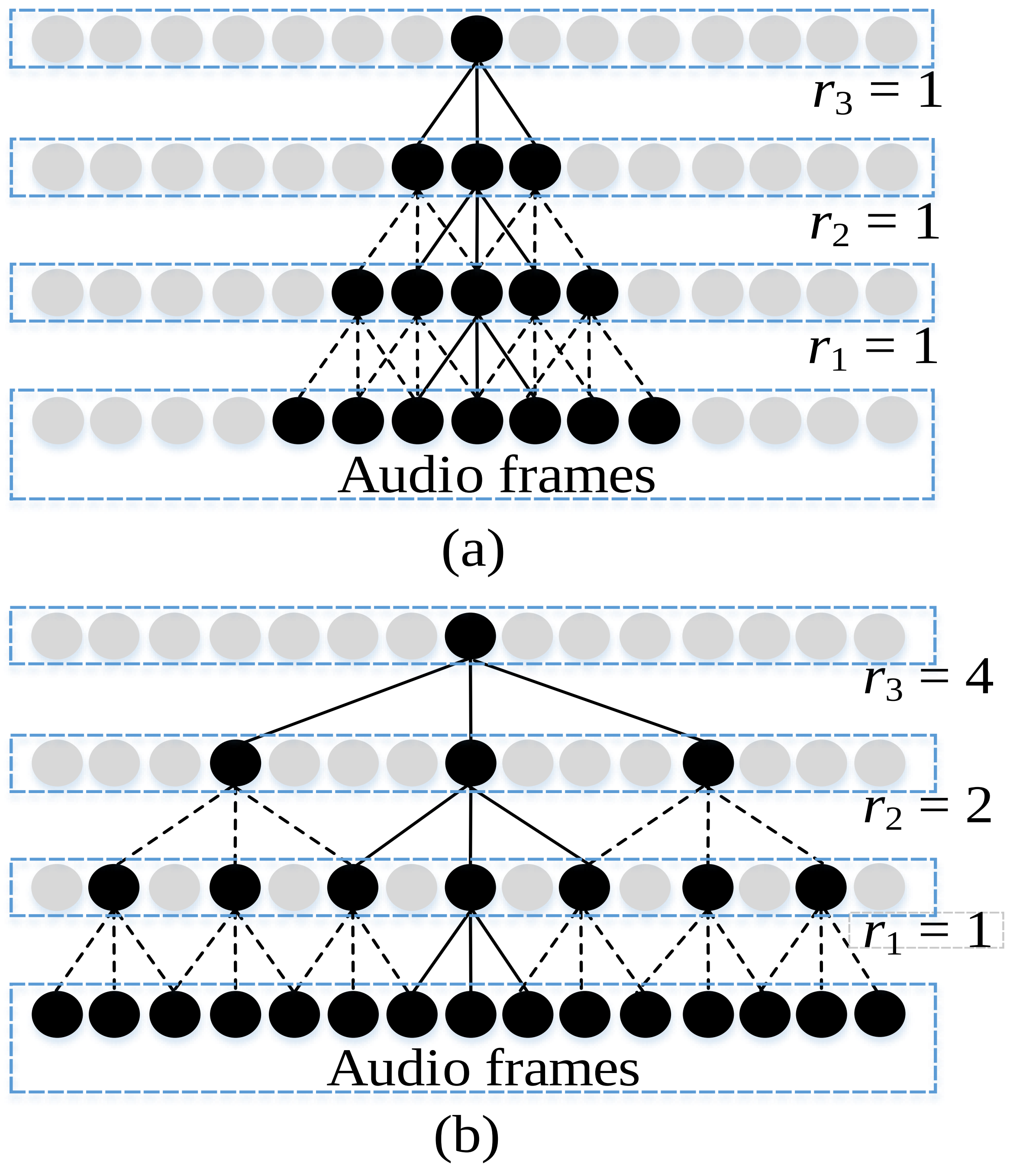


Fig. 3 (a) Conventional convolution, and (b) dilated convolution.

Experiments

We discuss the performance of individual CRNNs with different settings of convolutional layers for SED. The difference between baseline CRNN and dilated CRNN consists in the dilation rates only. That is, in each convolutional layer, the dilation rate is fixed to 1 for baseline CRNN and is variable for dilated CRNN. The F1 scores and ERs obtained by the CRNNs with different settings of convolutional layers are given in Tables 1 to 3 evaluated on the Synthetic 2016, the TUT Sound Events 2016, and the TUT Sound Events 2017, respectively. In the column of "Dilation rate", the digits denote dilation rates with the increase of number of convolutional layers.

Table. 1 The F1 scores and ERs of the baseline and dilated CRNNs.

	Baseline CRNN		Dilated CRNN	
Dataset	F1-Score	Error rate	F1-Score	Error rate
Synthetic 2016	58.20%	49.50%	59.40%	48.60%
DCASE 2016_H	41.10%	66.60%	42.80%	64.60%
DCASE 2016_R	63.40%	43.90%	65.40%	42.00%
DCASE 2017	58.20%	53.40%	60.70%	49.50%

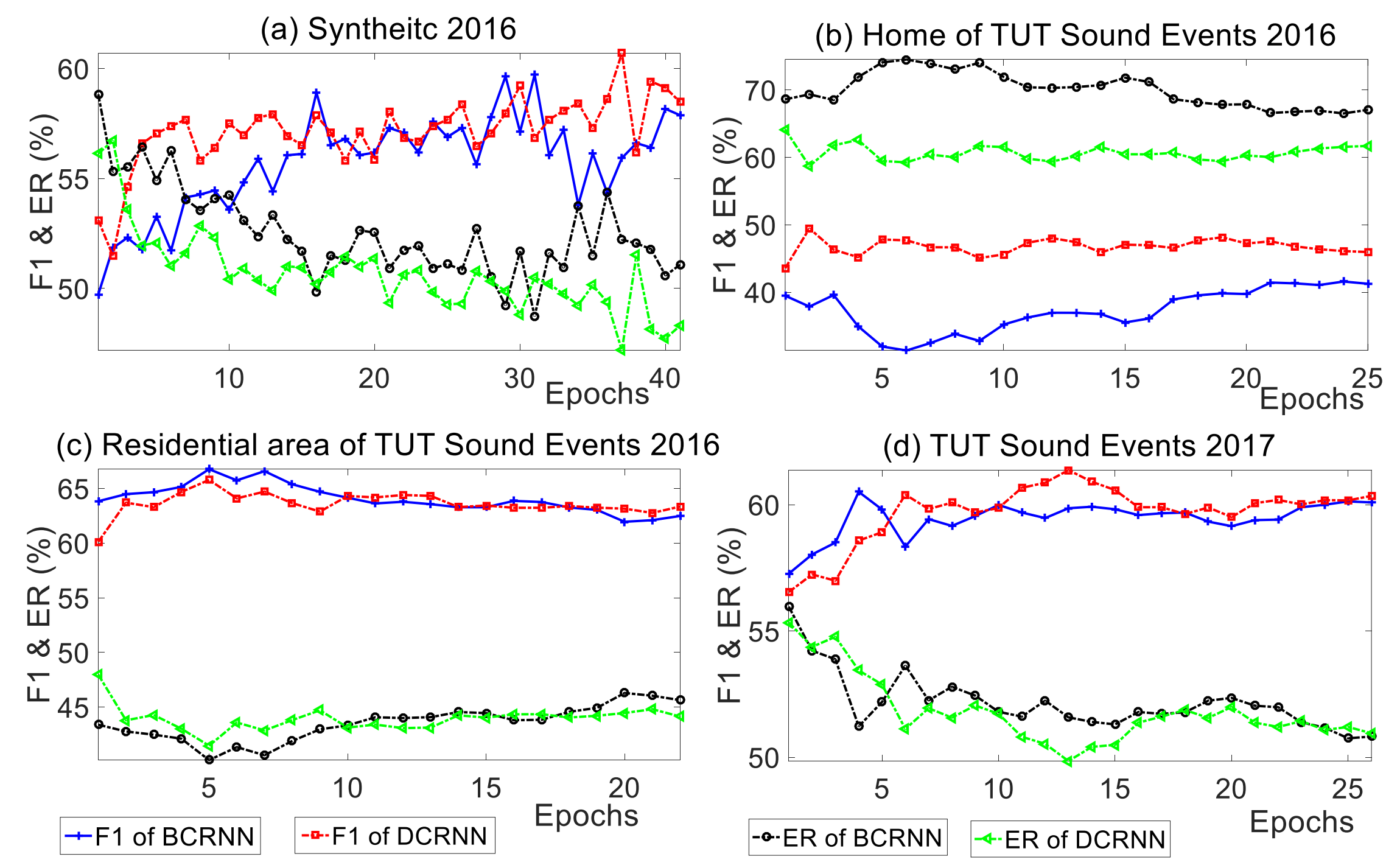


Fig. 4 The F1 scores and ERs of the baseline and dilated CRNNs.

Conclusions

In this work, we conducted a preliminary study for SED via investigating the effectiveness of dilation operations in convolutional layers of CRNNs. Evaluated on three public audio corpora, the experimental results showed that the dilated CRNNs always outperforms their equivalent baseline CRNNs in terms of both F1 score and ERs. That is, dilation operation with different rates has contribution to improve the performance of the CRNNs for the task of SED.