cochlear.ai

MARG — Music & Audio Research Group

Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks

Hyungui Lim[1], Jeongsoo Park[1,2], Kyogu Lee[2], Yoonchang Han[1]

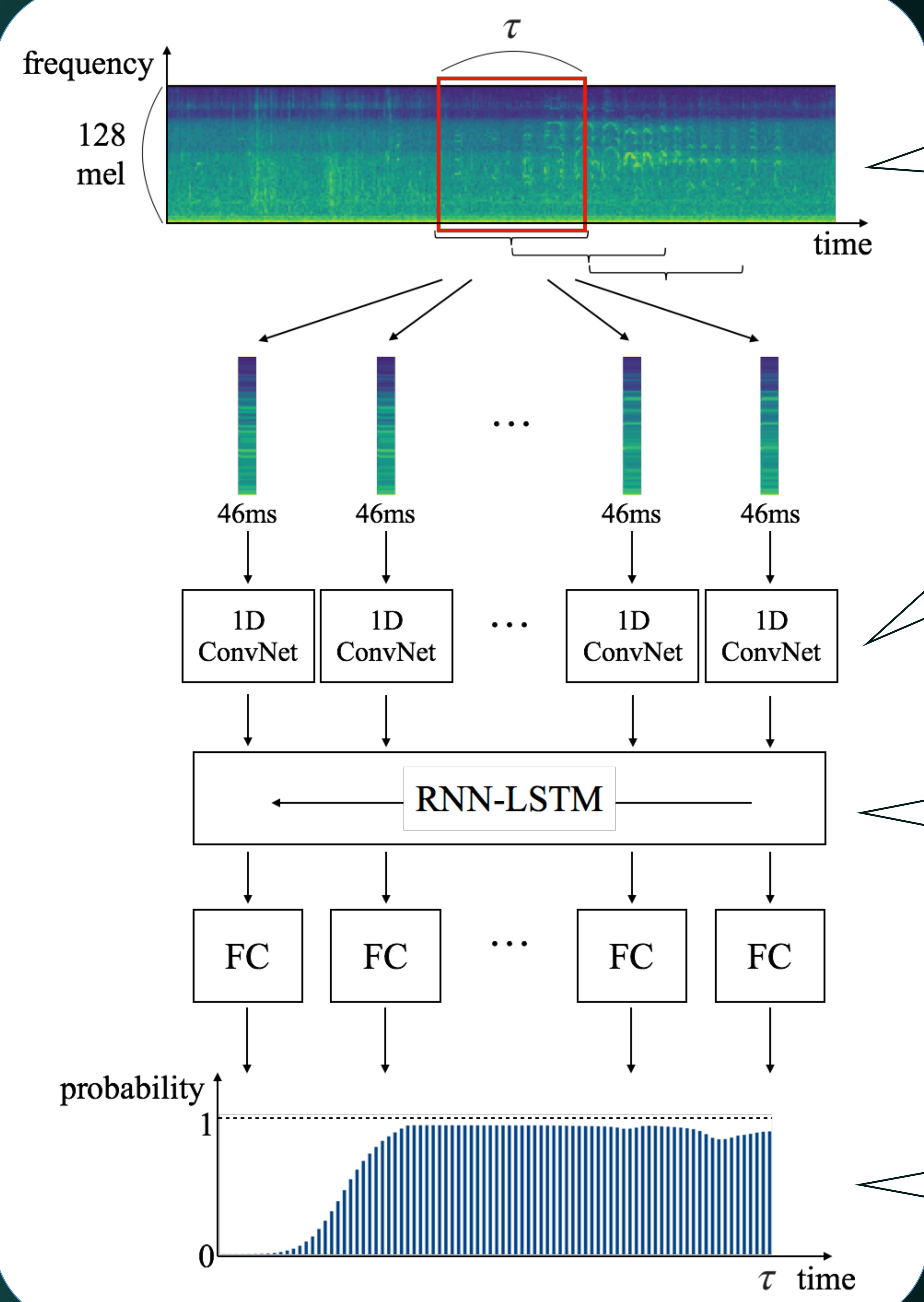
[1] Cochlear.ai, Seoul, Korea [2] Music and Audio Research Group, Seoul National University, Seoul, Korea

{hglim, jspark, ychan}@cochlear.ai, kglee@snu.ac.kr

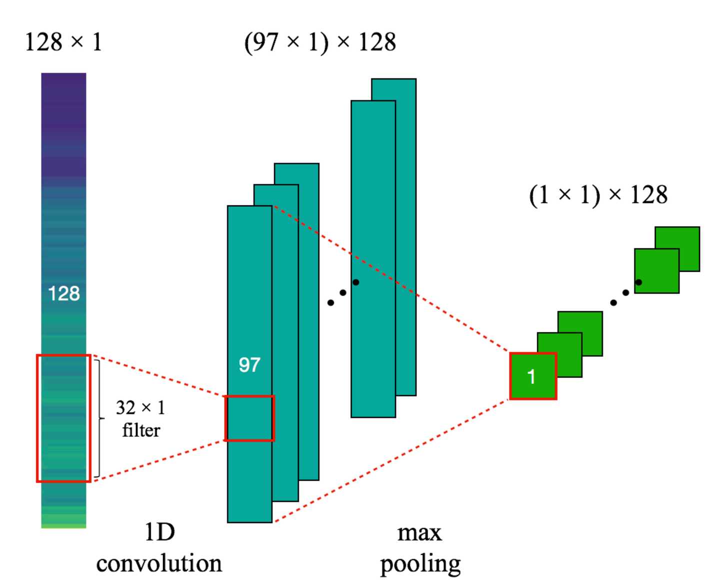
Introduction

- Rare sound event detection (**RSED**) task aims to detect certain emergency sounds (**baby crying, glass breaking, gunshot**) and their onset times precisely.
- We apply **1D CRNN** which is a combination of 1D convolutional neural network (1D ConvNet) and recurrent neural network (RNN) with long short-term memory units (LSTM) for each target event.
- Different input length (timestep) and different set of audio mixtures are applied to combine the results to imporve performance.

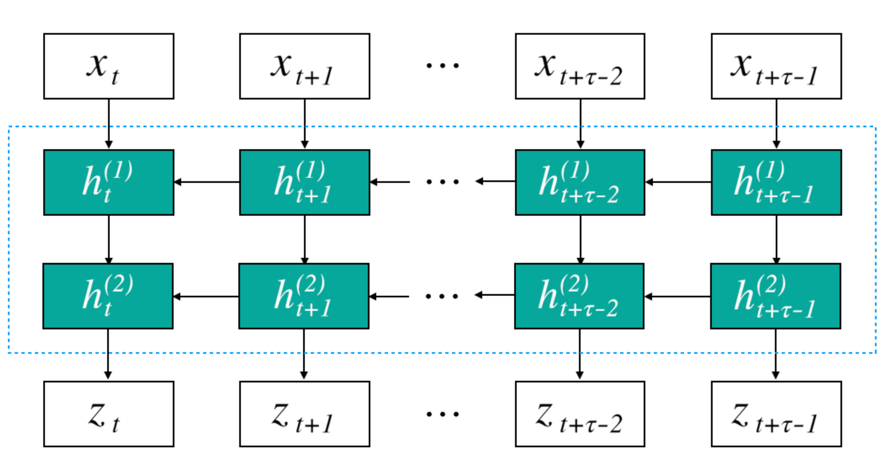
Proposed Method



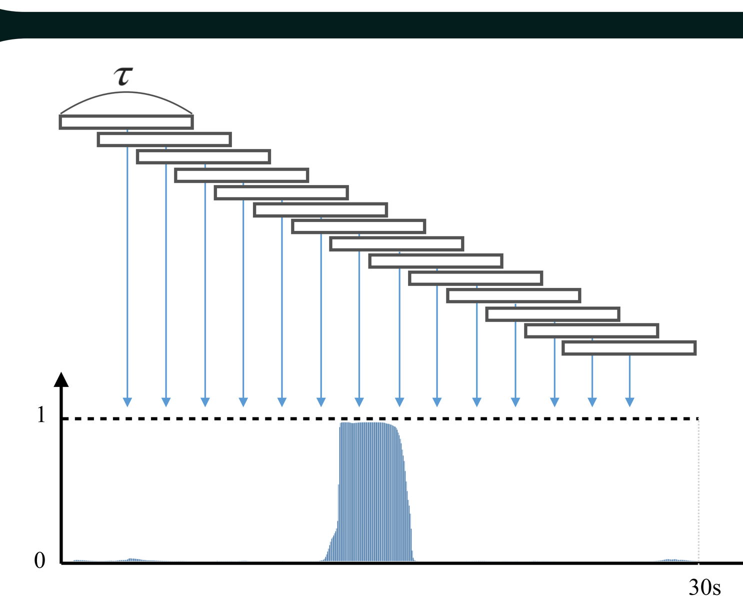
- Log-amplitude mel-spectrogram is extracted from audio signal.
  - window size: 46 ms / hop size : 23 ms / mel-filter banks : 128
- The mel-spectrogram is divided into a chunk with the size of a timestep (τ frames).
  - baby crying : 50, 100 / glass break : 5 / gunshot : 10, 14, 20, 50



- We apply spectral-side 1D ConvNet that enables frame-level investigation by filtering the spectral components of each frame.
  - filter size : 32 / # of filters : 128
  - Batch Normalization (BN)
  - activation : rectified linear unit (ReLU)
- Max-pooling is applied to each filter output to extract representative value.



- 128 features from the ConvNet pass through the RNN and are converted to 128 outputs for each frames.
- We apply unidirectional backward RNN-LSTM.
  - # of layers : 2 / # of units : 128
  - activation : hyperbolic tangent (tanh)



- The time-distributed output layer consists of one sigmoid unit. It represents a probability sequence of the target event during the timestep (τ).
- Sliding ensemble method combines the probability sequences by sliding the prediction chunk with a hop size of one frame.

Dataset

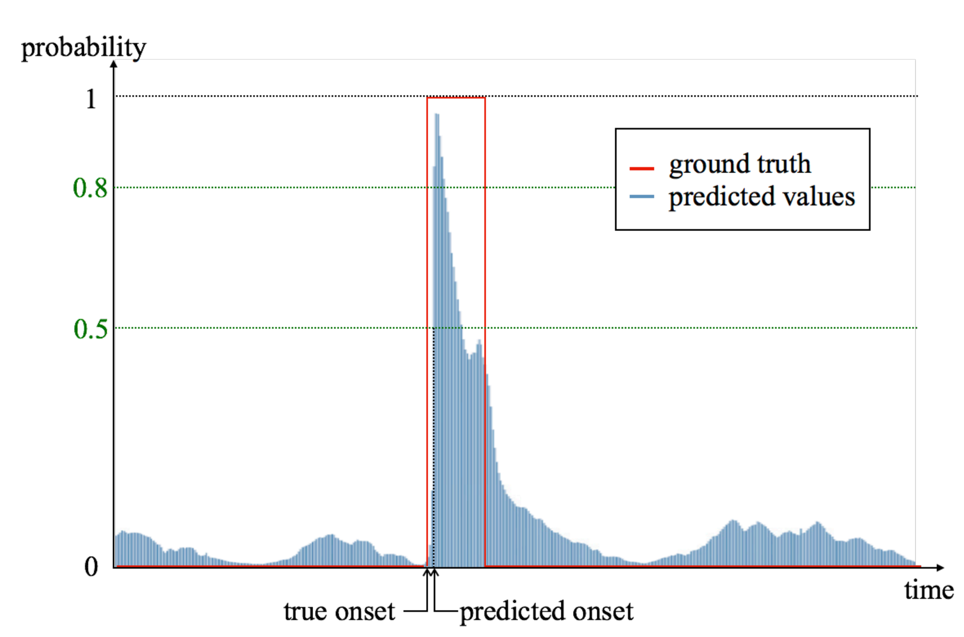
- Training set
  - 4 sets (S_1, S_2, S_3, S_4) of 15,000 synthesized audio mixtures (5,000 per event class)
- Test set
  - 1,500 given audio mixtures (500 per event class)

Ensemble method

Event	Ensemble method
Baby crying	$\left(p_1^{(100)} + 2p_2^{(50)} + p_3^{(50)} + p_3^{(100)}\right)/5$
Glass breaking	$\left(p_1^{(5)} + p_3^{(5)}\right)/2$
Gunshot	$\left(2p_1^{(14)} + p_1^{(50)} + p_3^{(10)} + p_4^{(10)} + p_4^{(20)}\right)/6$

- p_a^b: probability sequence calculated by the model using a mixture set of S_a and a timestep size of b.

Decision Making



- Presence of event
  - maximum probability value > 0.8 (0.5 for 'gunshot')
- Onset time of event
  - first index of the value greater than 0.5 before 50 frames (200 for 'baby crying') from the maximum probability value.

Results

	ER		F-score	
	dev	eval	dev	eval
Baby crying	0.05	0.15	97.6	92.2
Glass breaking	0.01	0.05	99.6	97.6
Gunshot	0.16	0.19	91.6	89.6
Overall	0.07	0.13	96.3	93.1

- The results have achieved the 1st place in the challenge.

Conclusion

- The approach of separating spectral/temporal processing using 1D ConvNet and RNN-LSTM has shown promising results.
- There are three main factors that improve performance.
  1. A large amount of synthesized audio data.
  2. Frame-wise detection which is effective in finding the precise onset time.
  3. The internal/external ensemble methods which reduce a lot of noise.