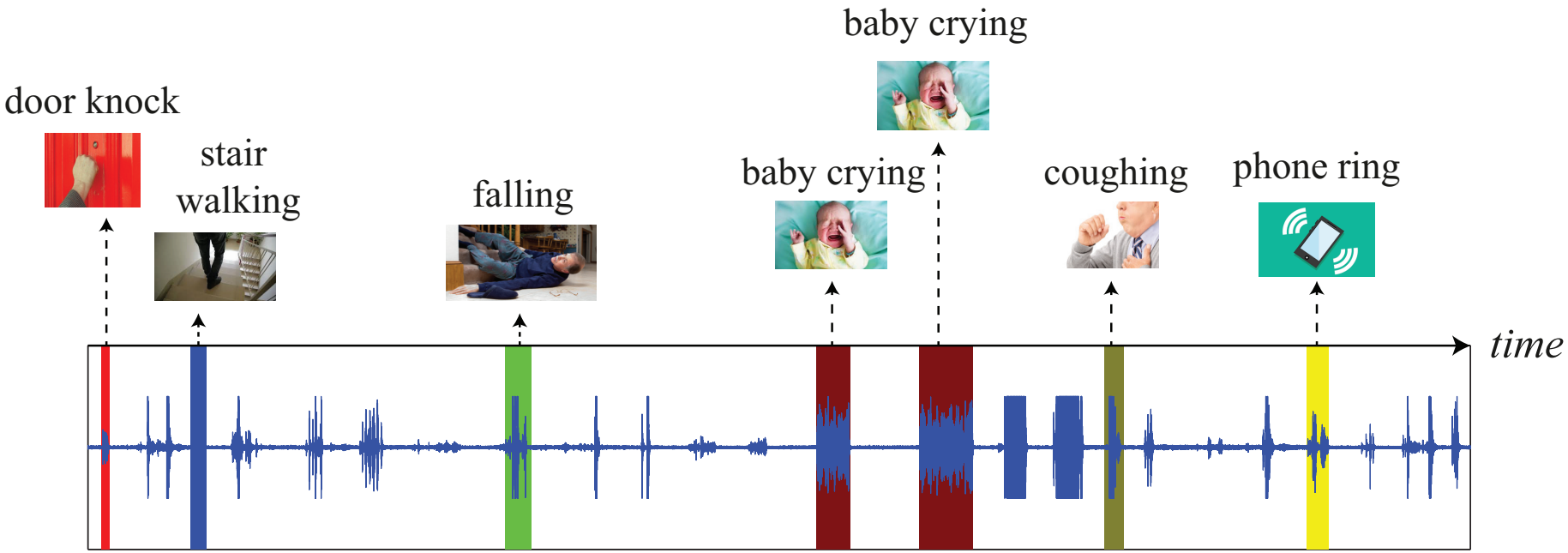
Modelling Overlapping Sound Events: a multi-label or multi-class problem

Huy Phan

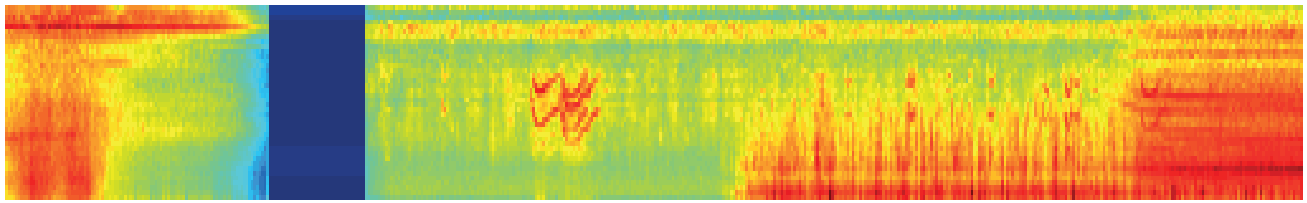
Centre for Digital Music (C4DM)
Queen Mary University of London

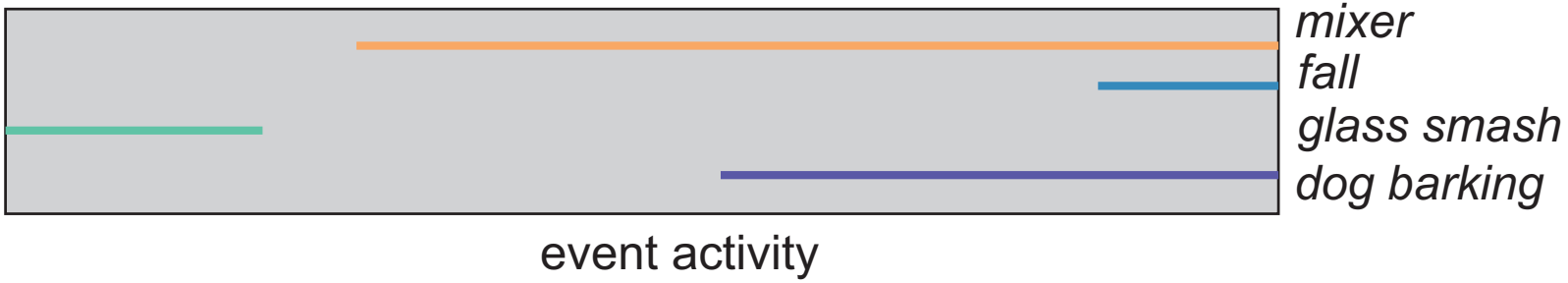
December 9, 2021

Audio event detection (AED)

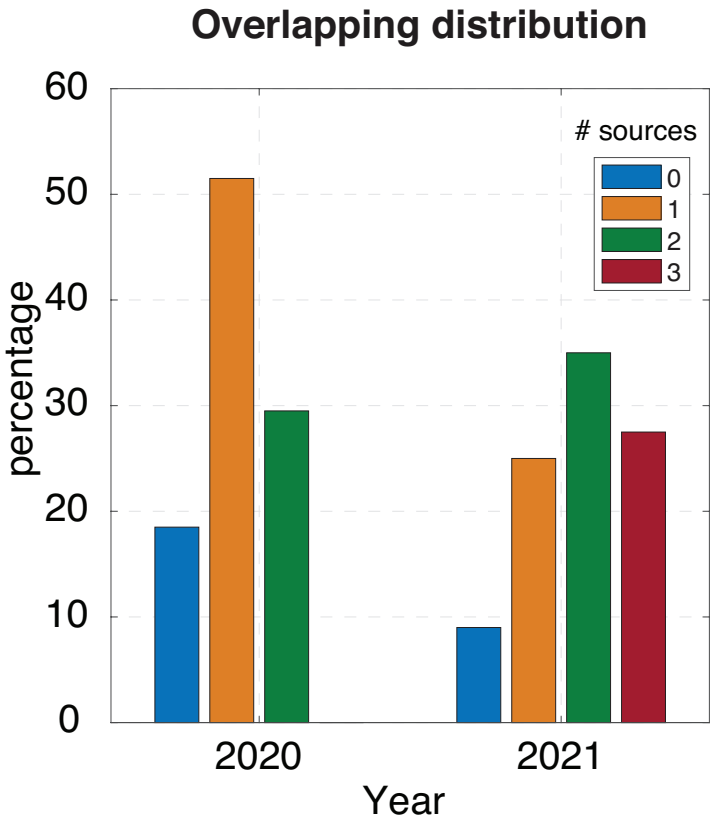


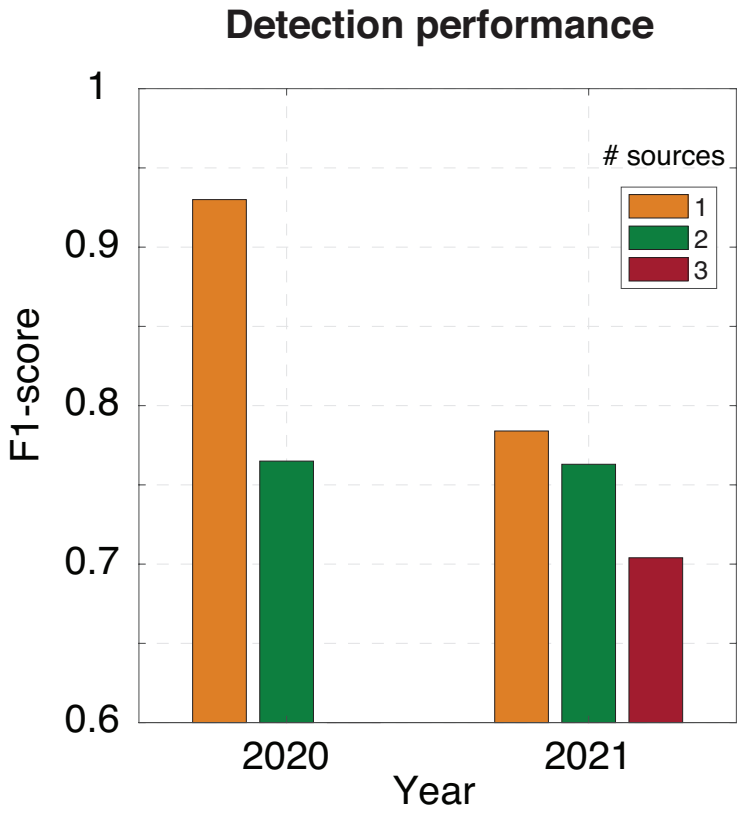
Overlapping audio events





AED in the DCASE challenge

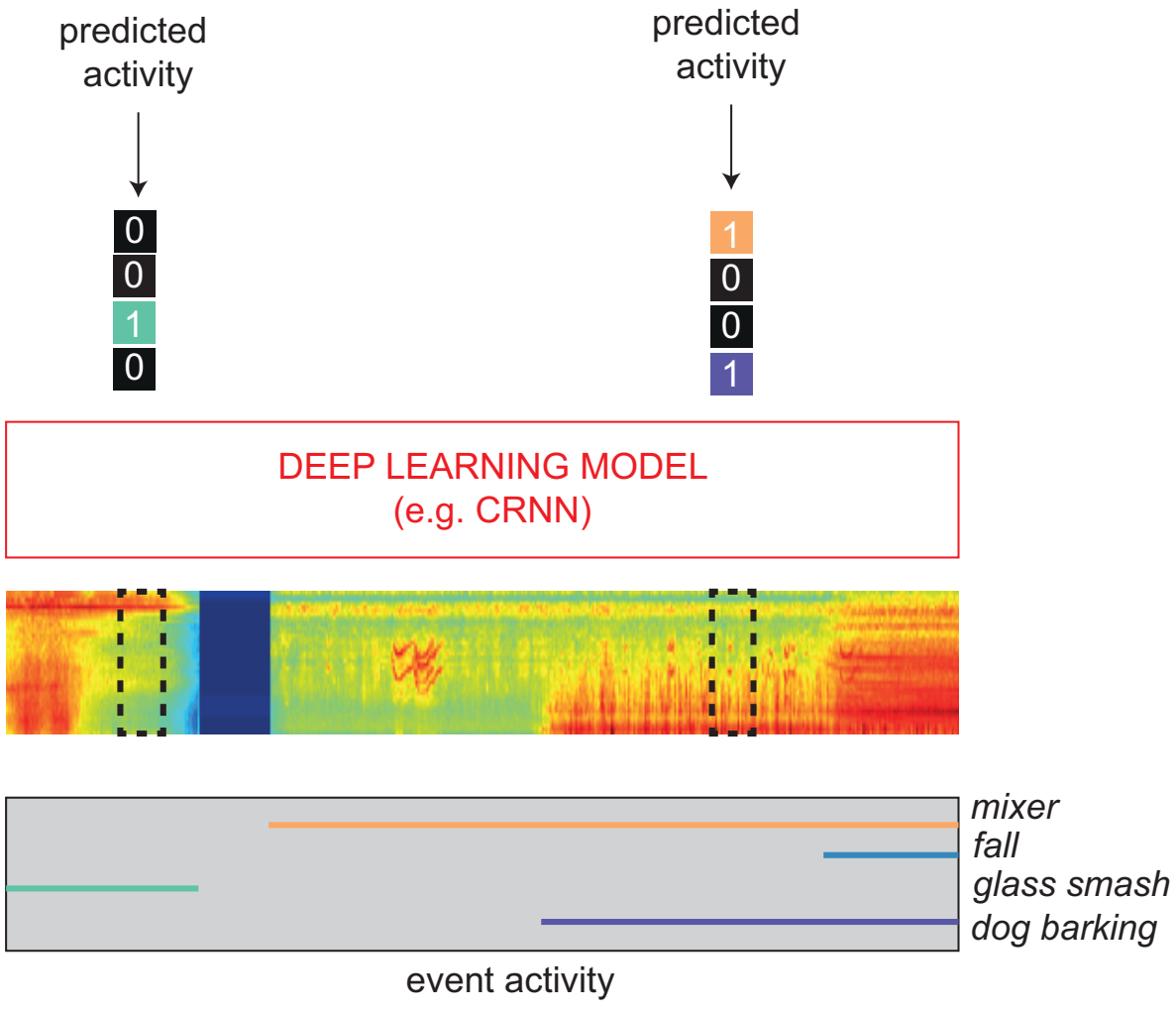




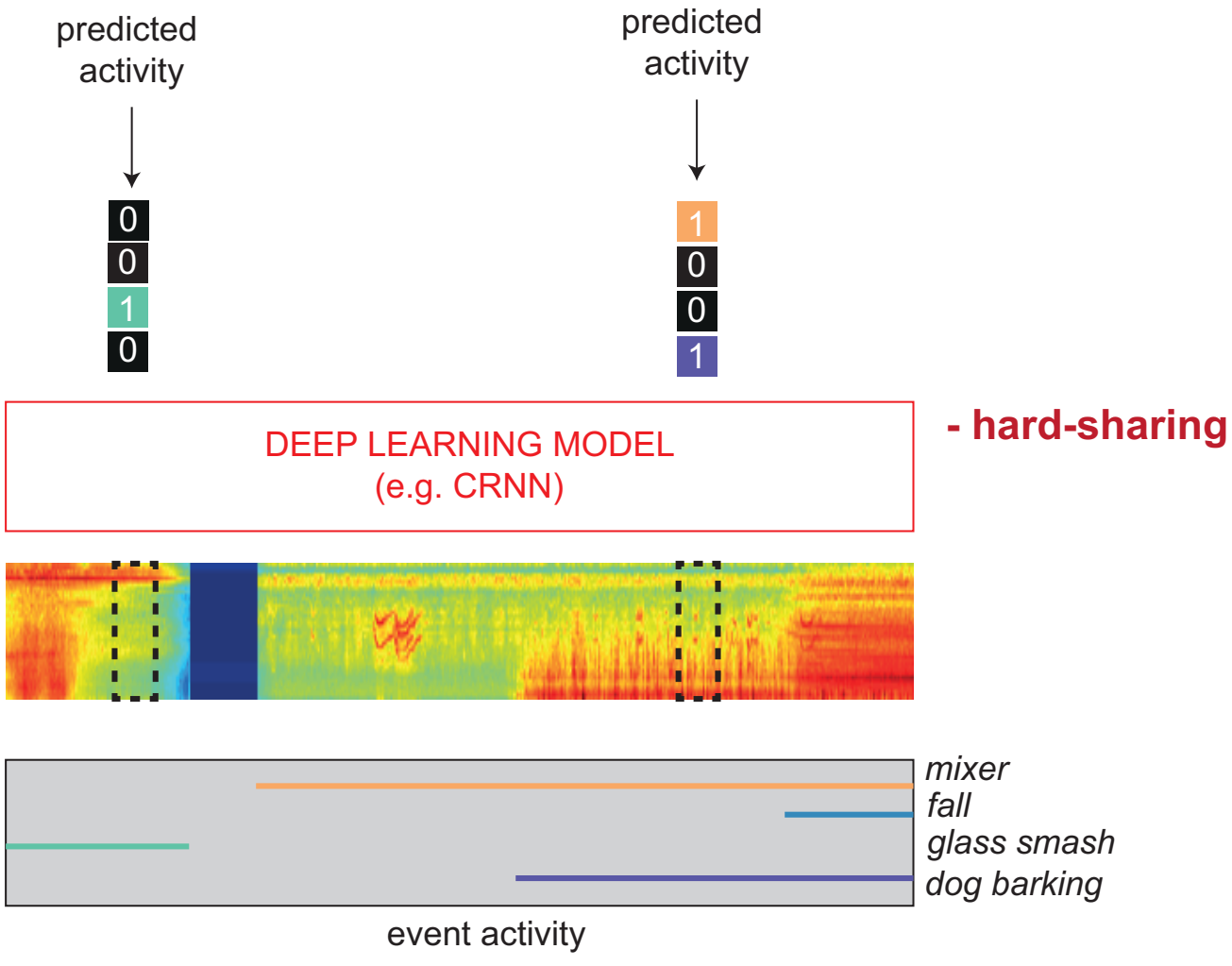
`http://dcase.community/`

Nguyen *et al.*, DCASE Workshop, 2021

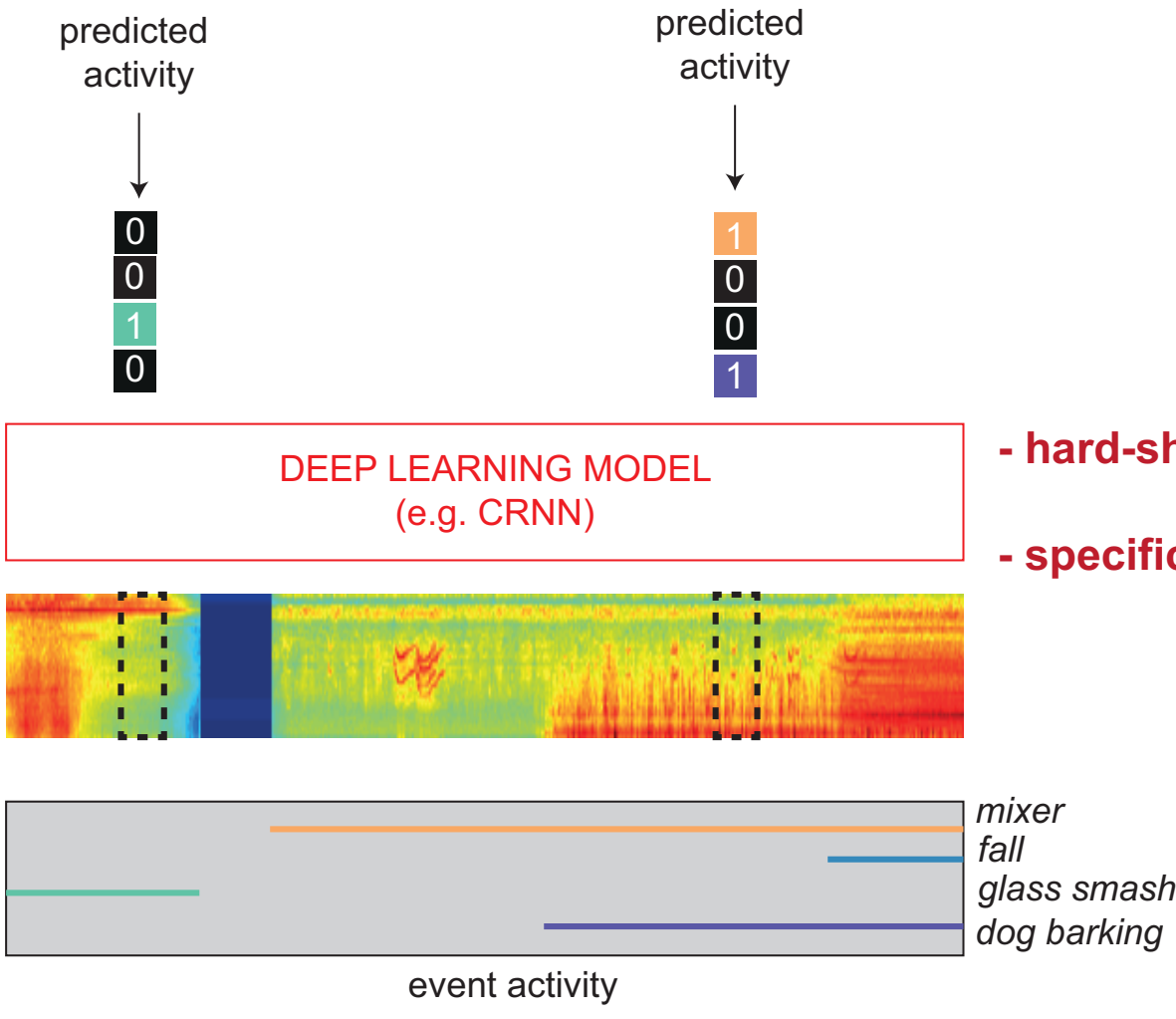
Multi-label modelling for (overlapping) AED



Multi-label modelling for (overlapping) AED



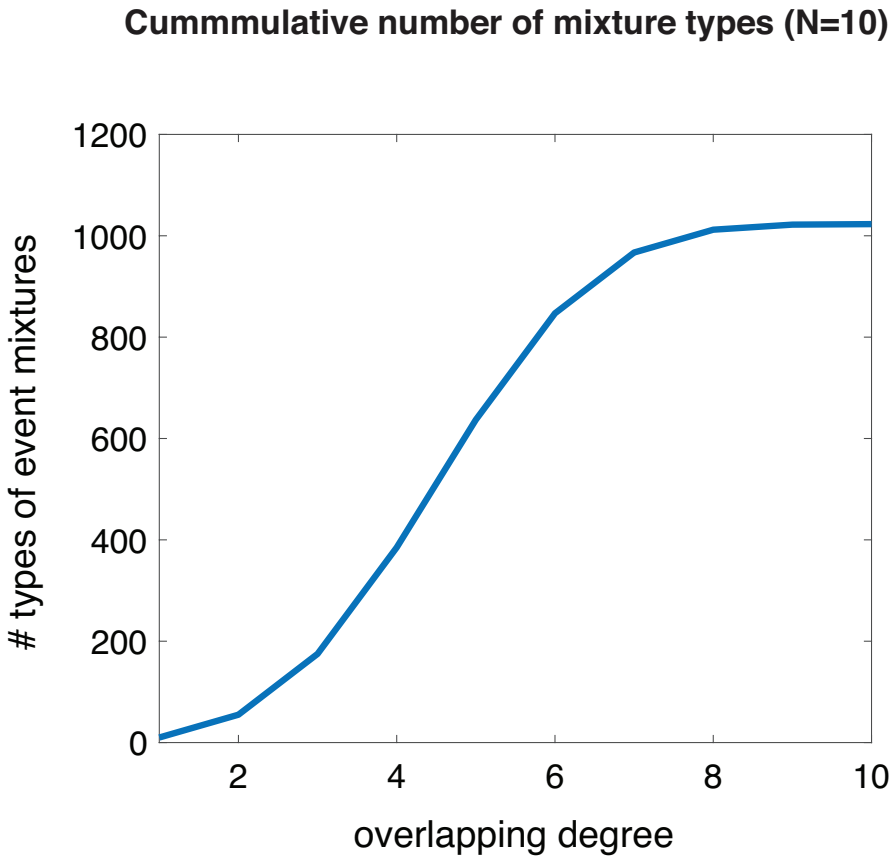
Multi-label modelling for (overlapping) AED



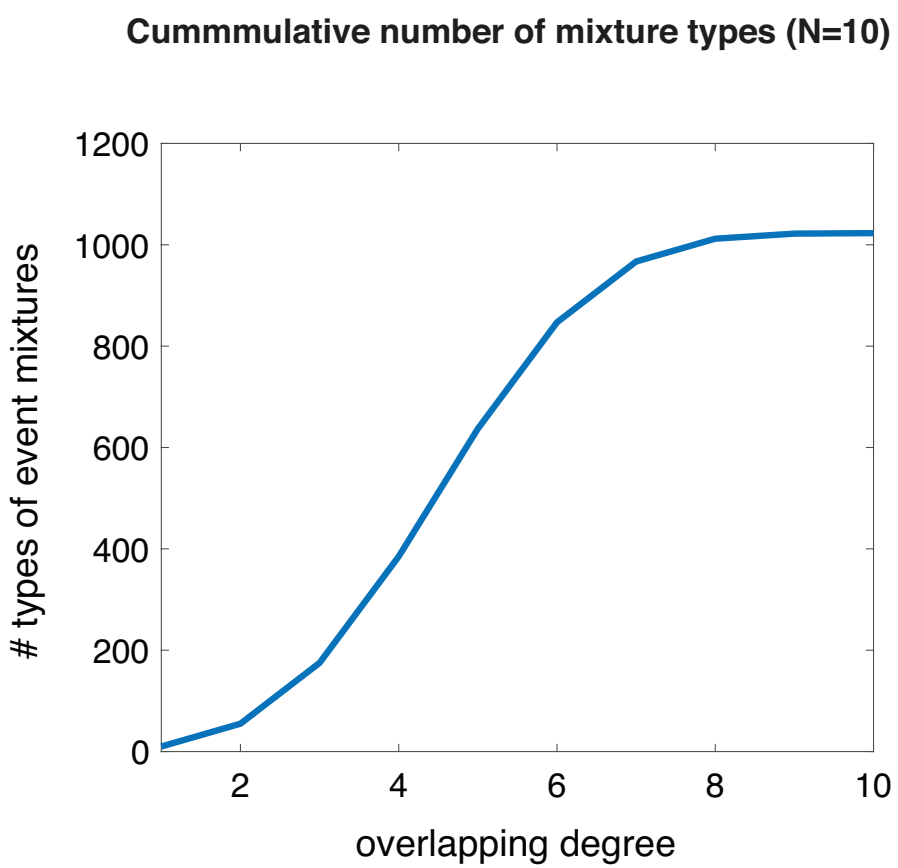
- hard-sharing

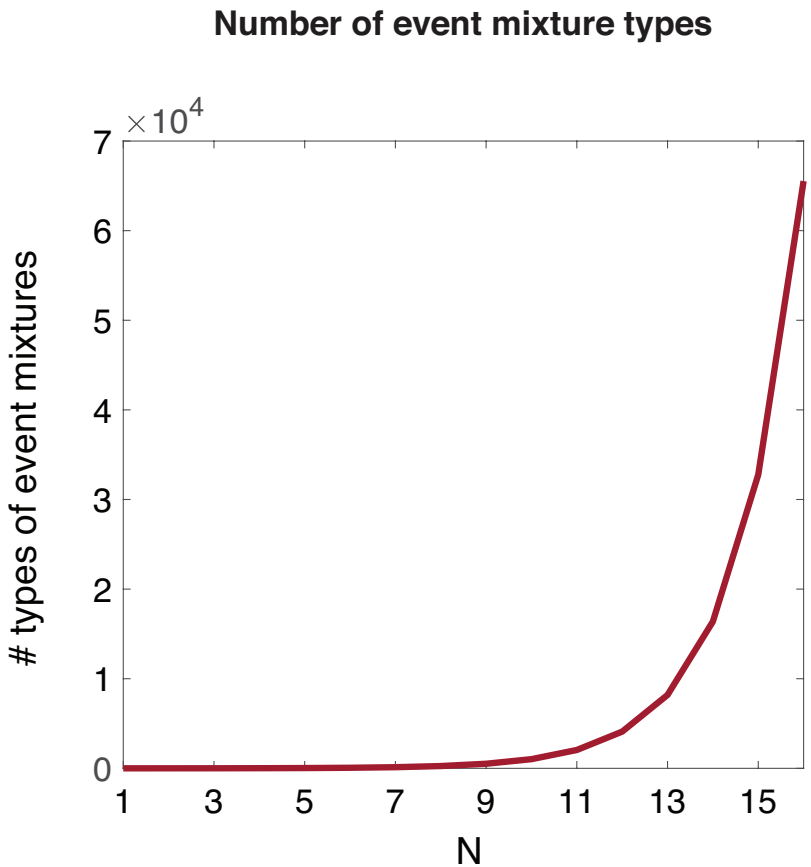
- specific to training mixtures

Exponential number of event mixtures



Exponential number of event mixtures





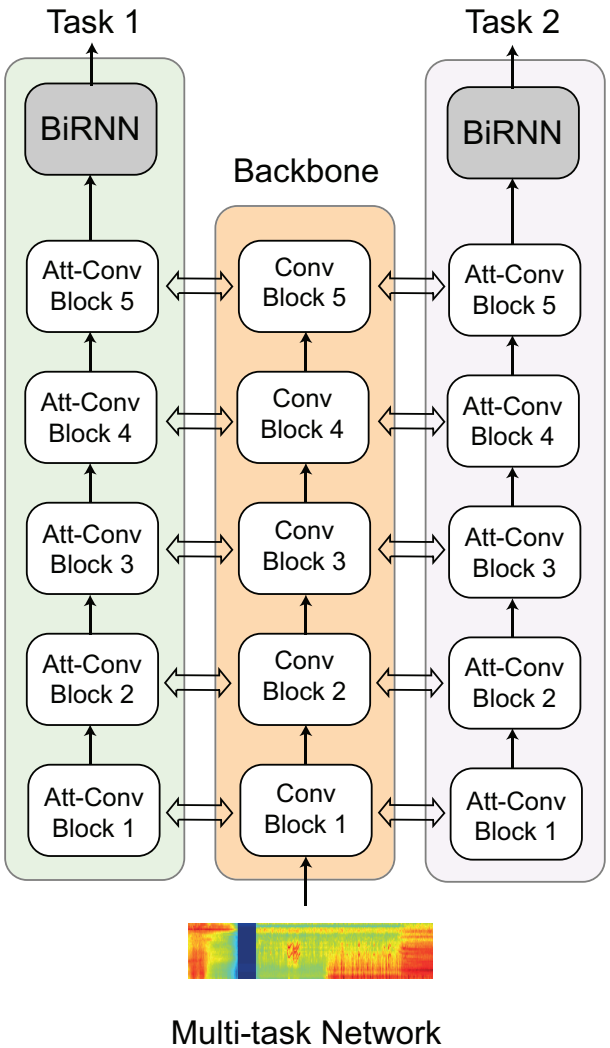
Multitask decomposition

- Decompose Y categories into N non-overlapping groups, $2 \leq N \leq Y$
- $\{Y_1, Y_2, \ldots, Y_N\}$ categories in the groups, $Y_1 + Y_2 + \ldots + Y_N = Y$
- Treat detection of event categories in a group as a task
  $\Rightarrow$ Breaking down the original task into N smaller tasks
- For n-th task, $2^{Y_n} << 2^Y$
  $\Rightarrow$ Allowing to model all possible mixture types as **multi-class** classification problem

Example

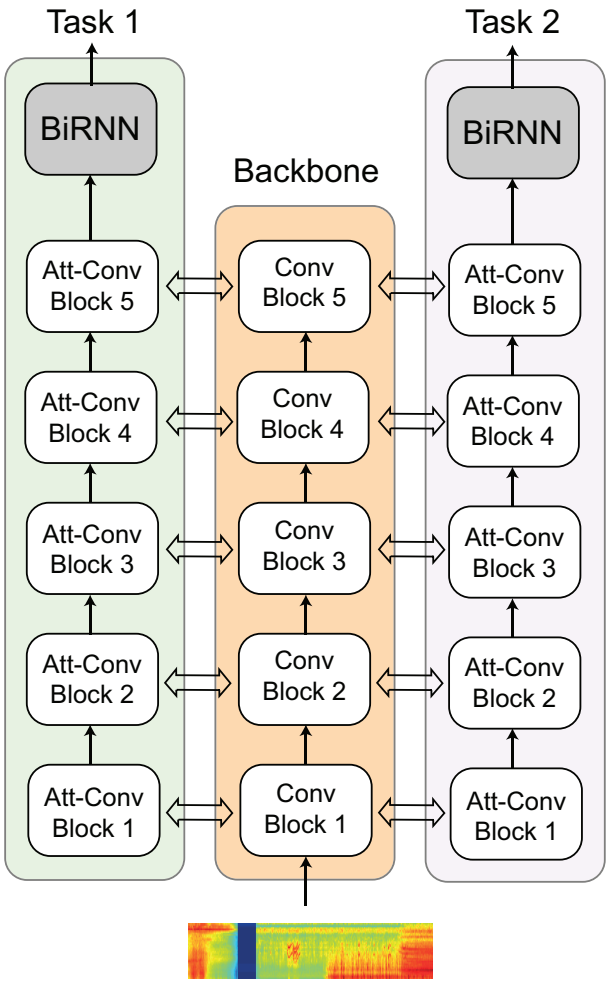
- Six categories: *baby crying, dog barking, cat meowing, footsteps, phone ringing, mixer*
- {baby crying, dog barking}
  - Classes: baby crying; dog barking; baby crying $\oplus$ dog barking
- {cat meowing, footsteps}
  - Classes: cat meowing; footsteps; cat meowing $\oplus$ footsteps
- {phone ringing, mixer}
  - Classes: phone ringing; mixer; phone ringing $\oplus$ mixer

Multitask modelling

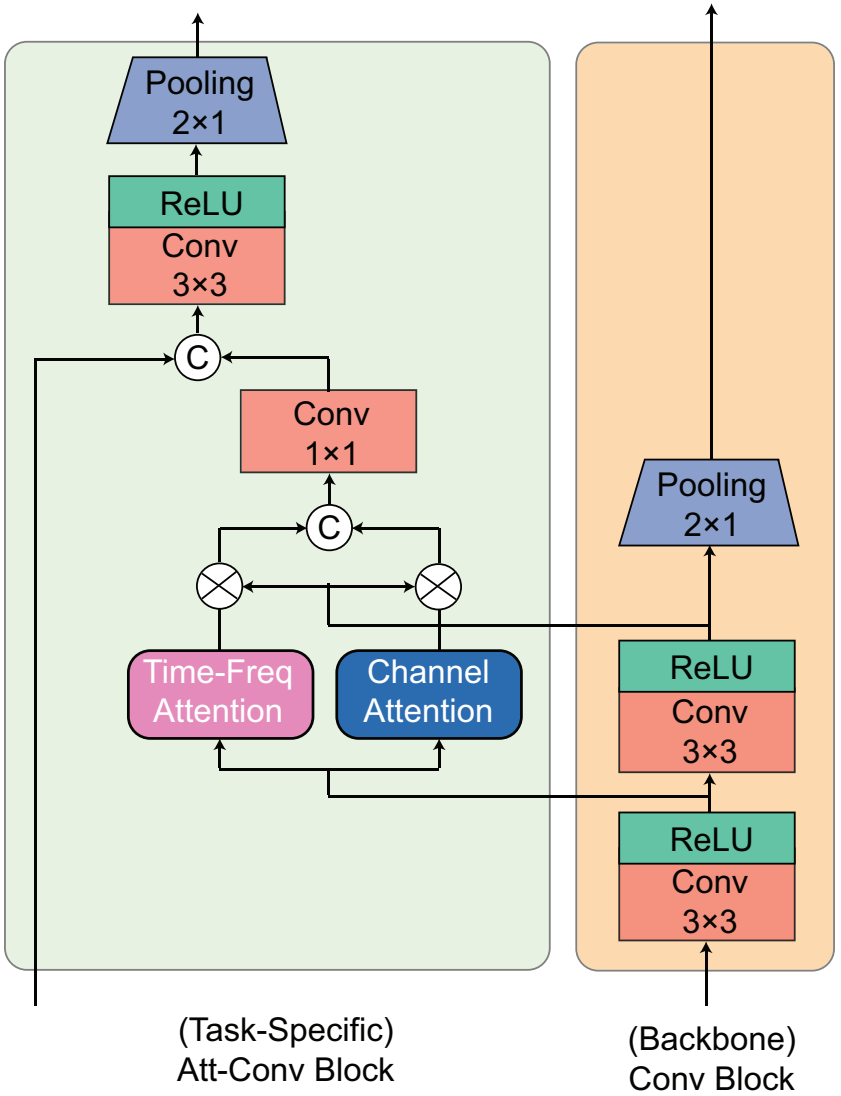


Multi-task Network

Multitask modelling



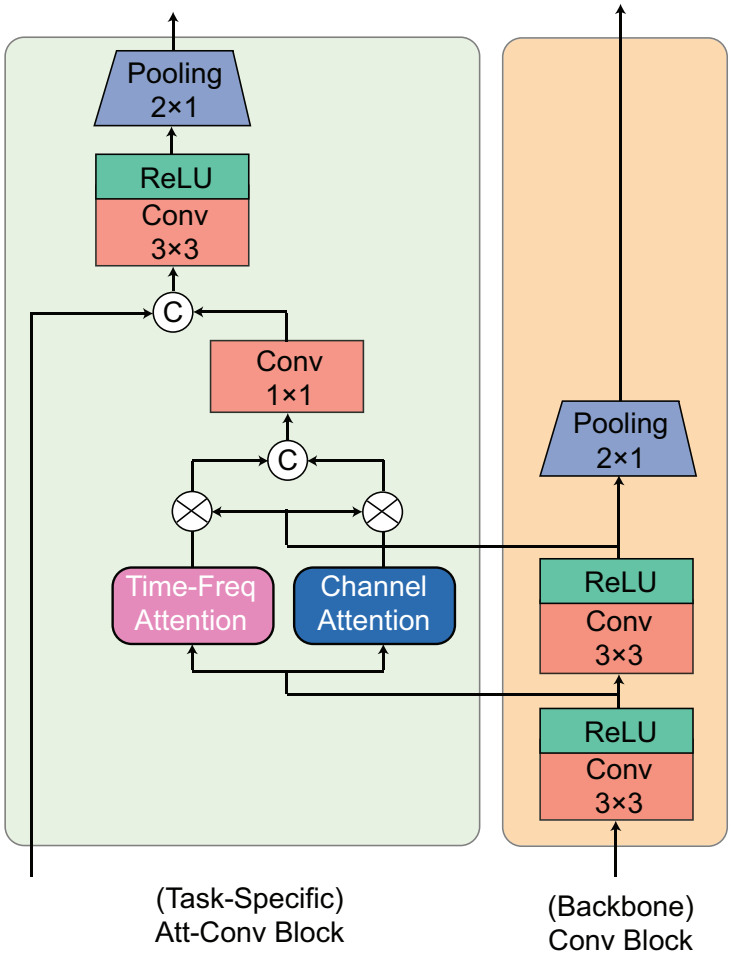
Multi-task Network



(Task-Specific) Att-Conv Block

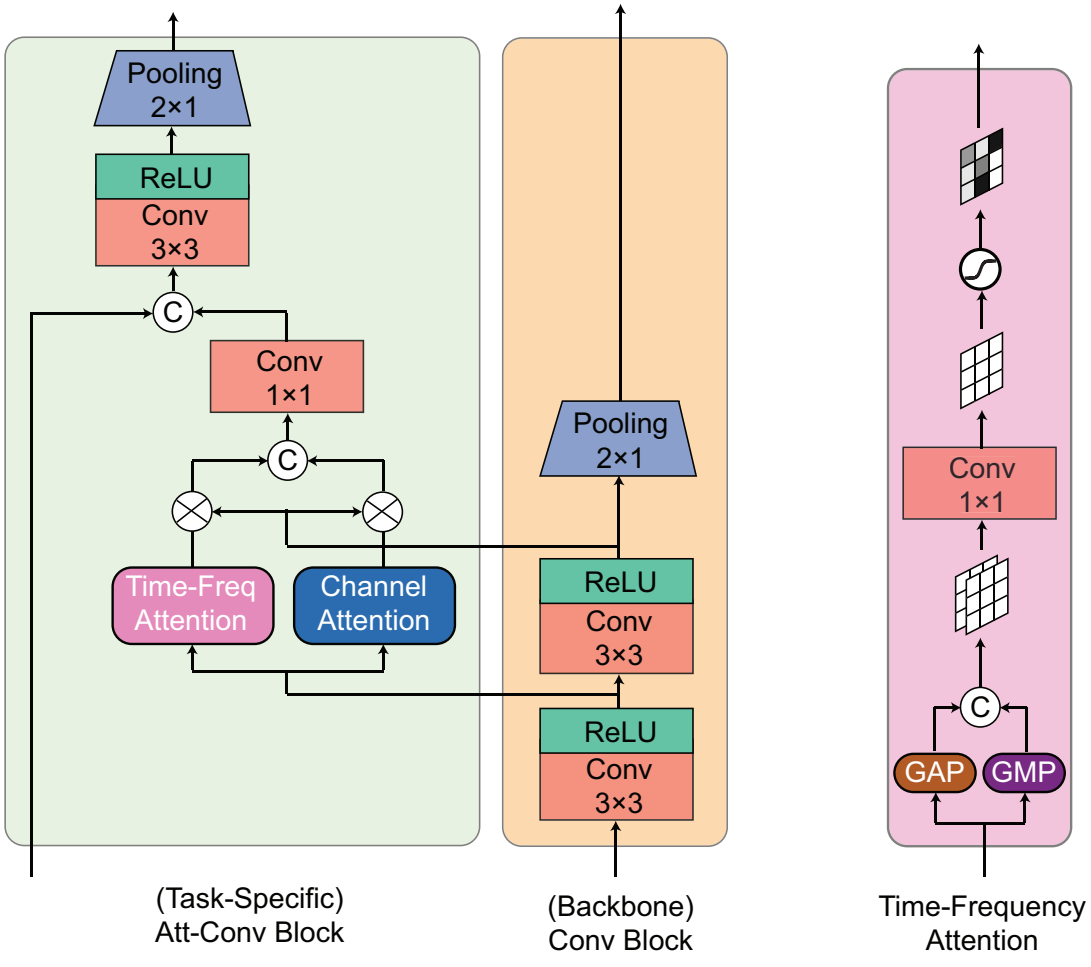
(Backbone) Conv Block

Time-frequency and channel attentions



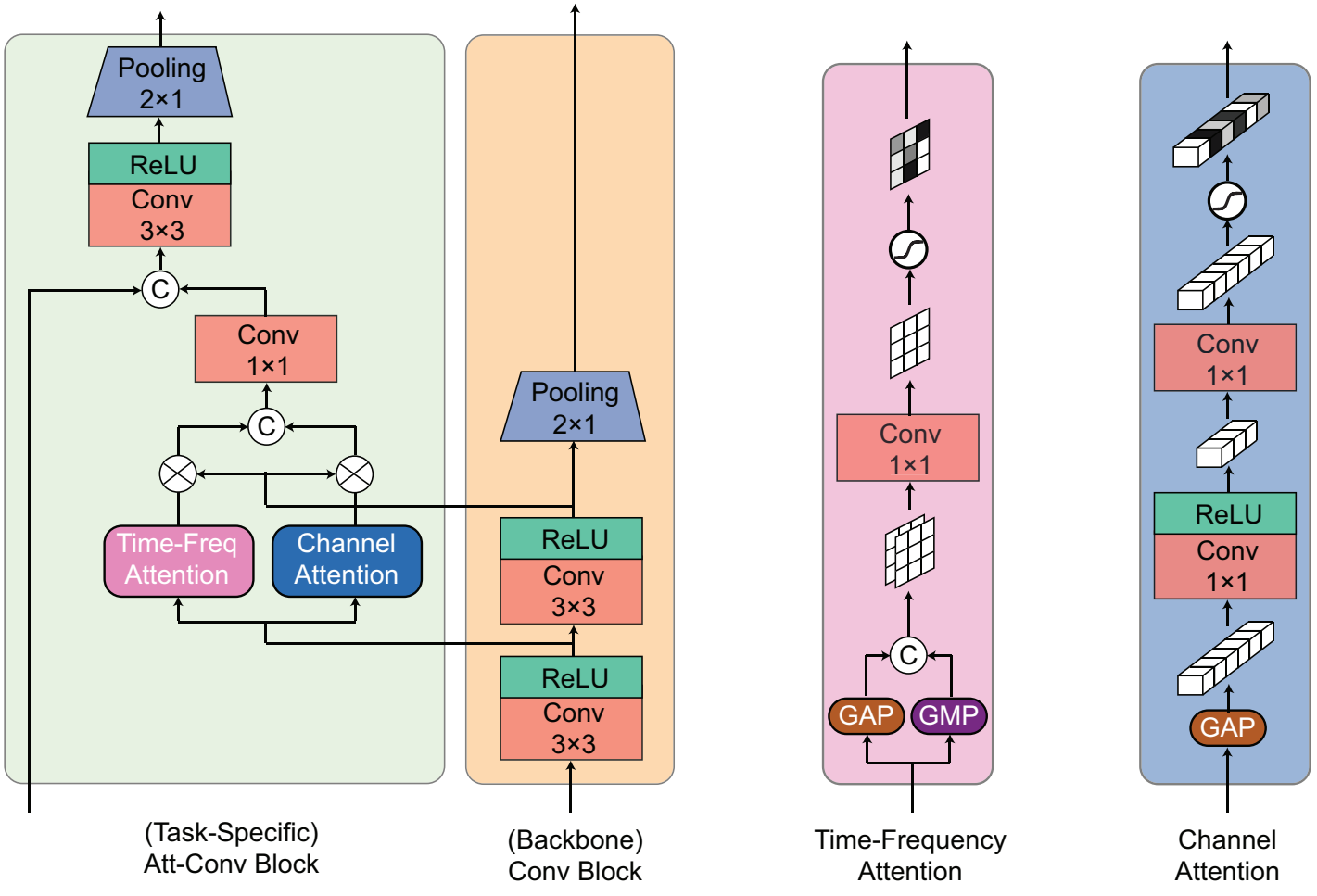
$$\mathbf{M}^* = (\mathbf{m}_{\text{tf}} \otimes \mathbf{M}_2^{\text{bb}}) \oplus (\mathbf{m}_{\text{c}} \otimes \mathbf{M}_2^{\text{bb}})$$

Time-frequency and channel attentions



$$\mathbf{M}^* = (\mathbf{m}_{\text{tf}} \otimes \mathbf{M}_2^{\text{bb}}) \oplus (\mathbf{m}_{\text{c}} \otimes \mathbf{M}_2^{\text{bb}})$$

Time-frequency and channel attentions



$$\mathbf{M}^* = (\mathbf{m}_{\text{tf}} \otimes \mathbf{M}_2^{\text{bb}}) \oplus (\mathbf{m}_{\text{c}} \otimes \mathbf{M}_2^{\text{bb}})$$

TUT-SED-Synthetic-2016 dataset

- 16 event categories
- 100 overlapping mixtures created from 994 monophonic instances
  - 566 mins in total
  - 60 for training, 20 for validation, 20 for evaluation
- Maximum overlapping degree of 6

Cakir *et al.*, TASLP, 2017

Multi-task decomposition

Event categories	Multi-task			
	2 tasks	4 tasks	8 tasks	16 tasks
alarms & sirens (as)	as	as	as	as
baby crying (bc)	bc	bc	bc	bc
bird singing (bs)	bs	bs	bs	bs
bus	bus	bus	bus	bus
cat meowing (cm)	cm	cm	cm	cm
crowd applause (ca)	ca	ca	ca	ca
crowd cheering (cc)	cc	cc	cc	cc
dog barking (db)	db	db	db	db
footsteps (fs)	fs	fs	fs	fs
glass smash (gs)	gs	gs	gs	gs
gun shot (gsh)	gsh	gsh	gsh	gsh
horsewalk (hw)	hw	hw	hw	hw
mixer (mx)	mx	mx	mx	mx
motorcycle (mc)	mc	mc	mc	mc
rain	rain	rain	rain	rain
thunder (td)	td	td	td	td

Cakir *et al.*, TASLP, 2017

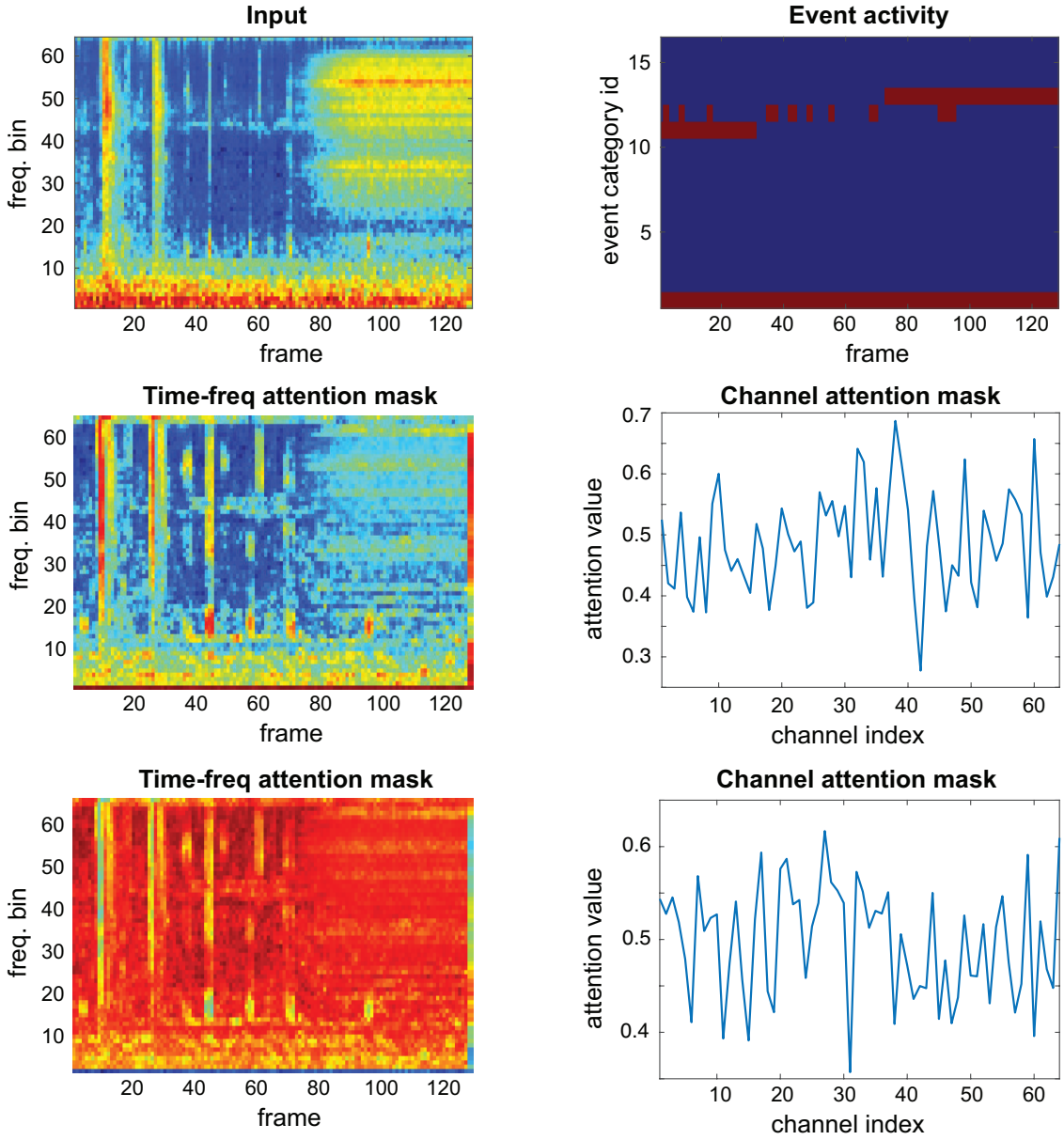
Frame-wise F1-scores results

Event type	Multi -label	Multi-task			
		2 tasks	4 tasks	8 tasks	16 tasks
alarms & sirens	65.1	**68.4**	**70.8**	**73.2**	**67.0**
baby crying	52.3	**56.5**	49.1	**54.8**	**55.3**
bird singing	49.4	**50.5**	**52.0**	48.3	45.8
bus	53.7	**53.9**	**58.4**	**65.0**	**58.8**
cat meowing	28.4	**46.8**	**49.0**	**44.1**	**48.0**
crowd applause	70.9	70.2	**73.5**	**73.1**	**73.3**
crowd cheering	69.6	**73.9**	**75.1**	**71.1**	**74.8**
dog barking	74.1	**76.6**	**77.3**	**78.2**	**78.4**
footsteps	45.3	**48.8**	**50.7**	**51.8**	**50.5**
glass smash	80.1	77.0	**82.0**	**82.6**	**83.9**
gun shot	79.7	67.9	76.3	**82.4**	**83.9**
horsewalk	44.7	44.6	**45.6**	44.3	44.2
mixer	69.9	**80.9**	**75.9**	**75.5**	**70.9**
motorcycle	50.8	43.2	46.1	**55.7**	**55.2**
rain	77.9	69.7	77.7	73.9	**82.5**
thunder	60.1	56.9	**62.2**	58.7	**61.3**
Average	60.8	**61.6**	**63.9**	**64.5**	**64.6**
Overall	63.1	**64.3**	**66.0**	**66.8**	**65.9**

Influence of overlapping degree

Overlapping degree	Multi -label	Multi-task			
		2 tasks	4 tasks	8 tasks	16 tasks
1	70.5	**71.4**	**73.8**	**74.0**	**73.4**
2	56.7	**58.2**	**59.8**	**60.3**	**62.0**
3	48.9	**49.0**	**52.1**	**53.7**	**52.8**
4	44.9	42.2	43.7	**48.1**	**47.1**
5	34.5	**36.1**	**39.9**	**38.0**	**36.0**
6	26.7	**28.1**	**34.0**	**32.0**	**30.5**

Task-specific attention masks



Conclusion & Discussion

- Consider all possible event mixtures as classes
- Multitask decomposition to circumvent the combinatorial explosion
- Multitask network architecture
- Future work
  - Optimal multitask decomposition?
  - Multitask architectures and training

Thank you for your attention