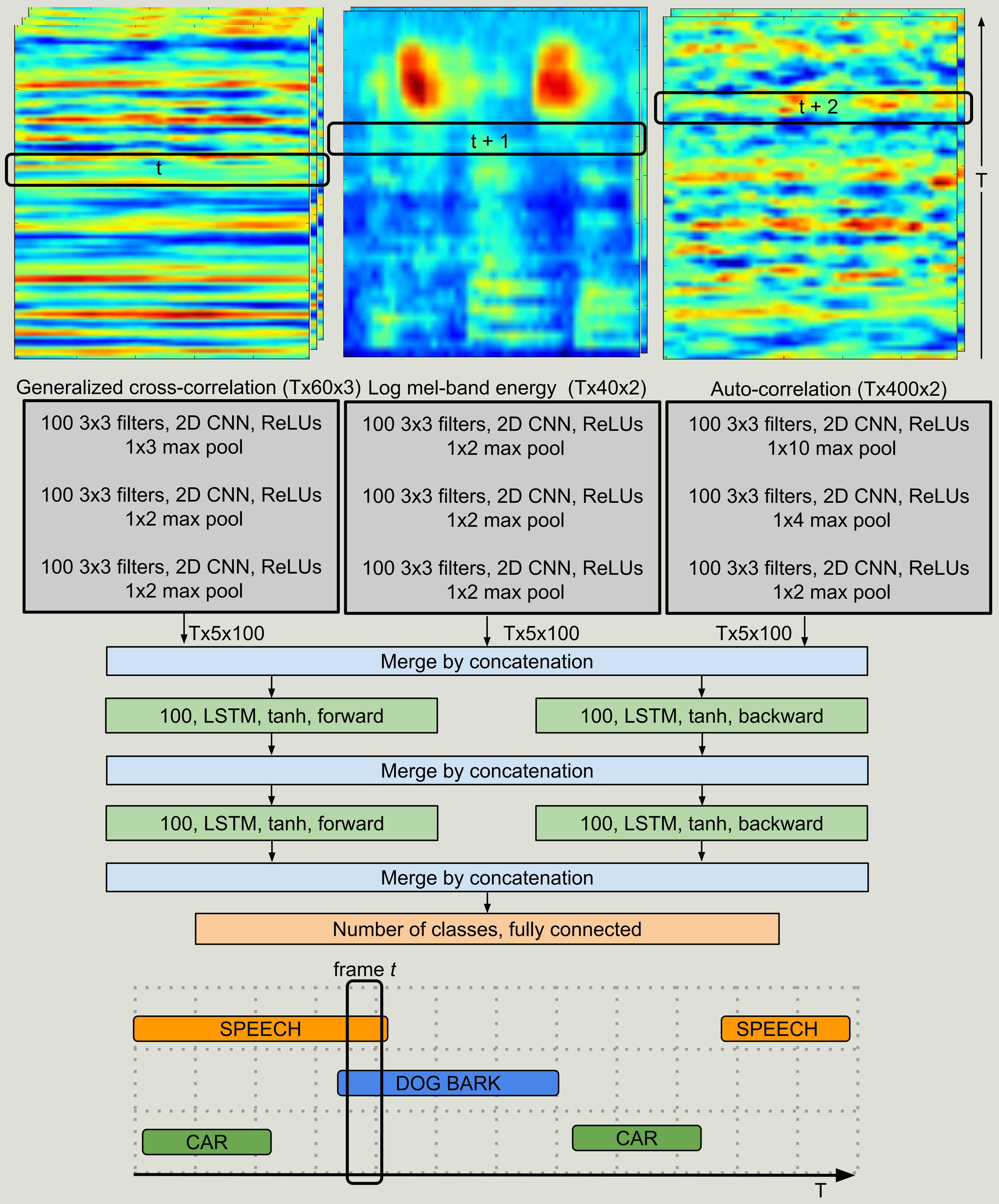
Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network

Sharath Adavanne, Pasi Pertilä, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology, Finland

Introduction

- Real life auditory scenes have many overlapping sound events, making it hard to recognize with just mono channel audio.
- We propose to train the SED systems to learn spatial information from binaural audio in order to distinguish overlapping sounds events better.



Convolutional bi-directional recurrent neural network (CBRNN) architecture for multichannel audio feature.

Spatial features

- Interaural intensity difference (IID)
  - Spatially separated sound events have different intensities in the binaural channels.
  - Represented using 40 log mel-band energies extracted from each of the binaural channels (*mel*).
- Interaural time difference (ITD)
  - Spatially separated sound events have different time difference of arrival (*TDOA*) values. Furthermore, temporally overlapping sound events do not always have the same frequency spread.
    - High level feature : *TDOA* - picked in five mel-bands.
    - Low level feature : Generalized cross-correlation with phase based weighting (*GCC-PHAT*) - single band.
- Perceptual feature
  - Overlapping sound events do not always have the same dominant frequencies.
    - *dom-freq* - Top three dominant frequencies and their magnitudes in 100-4000 Hz range.
    - *ACR* - auto-correlation magnitudes in 107.5-4410 Hz range.

Dataset

TUT-SED 2009

- Ten contexts - beach, office, restaurant, basketball, street etc.
- 9-16 classes and 8-14 recordings varying from 10-30 minutes for each context.
- Classes like cheering, applause, bird, laughter, music etc.
- Sum length of 19 hours.

TUT-SED 2016

- Development set of publicly available TUT-SED 2016 database.
- Two contexts - home (10 clips with 11 classes) and residential area (12 clips with 7 classes).
- Classes like cutlery, water tap running, wind blowing etc.
- Sum length of around an hour.

Both datasets consisted of audio recordings collected using in-ear microphones. All tests were done in context-independent manner.

Results

- Error rate (ER) and F-score achieved using binaural spatial features and CBRNN on TUT-SED 2009 and 2016 datasets.

Feature combination	TUT-SED 2009		TUT-SED 2016	
	ER	F	ER	F
CRNN baseline [Cakir 2017]	0.49	68.8	**0.93**	31.3
mel-monaural	0.49	68.0	1.03	29.7
mel-concat	0.44	70.3		
mel	**0.43**	71.1	0.99	32.3
mel + *TDOA*	0.45	70.9	0.95	**35.8**
mel + *GCC-PHAT*	0.44	71.1	0.95	34.6
mel + *dom-freq*	**0.43**	**71.7**	0.98	32.8
mel + *ACR*	0.44	71.2	0.98	33.8
mel + *TDOA* + *dom-freq*	0.44	71.0	1.01	33.3
mel + *GCC-PHAT* + *ACR*	0.45	70.9	0.99	33.6

- By using binaural over monaural features, F-score improved by 2.7% for TUT-SED 2009 and 6.1% for TUT-SED 2016.
- Comparable performance of using *GCC-PHAT* instead of *TDOA* or *ACR* instead of *dom-freq* shows that network learns equivalent high-level features information from just the low-level features.
- Other observations
  - *dom-freq* / *ACR* and *mel* useful for indoor and sound intense contexts (bus, hallway, office, and basketball)
  - *TDOA* / *GCC-PHAT* and *mel* are seen to help in outdoor contexts (beach and street).

Conclusions

- Binaural spatial features was shown to recognize sound events better than monaural features.
- Network architecture proposed to handle multiple feature classes and easily scalable to multichannels.
- Network was shown to learn high-level equivalent information from simple low-level features.

http://www.cs.tut.fi/~adavanne/ sharath.adavanne@tut.fi