A PLLR and Multi-stage Staircase Regression Framework for Speech-based Emotion Prediction

Zhaocheng(David) Huang[1,2], *Julien Epps*[1,2]

[1] The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia

[2] Data61, CSIRO, Australia

zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au

1. Introduction

- Phonetic variability is found to be detrimental in emotional speech processing, which is why phonetic features have been rarely used for speech based emotion recognition.
- Approaches to mitigate the variability involve:
  - functionals
  - lexical normalization
  - phone-specific features or models
- It is also found that some phones are emotionally discriminative.
  - e.g. features extracted from vowels are more helpful for emotion classification than those from consonants [11].
- Investigation in this study involves:
  - Direct use of phonetic features, i.e. the PLLR features, for speech-based emotion prediction
  - Exploitation of discriminative nature of phones using a Staircase Regression (SR) framework

2. Phone Log-likelihood Ratio (PLLR) Features

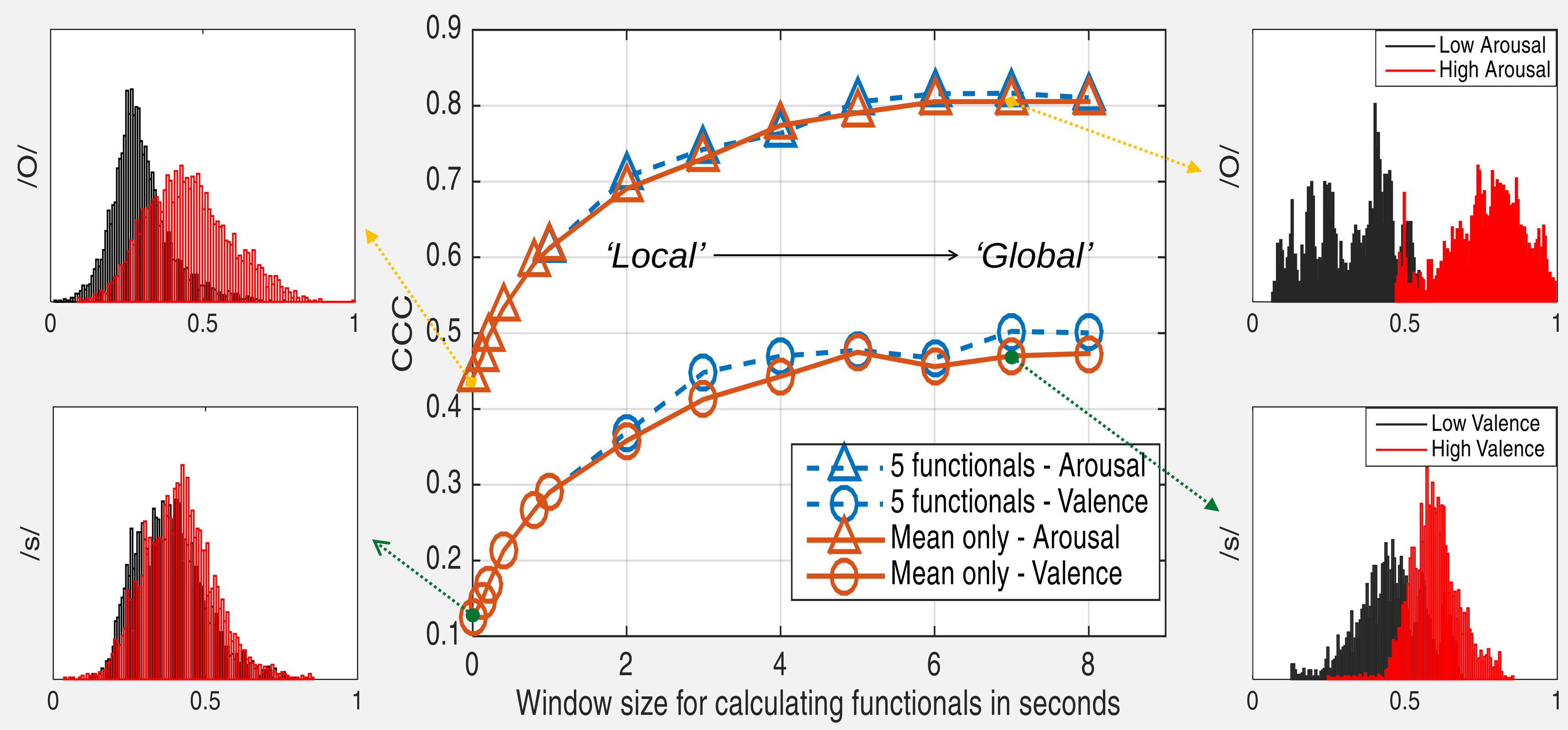
- Given a phone decoder with M phones, each of which has been modelled by one Hidden Markov Model (HMM) with S states, the posterior probability for each state s ($1 < s < S$) of each phoneme model m ($1 < m < M$) at each frame t is denoted as $p_{t,s}(m)$. Then the posterior probabilities of each phone are summed across all states before calculating the PLLR features [24]:

$$p_t(m) = \sum_{\forall s} p_{t,s}(m)$$

$$PLLR_t(m) = \log \frac{p_t(m)}{\frac{1}{(M-1)} \sum_{\forall j \neq m} p_t(j)}$$

- The ratio $PLLR_t(m)$ provides a probabilistic measure for the presence of phoneme m.
- In the emotion prediction context, PLLR features
  1. provide an indication of the most relevant phone for a given frame (allowing phone-specific modelling)
  2. provide a kind of 'positioning' of the current frame among all phones.

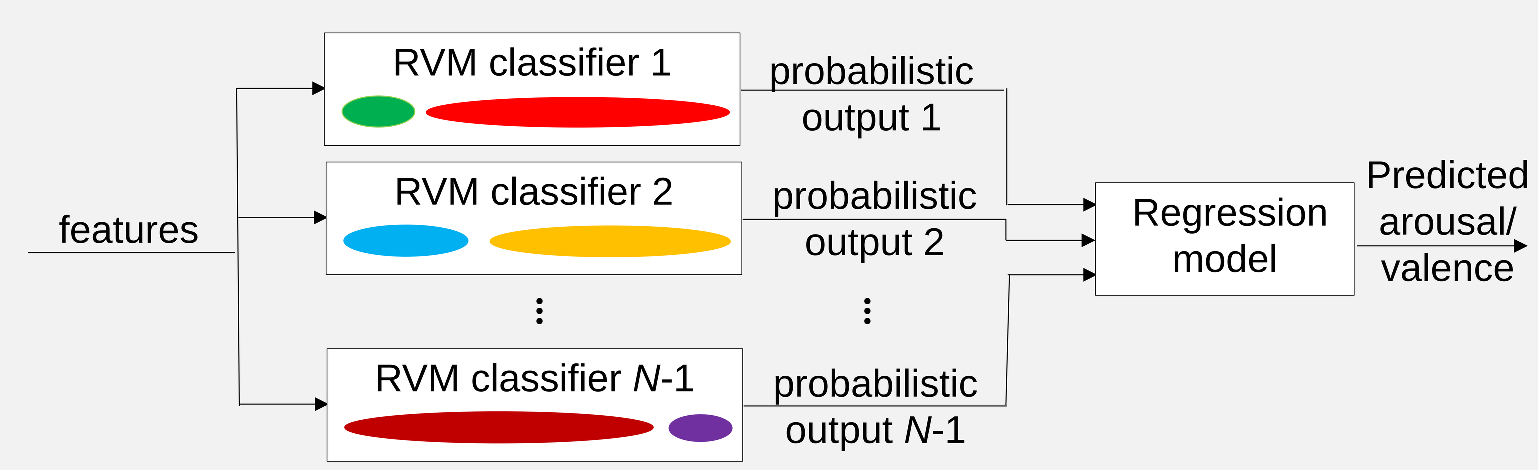
3. Global Functionals vs Local Functionals of PLLR Features

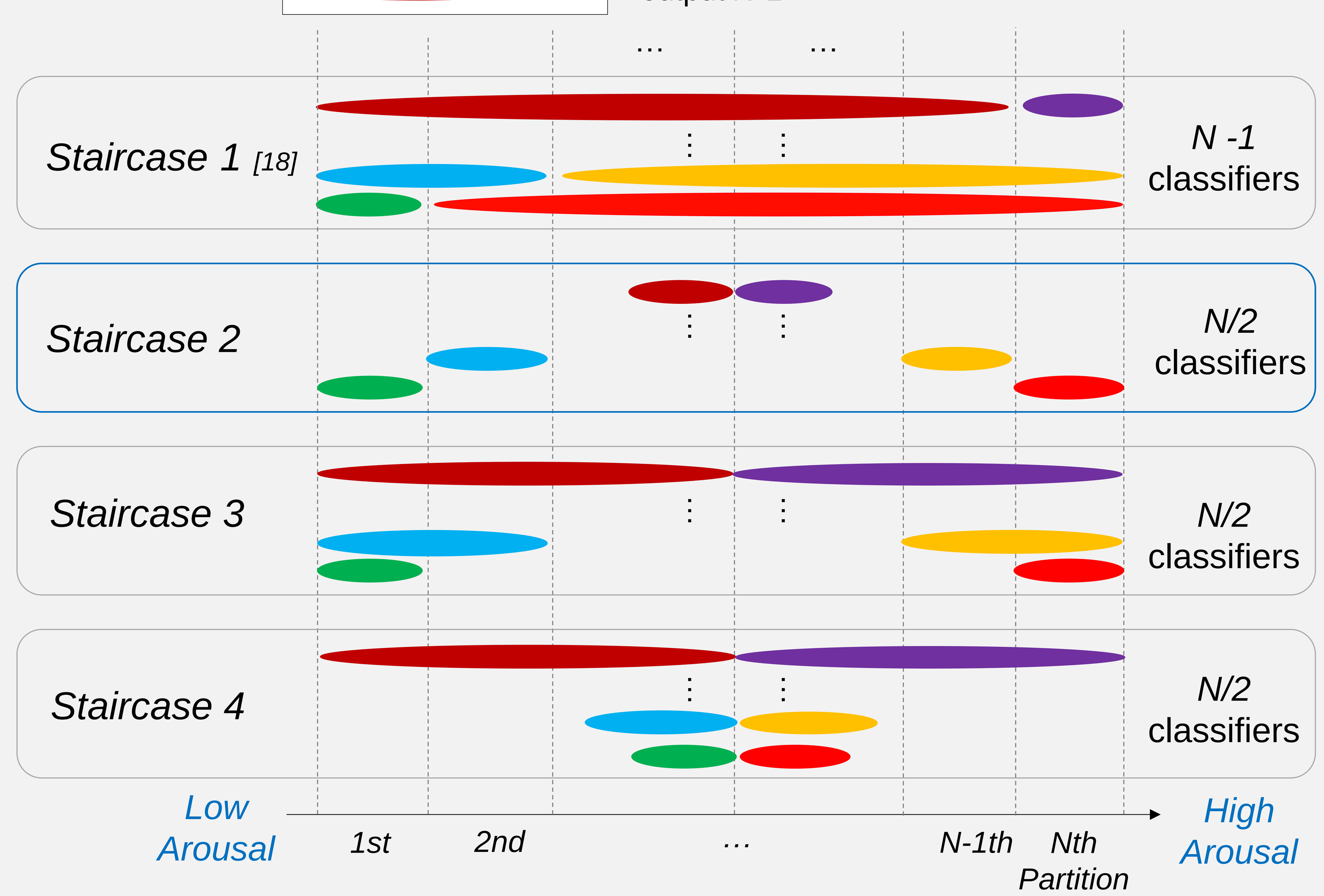


- Back-end regressor: Relevance Vector Machine (RVM)
- 5 functionals: mean, standard deviation, 20% percentile, 80% percentile, 20%-80% percentile
- /O/ phone for arousal and /s/ phone for valence
- PLLR features become more ***discriminative*** as window size increases, which is further exploited in the ***Staircase Regression (SR)*** framework

4. Staircase Regression Framework (RVM-SR)

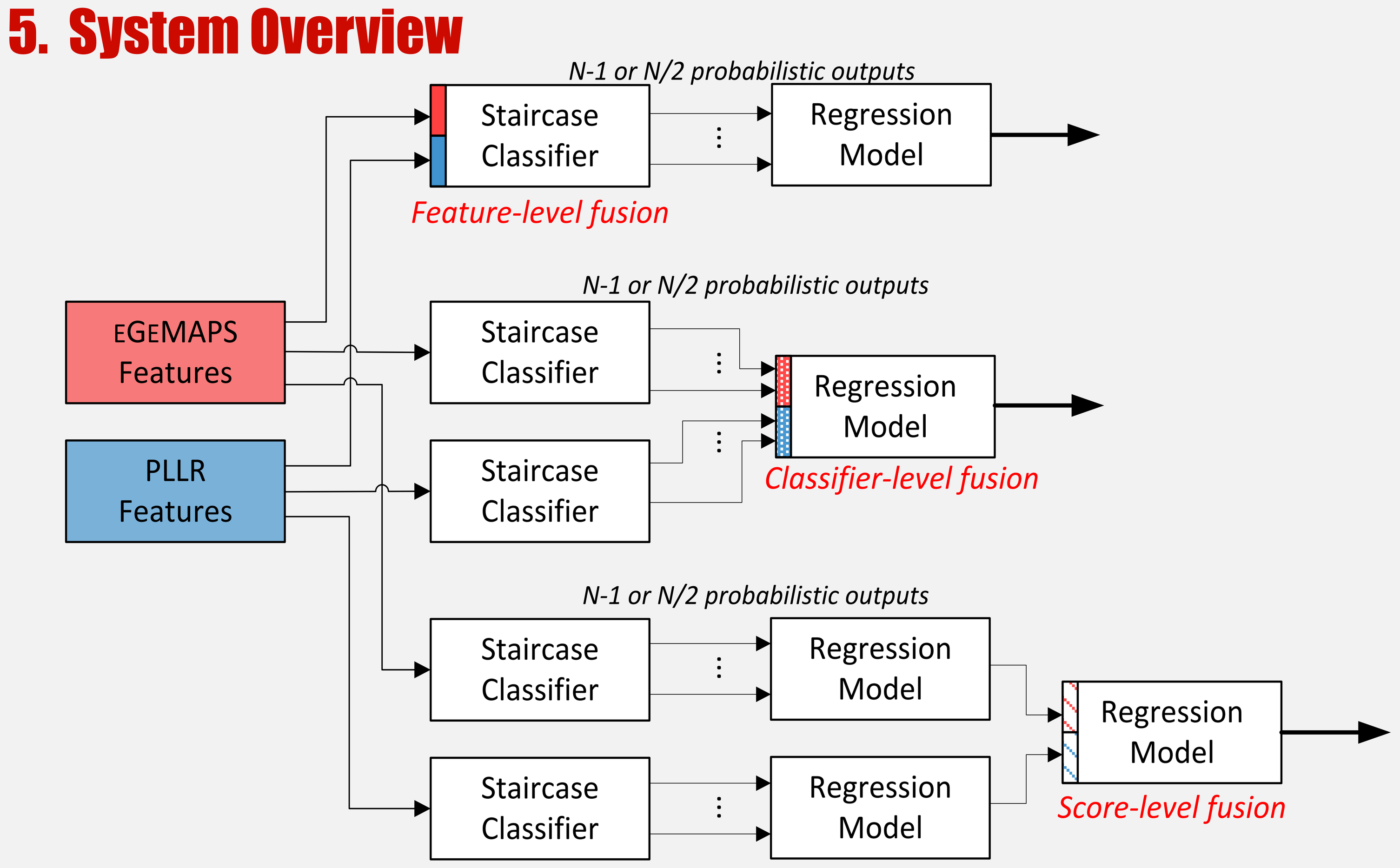
- Using scores of low-high classifiers for regression





- *Staircase 2* worked best; it contains more extreme low-high pairs.

5. System Overview



- The 88-dimensional EGEMAPS Features [6] are considered as a baseline feature set. The 59-dimensional PLLR were extracted using the BUT phone recognizer (Hungarian) [30].

6. Key Results

		Arousal	Valence
SVR	EGEMAPS ([5])	0.796	0.455
	PLLR	0.838	0.438
RVM	EGEMAPS	0.794	0.430
	PLLR	0.821	0.473
	feature-level fusion	0.848	0.502
RVM-SR (*Staircase 2*)	EGEMAPS	0.794	0.286
	PLLR	0.846	***0.508***
	feature-level fusion	0.860	0.463
	classifier-level fusion	0.849	0.437
	score-level fusion	***0.861***	0.500

- Performance measure is in Concordance Correlation Coefficient (CCC).
- Corpus: Audio/Visual Emotion Challenge (AVEC) 2016 dataset.
- Partition number in RVM-SR: 20 for arousal and 40 for valence.

7. Conclusion

- The pure phonetic PLLR features have been investigated, for the first time, for speech-based emotion prediction, outperforming the EGEMAPS feature set.
  - Global PLLR features showed significant improvements over the local PLLR features, which is presumably due to mitigation of frame-to-frame variability.
  - The promising results of PLLR features are fairly ***surprising***, opening new possibilities in this area.
- The Staircase Regression(SR) framework is helpful to exploit the discriminative nature of PLLR features.