Low resource automatic intonation classification using gated recurrent unit (GRU) networks pre-trained with synthesized pitch patterns

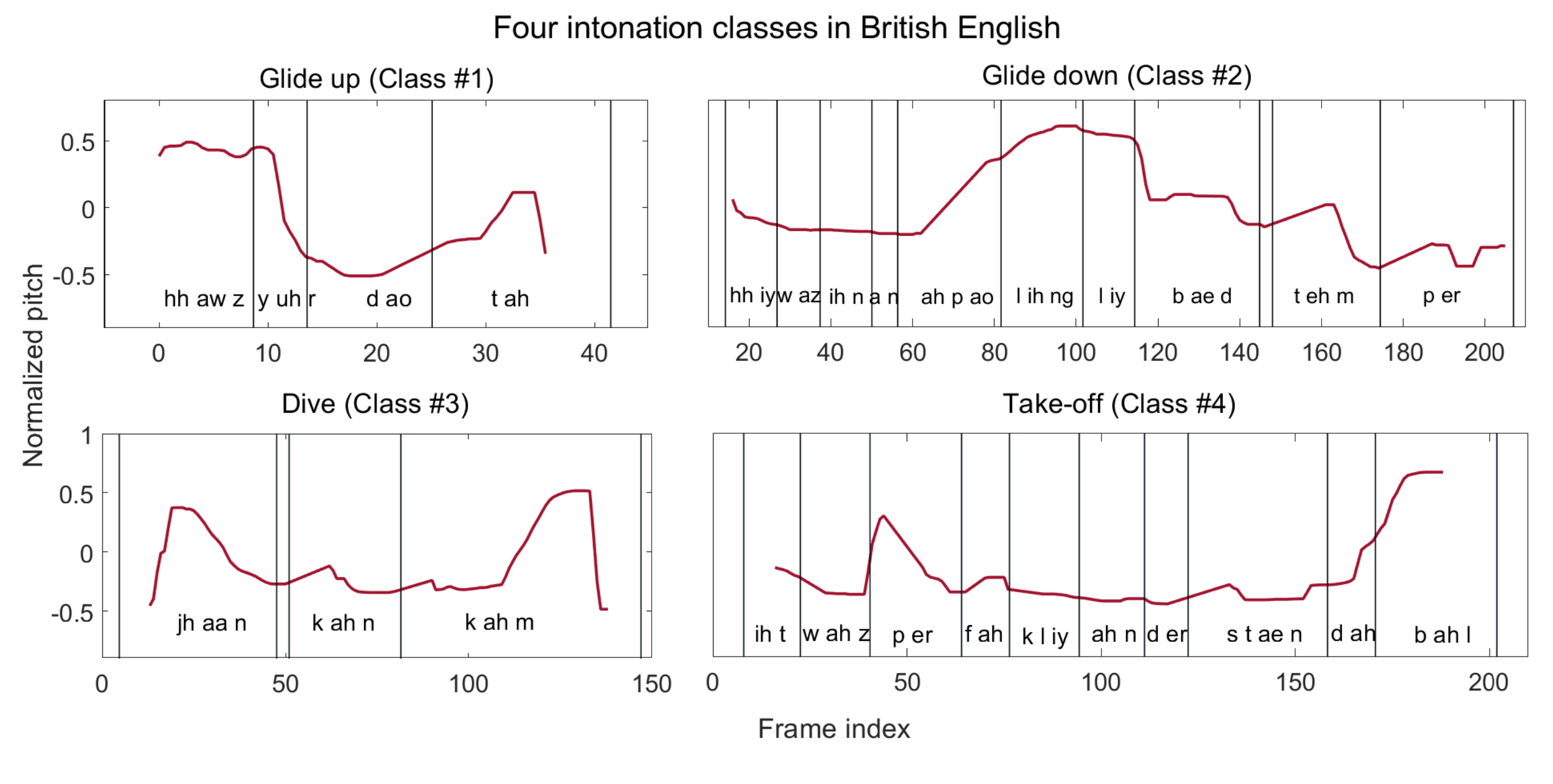
Atreyee Saha[1], *Chiranjeevi Yarra*[2], *Prasanta Kumar Ghosh*[2]

[1]Electrical Engineering, Jadavpur University, Kolkata 700032, India

[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India

Introduction

- Intonation refers to the modulation of pitch in the speech signal.
- It is modelled with temporal structures in either pitch contour or the tone sequence in an utterance.



- Why is intonation important?
  - It is believed to be an emotional indicative of the speaker.
  - For second language (L2) learners, intonation is important in conveying the meaning of an utterance [2].
- Objective:
  - To classify intonation in British English into one of the four classes under a low resource scenario.

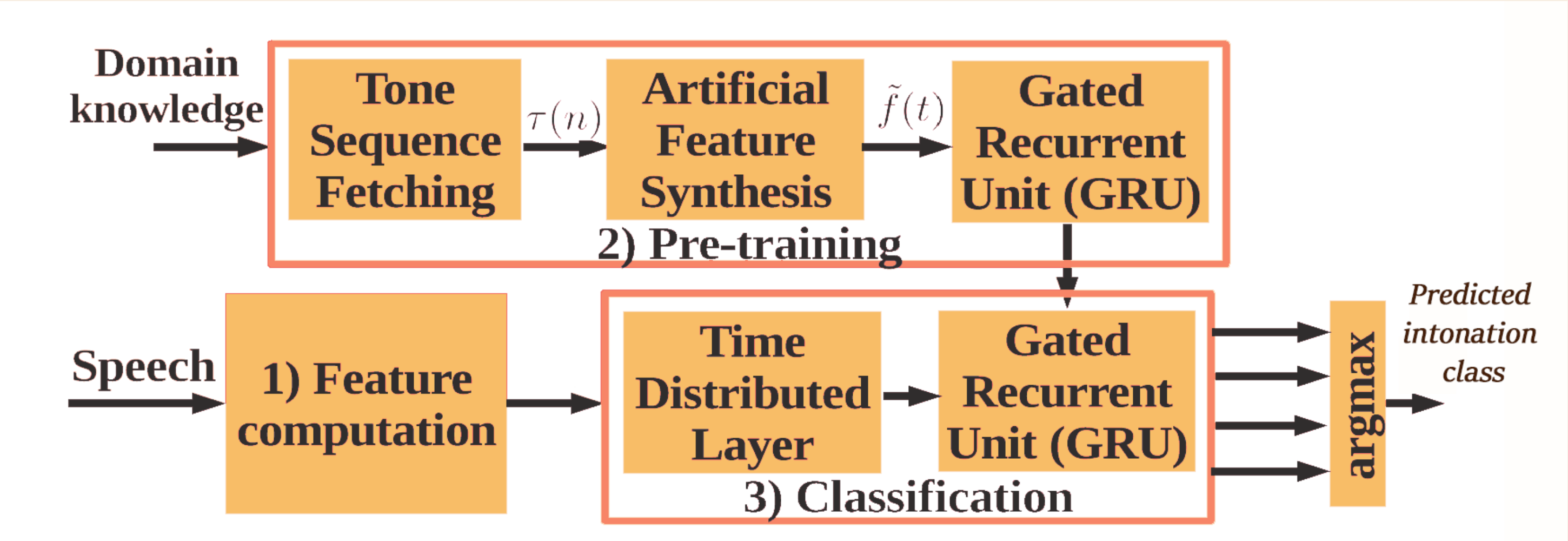
Key contributions

- Handling pitch estimation errors
  - Considered a 3-dimensional feature $f(t) = [f_1(t), f_2(t), f_3(t)]^T$, which includes the confidence score ($f_3(t)$) associated with estimated pitch values ($f_1(t)$). $f_2(t)$ is the first order difference of $f_1(t)$.
- Incorporation of tone sequence modeling
  - Estimation of tone sequence from the text of an utterance is costly and cumbersome and prone to errors.
  - Considered domain specific knowledge in modelling.
- Addressing low resource scenario
  - Proposed a domain specific knowledge based pre-training scheme.
  - Considered a gated recurrent unit (GRU) network, since it involves less number of trainable parameters.

Methodology

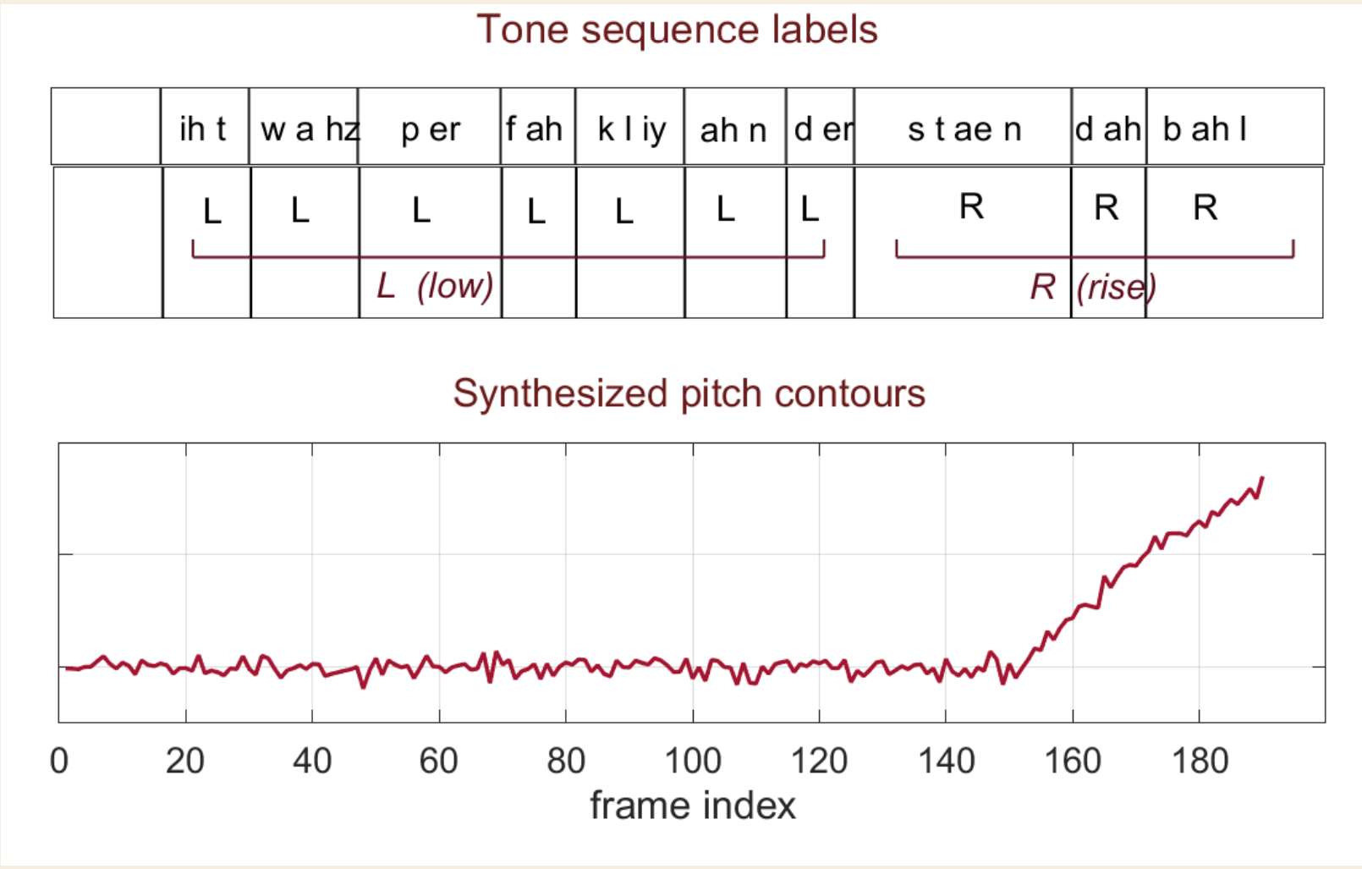
- Tone sequences $\tau(n)$ of each intonation class is collected based on the knowledge of relation between the temporal dependencies in tone sequences and the intonation classes.
- Pre-training is performed using synthetic pitch contours $\tilde{f}_1(t)$ that closely approximate actual pitch contours.
- During training, the weights of the time distributed layer (TDL) are learnt in a way so that the score values are used to minimise the errors in pitch estimation.

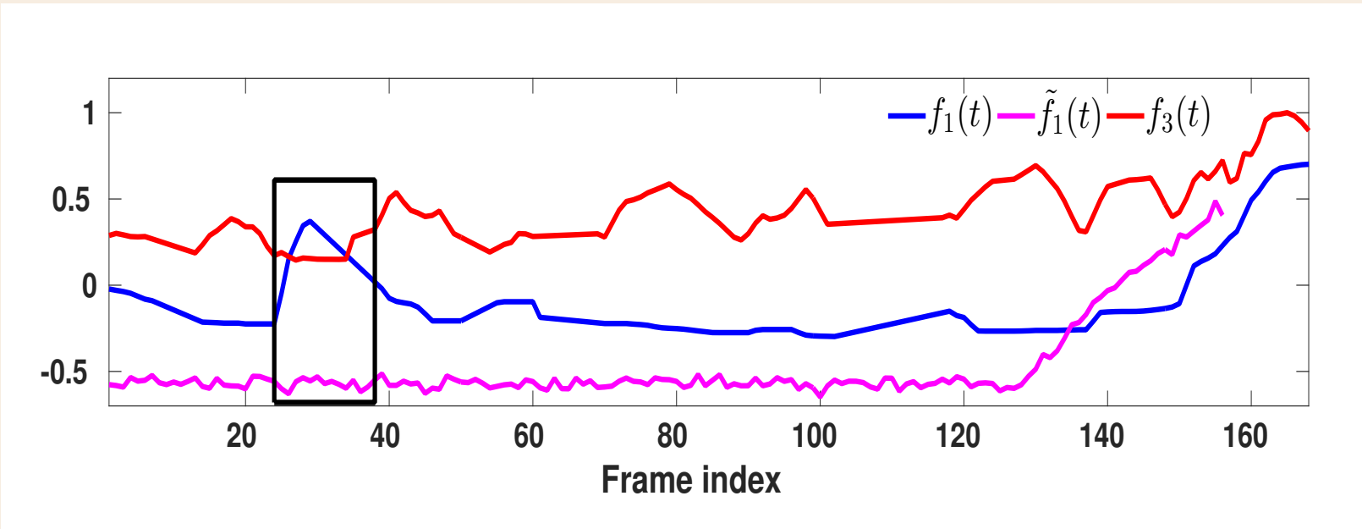
Methodology (contd.)



Proposed Approach

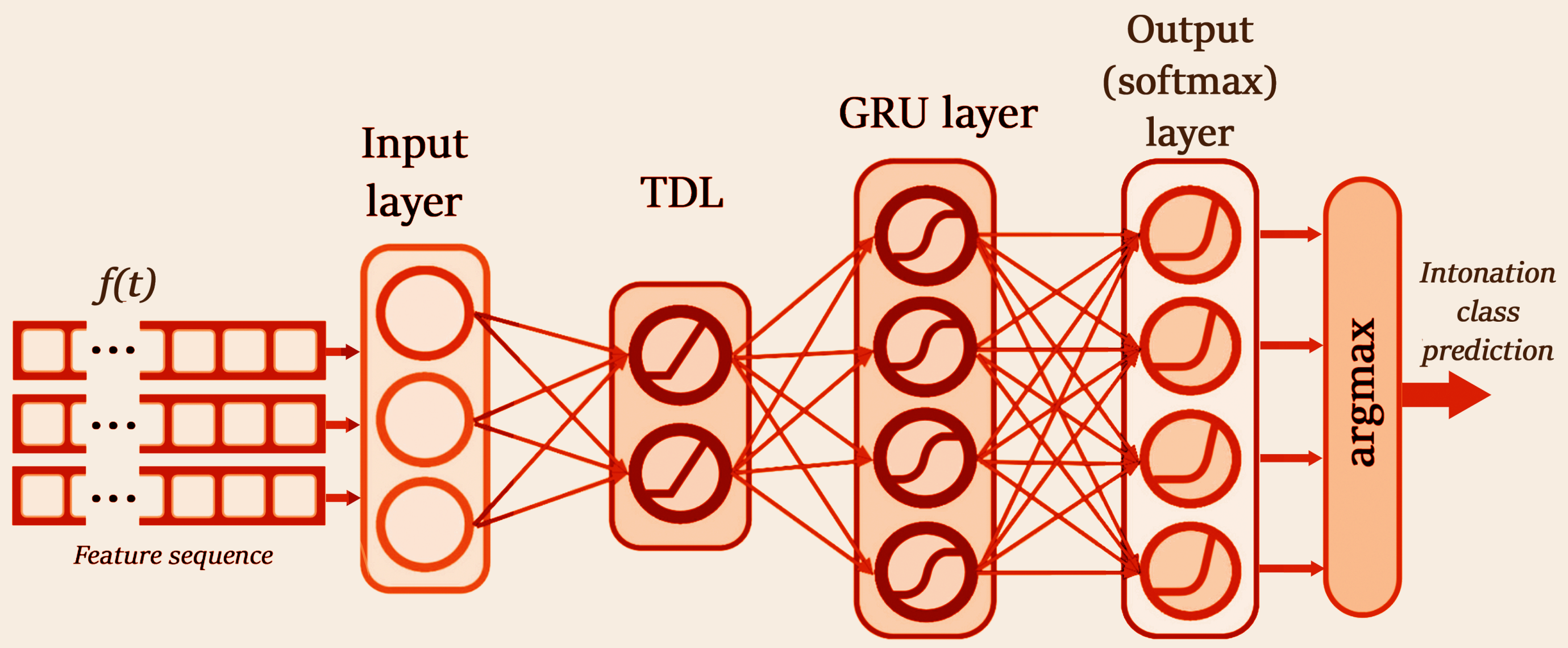
Feature synthesis





- $\tilde{f}_1(t)$ for an exemplary sentence "It was perfectly understandable" belonging to Take-off class.
- Consecutive syllable segments belonging to the same tone form a sub-segment.
- Interpolation is carried out using the individual sub-segments added with Gaussian noise at 20dB SNR.
- It is hypothesized that unwanted errors in pitch estimation are minimized using TDL before applying to GRU network.

Experimental setup



- Speech data is considered from spoken English training material [2] used for teaching British English.
- SWIPE algorithm used to estimate pitch and to obtain confidence scores [3].
- Baseline scheme and $f(t)$ are considered following the work of Yarra *et al.*[1].
- 10-fold cross validation setup.
- The unweighted average recall (UAR) as performance measure.

Results

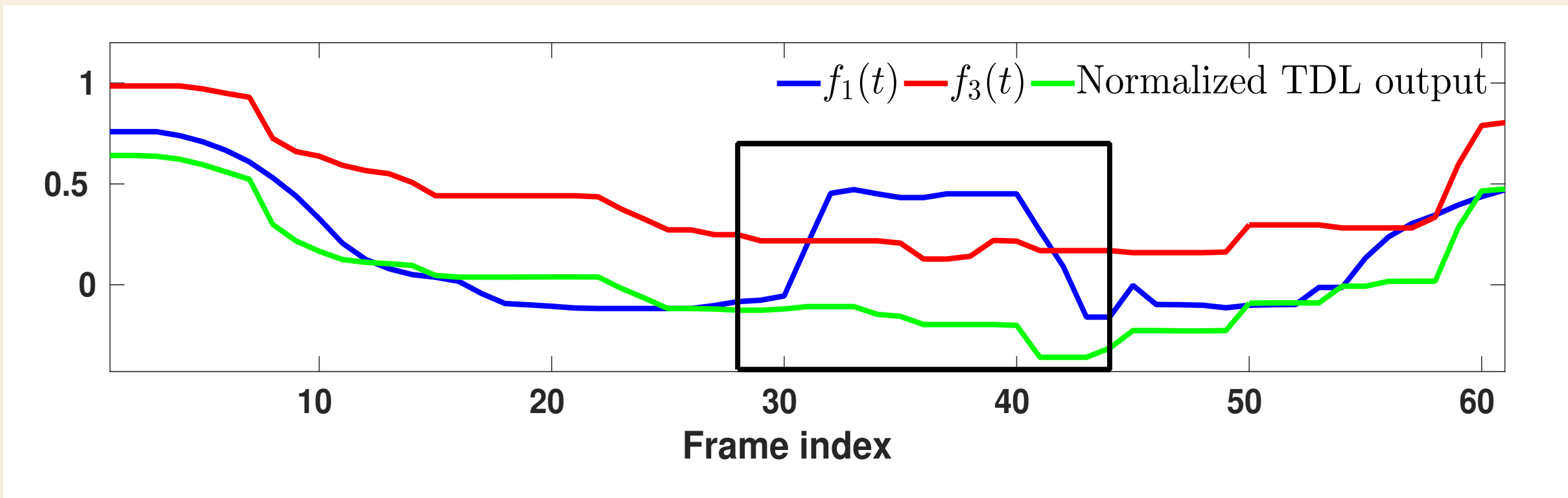
Performance in average UAR (SD)

	Baseline	*Proposed approach*			
		with pre-training		w/o pre-training	
		with TDL	w/o TDL	with TDL	w/o TDL
test	61.77 (8.6)	67.78 (9.8)	63.64 (6.9)	63.54 (5.4)	60.45 (7.3)
dev	62.32 (7.2)	67.73 (8.5)	62.67 (5.5)	63.59 (6.6)	60 (6.0)

Confusion matrix

Class	*Baseline*				*Proposed*			
	#1	#2	#3	#4	#1	#2	#3	#4
#1	62.5	5.00	30.0	2.5	15.0	10.0	70.0	5.0
#2	22.1	61.7	14.5	1.7	3.2	82.5	14.3	0.0
#3	22.2	17.2	54.6	6.0	2.4	14.6	78.1	4.9
#4	0.00	6.7	25.0	68.3	5.7	0.0	17.1	77.1

- The average UAR with the baseline is found to be 6.01% and 5.41% lower than that using the proposed approach on test and development sets respectively.
- Average UARs obtained with the proposed approach are higher when the classifier is pre-trained.
- Significant improvement (decrement) is found in the diagonal (off-diagonal) entries in the confusion matrix with the proposed approach compared to the baseline in all classes except Glide-up.



- Illustration of removal of unwanted pitch estimation errors at the output of TDL.

Conclusion

- Experiments on intonation classification are carried out on British English text implementing GRU network considering pre-training with synthetic pitch contour and input from TDL.
- Overall improvement in the accuracy compared to the baseline scheme.
- Further investigations are required to incorporate complementary properties of the proposed and baseline schemes for satisfactory discrimination between the Glide-up and Dive classes.
- Future works also include the use of linguistic features and data augmentation.

Acknowledgement

Authors thank the Department of Science and Technology, Government of India and the Pratiksha Trust for their support.

References

1. C. Yarra and P. K. Ghosh, "Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, pp. EL471–EL476, 2018.
2. J. D. O'Connor, *Better English Pronunciation.* Cambridge University Press, 1980.
3. A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

chiranjeeviy@iisc.ac.in, prasantg@iisc.ac.in