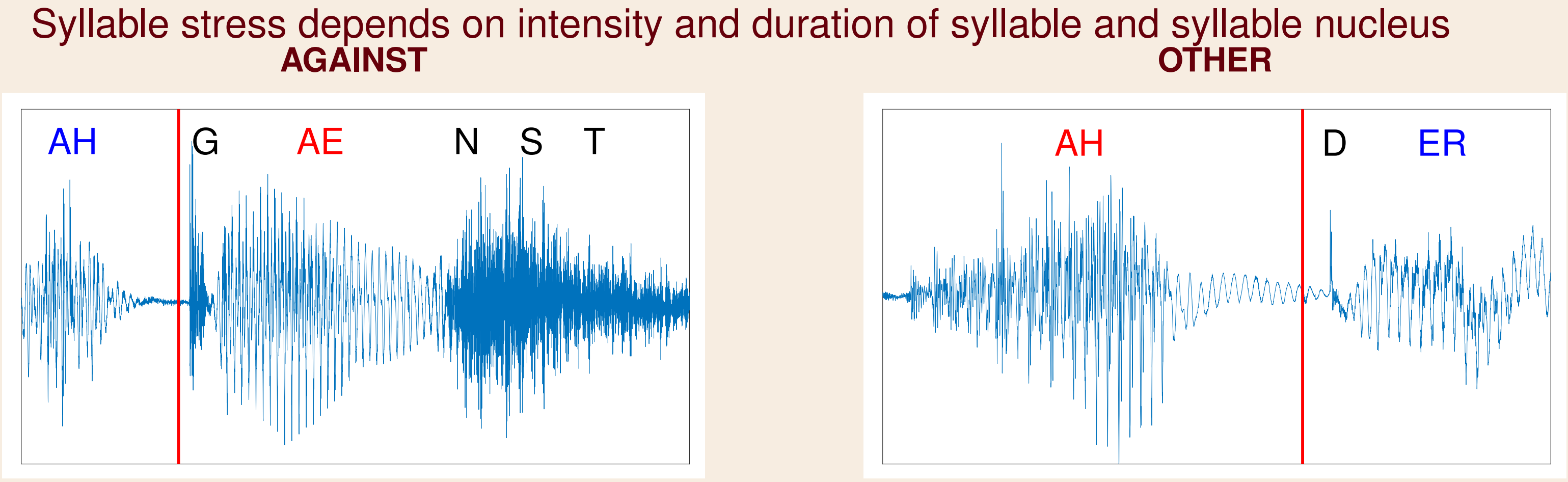
ASR inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features

Manoj Kumar Ramanathi, Chiranjeevi Yarra, Prasanta Kumar Ghosh

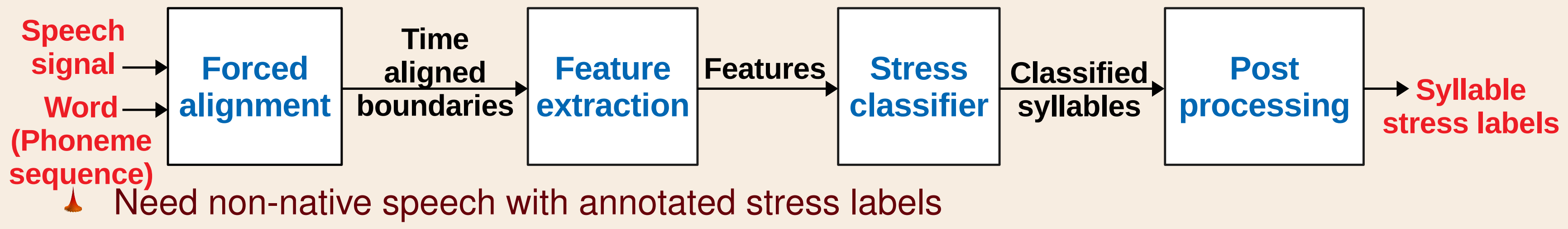
Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore

PROBLEM STATEMENT

- **Task:** To detect syllable stress in polysyllabic words spoken by non-native English speakers
- Syllable stress depends on intensity and duration of syllable and syllable nucleus

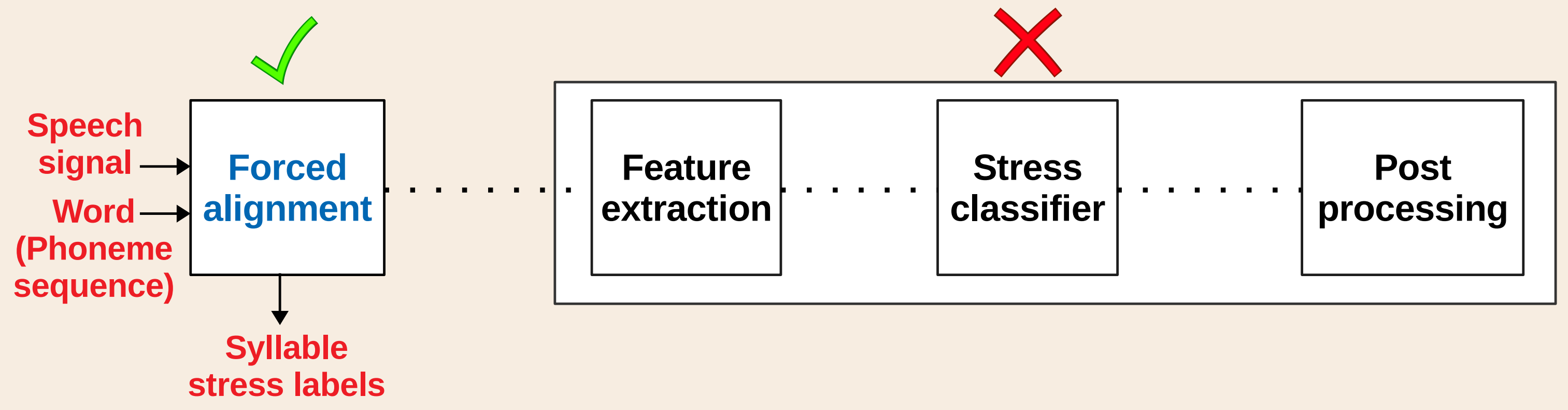


Typical approach:

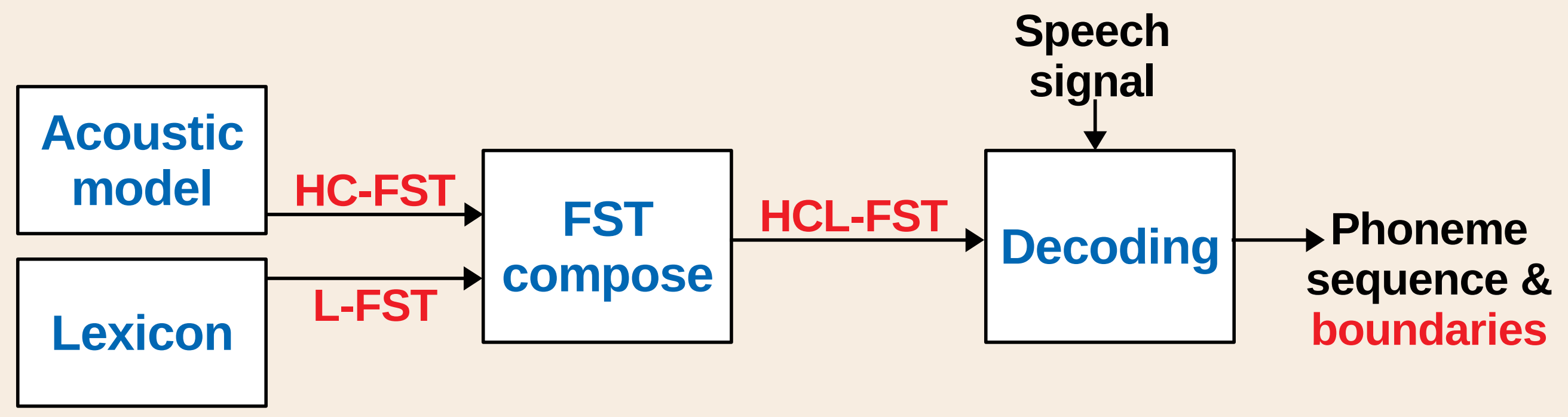


- Need non-native speech with annotated stress labels
- Inaccurate syllable/syllable nuclei boundaries due to typical forced alignment

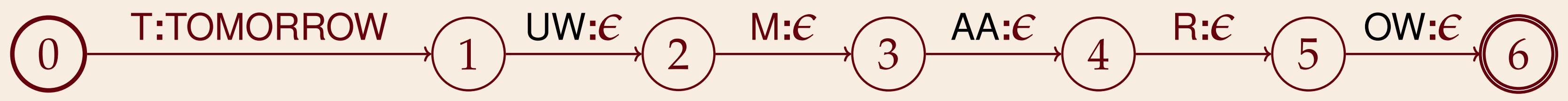
Our approach:



TYPICAL FORCED ALIGNMENT



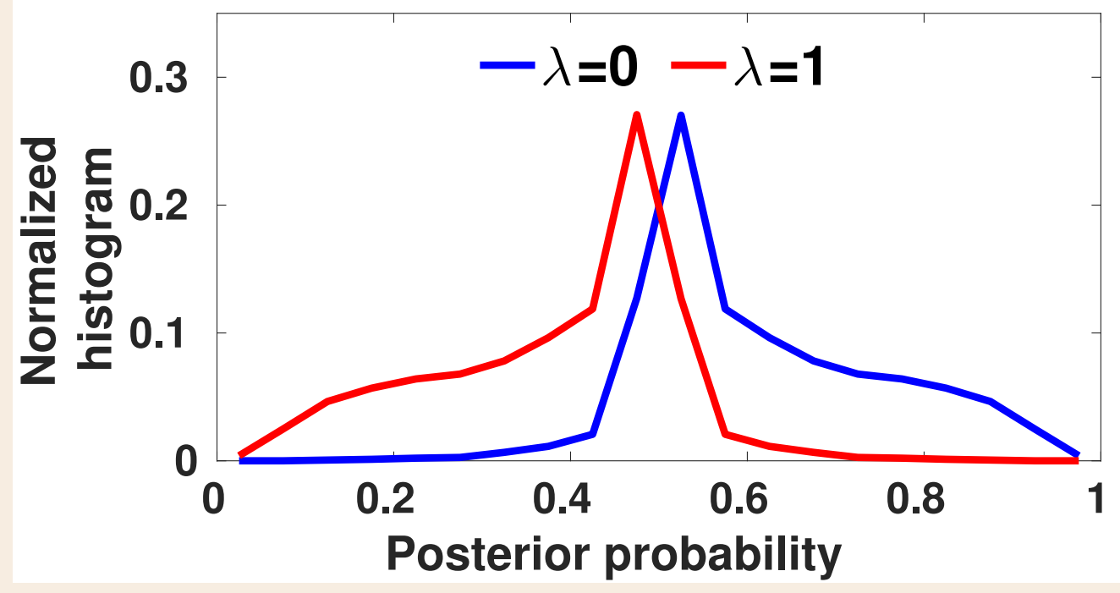
- Forced alignment is done through decoding using acoustic model and lexicon
- Acoustic model and lexicon are represented through finite state transducers (FSTs)
- Exemplary L-FST for the word **TOMORROW** (**T UW M AA R OW**)



MOTIVATION

- p - syllable nucleus phoneme (e.g., UW), λ - stress label (0,1), **O** - acoustic segment of p
- $\mathcal{P}(\lambda)$ - prior, $\mathcal{P}(\mathbf{O}|p,\lambda)$ - likelihood, $\mathcal{P}(\lambda|\mathbf{O},p)$ - posterior probability

$$\mathcal{P}(\lambda|\mathbf{O},p) = \frac{\mathcal{P}(\mathbf{O}|p,\lambda)\mathcal{P}(\lambda)}{\sum_{\lambda\in\{0,1\}} \mathcal{P}(\mathbf{O}|p,\lambda)\mathcal{P}(\lambda)}$$



Distribution of posterior probability of unstressed acoustic segments

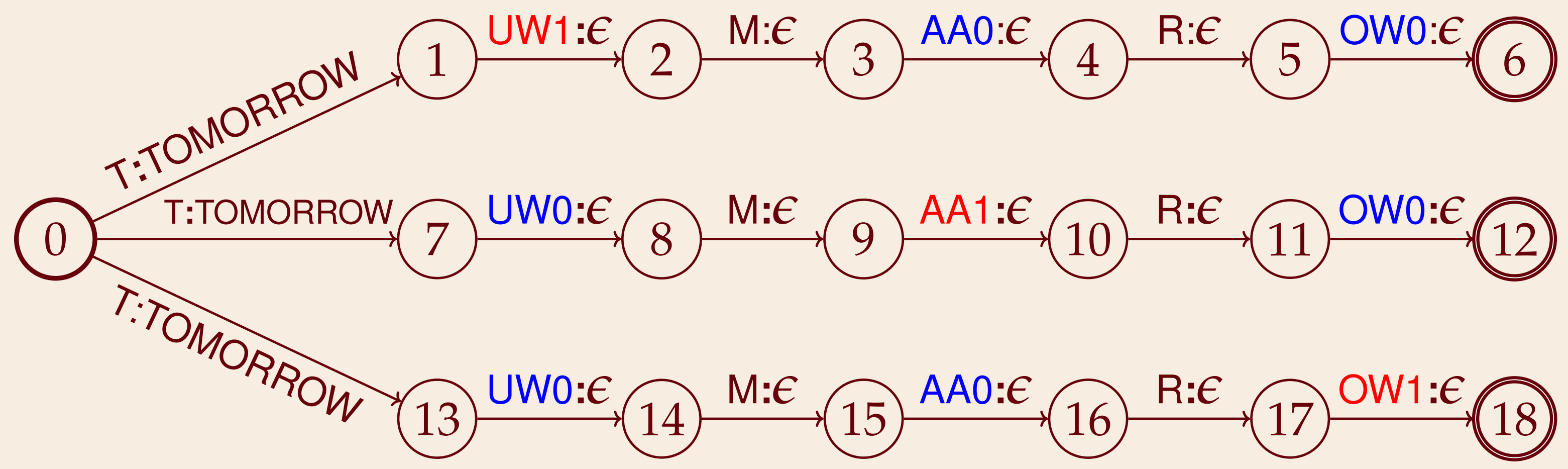
PROPOSED APPROACH

Modified forced alignment:

- Acoustic model with phoneme set containing stressed and unstressed vowel phonemes (syllable nuclei)
- Lexicon with **Stress Encoded Syllable Nuclei** (**SESN**): label 0 for no stress ⟹ UW0 for unstressed UW; label 1 for stress ⟹ UW1 for stressed UW

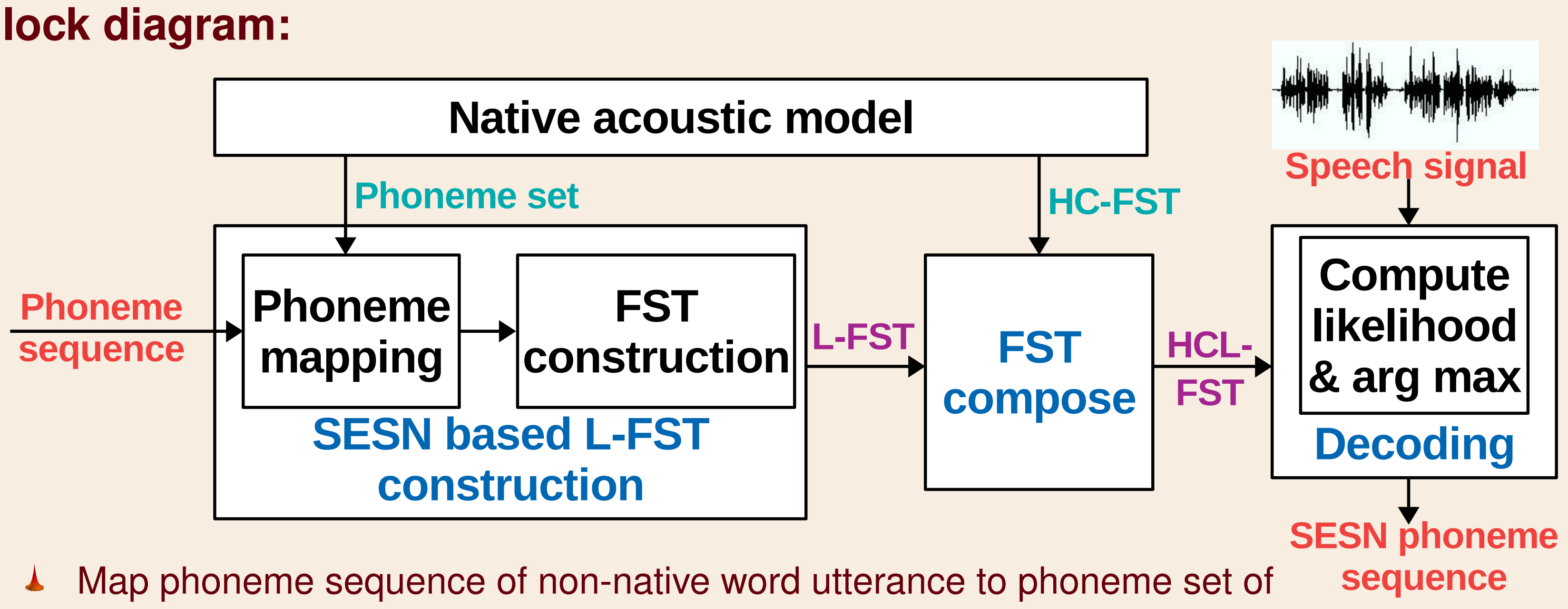
T UW1 M AA0 R OW0
T UW0 M AA1 R OW0
T UW0 M AA0 R OW1

SESN based L-FST:



- Phoneme sequence with *N* syllable nuclei has *N* stress variant pronunciations and hence *N* paths in SESN based L-FST

Block diagram:



- Map phoneme sequence of non-native word utterance to phoneme set of native acoustic model
- Construct SESN based L-FST
- Compose HC-FST of acoustic model and constructed L-FST
- Decode the maximum likely SESN based phoneme sequence

EXPERIMENTAL SETUP

Native acoustic model data:

- 960 hours of LibriSpeech (**Libri**)
- 30 hours of LibriSpeech (**Libri-S**)
- 30 hours of Wall Street Journal (**WSJ**)

Non-native stress detection data:

- **ISLE** corpus[1] - English utterances by 10 Italians (**ITA**) and 11 Germans (**GER**)
- Manually annotated stress labels for polysyllabic words
- Distribution of bisyllabic (**B**), trisyllabic (**T**) and quadrisyllabic (**Q**) words which form overall (**O**) words across **ITA** and **GER**:

	B	T	Q	O
ITA	1873	433	82	2388
GER	2360	578	99	3037

- **Metric**: syllable stress detection accuracy

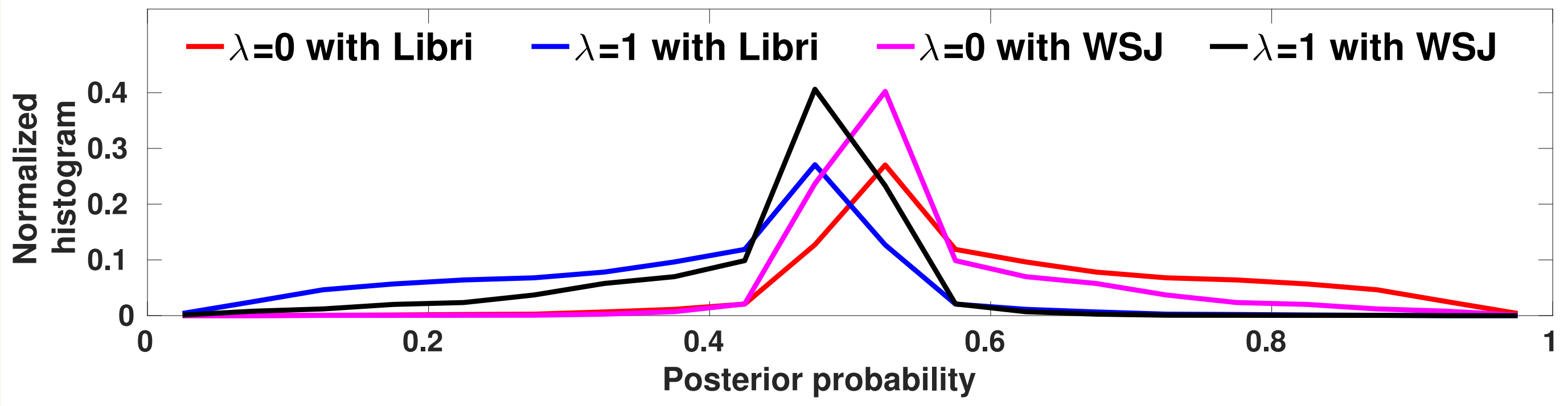
RESULTS & DISCUSSION

Comparison of proposed approach (PA) with baselines:

- Supervised baseline approaches - **BL-1**[2], **BL-2**[3]

	BL-1	BL-2	PA		
			Libri	Libri-S	WSJ
ITA	83.17	86.26	**85.24**	75.39	72.05
GER	85.81	87.53	**87.00**	79.32	75.33

- Although unsupervised, proposed approach performs on par with supervised baselines
- Sensitive to the amount of training data of acoustic models
- Number of syllable nuclei with posterior probability < 0.5 for $\lambda = 0$ is higher for WSJ than Libri



Performance across word lengths:

		ITA			GER		
		B	T	Q	B	T	Q
BL-2		88.85	86.97	77.21	89.13	84.31	73.58
PA	Libri	86.92	83.37	76.83	87.02	**89.27**	**74.24**
	Libri-S	76.13	72.13	**79.88**	78.22	81.43	**83.33**
	WSJ	71.70	72.90	72.56	75.04	76.93	**75.25**

- Performs better than BL-2 on words with more than two syllables

CONCLUSION

- Syllable stress detection in ASR framework performs on par with supervised baselines
- Train native acoustic model with phoneme set containing both stressed and unstressed syllable nuclei and construct lexicon with multiple phoneme sequences containing SESN

Future work:

- Methods for stress detection when the input phoneme sequence is unavailable
- Analysis of first language specific tendencies for mis-placing syllable stress

ACKNOWLEDGEMENT

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

REFERENCES

1. W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English", Proceedings of Language Resources and Evaluation Conference (LREC), vol. 2, pp. 957-964, 2000
2. J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 937-940, 2005
3. C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5845-5849, 2017

http://spire.ee.iisc.ac.in/spire/

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