Comparison of automatic syllable stress detection quality with time-aligned boundaries and context dependencies

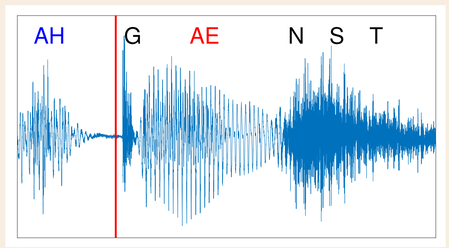
Chiranjeevi Yarra, Manoj Kumar Ramanathi, Prasanta Kumar Ghosh

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

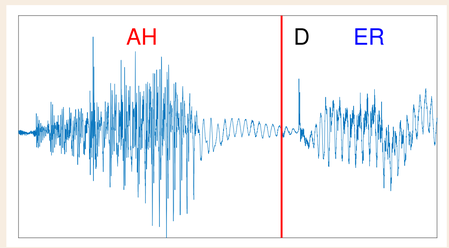
INTRODUCTION

- Syllable stress depends on intensity and duration of syllable and syllable nucleus

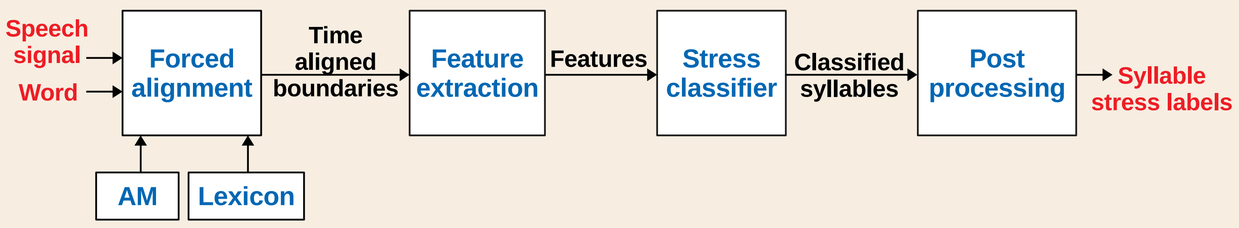
AGAINST



OTHER



- Syllable stress detection is carried out with classifier trained with stress labels and acoustic features within syllables
- Syllable boundaries are estimated through forced-alignment using acoustic model (AM) and lexicon



MOTIVATION

- Acoustic features are not only affected by stress prominence but also by context of the syllable nucleus[1]. Propose to use 19 dim. binary context feature vector

Context	Category
Syllable nucleus type	Low, Mid, High, Lax and Tense
Phoneme preceding the syllable nucleus	Nasals, Voiced stops, Unvoiced stops, Voiced fricative, Unvoiced fricatives and remaining consonants
Phoneme following the syllable nucleus	Nasals, Voiced stops, Unvoiced stops, Voiced fricative, Unvoiced fricatives and remaining consonants
Word Position	Pre-pausal, Not pre-pausal

- Syllable data from forced alignment require annotation of labels
  - costly and cumbersome
  - **Mismatch train-test condition**: Train with ground-truth syllable data and test with estimated syllable data

STUDY SUMMARY

- Train always with ground-truth syllable data and stress labels
- Test data for experiments is obtained as follows:

	Syllable data		Comments
	Transcription	Boundaries	
Exp-1	ground-truth	estimated	Used different AMs Analyzed the effect of AMs
Exp-2	estimated	estimated	Used native lexicon Analyzed mismatched train-test condition
Exp-3	estimated	estimated	Used native lexicon augmented with different percentages of non-native pronunciations Analyzed the effect of lexicons

- Both acoustic features alone and acoustic features augmented with the proposed context features are used in all the experiments

EXPERIMENTAL SETUP

Non-native stress detection data:

- ISLE corpus[2] - English utterances by 23 Germans (GER) and 23 Italians (ITA)
- Manually annotated stress labels for polysyllabic words
- 7586 & 7791 and 8586 & 4648 words in train and test data respectively for GER & ITA

Native AM data: AMs trained on LibriSpeech (LS), Wall Street Journal (WSJ) and Fisher English (FE) corpora

Lexicon:

- **Native lexicon:** by combining CMU, TIMIT, Beep and the lexicon used in preparing ISLE corpus
- **Lexicons:** by augmenting native lexicon with 0, 25, 50, 75, and 100 percentage of word pronunciations by GER and ITA

Acoustic features:[3]

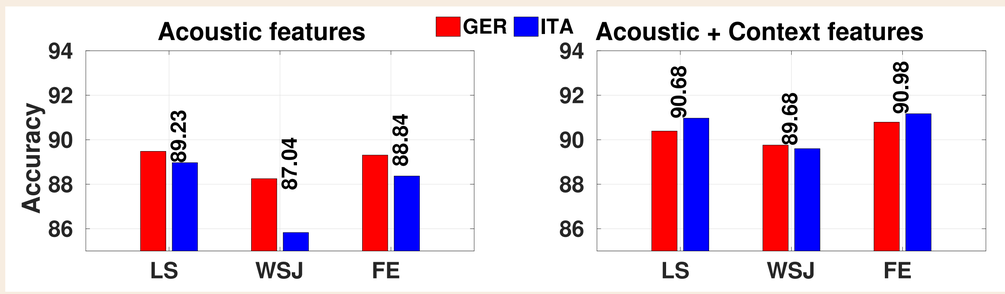
- 10 dim. syllable level features and 10 dim. syllable nuclei level features
- Features are based on strength, temporal variability and duration

Classifier: SVM classifier with RBF kernel and complexity parameter as 1.0

Metric: Syllable stress detection accuracy

RESULTS & DISCUSSION

Exp-1: Comparison based on AM:



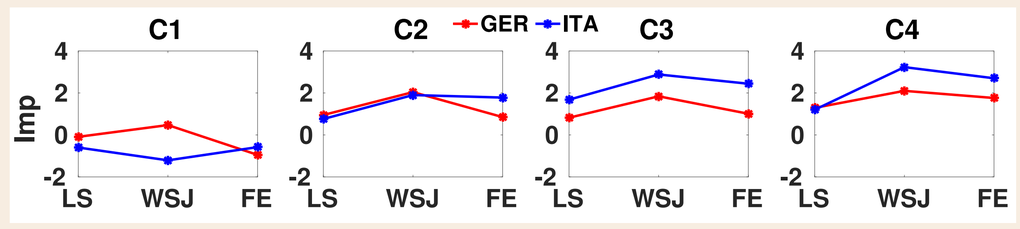
- Acoustic + context features perform better than acoustic features alone

Exp-2: Comparison based on mismatched train-test condition:

- Four combinations (C1, C2, C3, C4) of features are obtained from different types of syllable data:

Feature type	C1	C2	C3	C4
Acoustic	estimated	estimated	estimated	ground-truth
Context	none	estimated	ground-truth	estimated

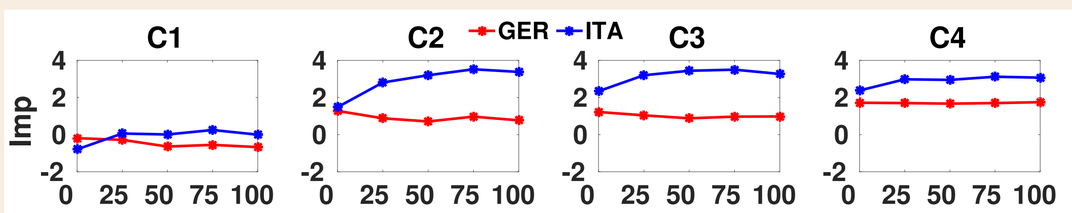
- Improvements (Imp) over accuracy obtained with acoustic features using ground-truth syllable data, which is 90.88%.



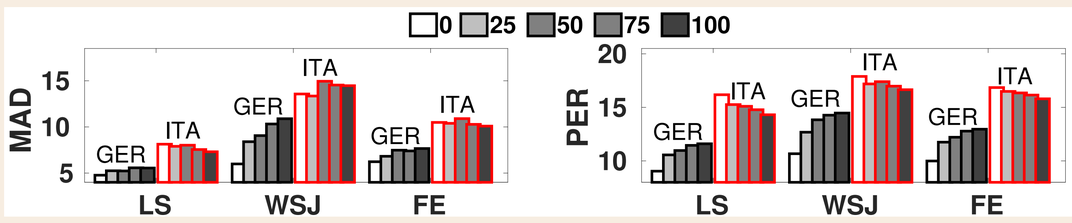
- The accuracy with acoustic + context based features are higher even those are obtained from estimated data.
- Comparing C4 and C3: Estimation of more accurate boundaries is important than the estimation of correct syllable transcriptions.

Exp-3: Comparison based on lexicon:

- Imp across percentage of word pronunciations averaged across LS, WSJ, FE



- Mean absolute difference (MAD) between the boundaries and phoneme error rate (PER) between the transcriptions of the estimated and the ground-truth syllable data is obtained as:



- Variations in averaged improvements are inversely proportional to variations in the MAD and the PER
- Increasing the size of lexicon improve the accuracy for a few types of non-native speakers

CONCLUSION

- Analyzed the accuracy of a supervised stress detection method as a function of the quality of the estimated syllable data.
- Proposed context features which are augmented with acoustic features
- Experiments are conducted on ISLE corpus using five types of lexicons, where the lexicons are generated by adding five different percentages of non-native pronunciations from the corpus.
- This study suggests the requirement for development of better strategies to obtain accurate aligned boundaries than accurate transcriptions.

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

REFERENCES

1. N. Umeda, "Vowel duration in American English", The Journal of the Acoustical Society of America, vol. 58, no. 2, pp. 434-445, 1975
2. 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
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

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