AUTOMATIC DETECTION OF SYLLABLE STRESS USING SONORITY BASED PROMINENCE FEATURES FOR PRONUNCIATION EVALUATION

Chiranjeevi Yarra[1], Om D. Deshmukh[2], Prasanta Kumar Ghosh[1]
[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India; [2]Xerox Research Center India, Bangalore, India. chiranjeevi/.yarra@ee.iisc.ernet.in, om.Deshmukh@xerox.com, prasantg@ee.iisc.ernet.in

Problem definition

TELEVISION

tɛ.lə.vi.ʒən

Sonority levels

Features are proposed by incorporating **relative sonority** levels in the **prominence** measures **for syllable stress detection.**

Compared to unstressed syllables, stressed syllables are perceptually more **prominent**

1. *Intensity*
2. *Duration*
3. *Pitch*

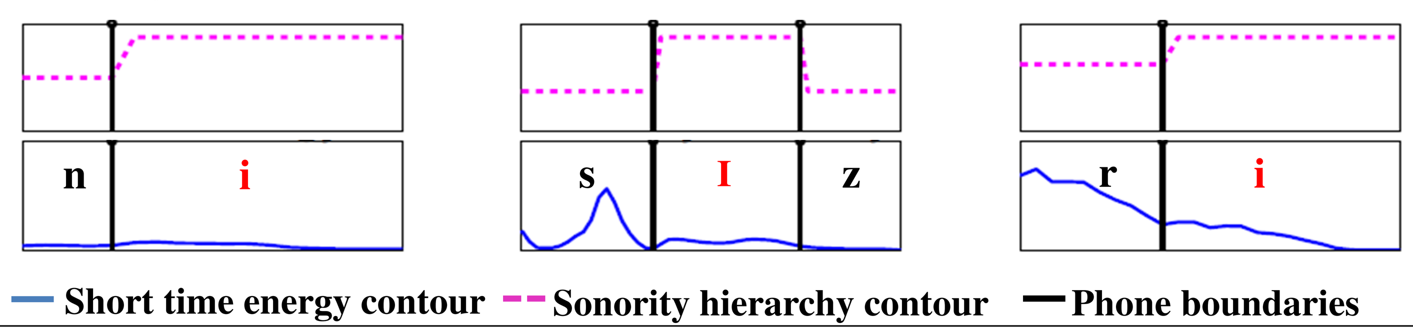
Features → Classifier

How sonority is useful?

Sonority is referred to as the carrying power of individual sounds in a word or a longer utterance. The carrying power is measured based on the sonorous hierarchy of various classes of sounds [1].

Below figure shows that the short-time energy contours have larger variabilities, but not the proposed hypothetical sonority contour.

Combined sonority cues and short-time energy could discriminate stressed and unstresed syllables better.



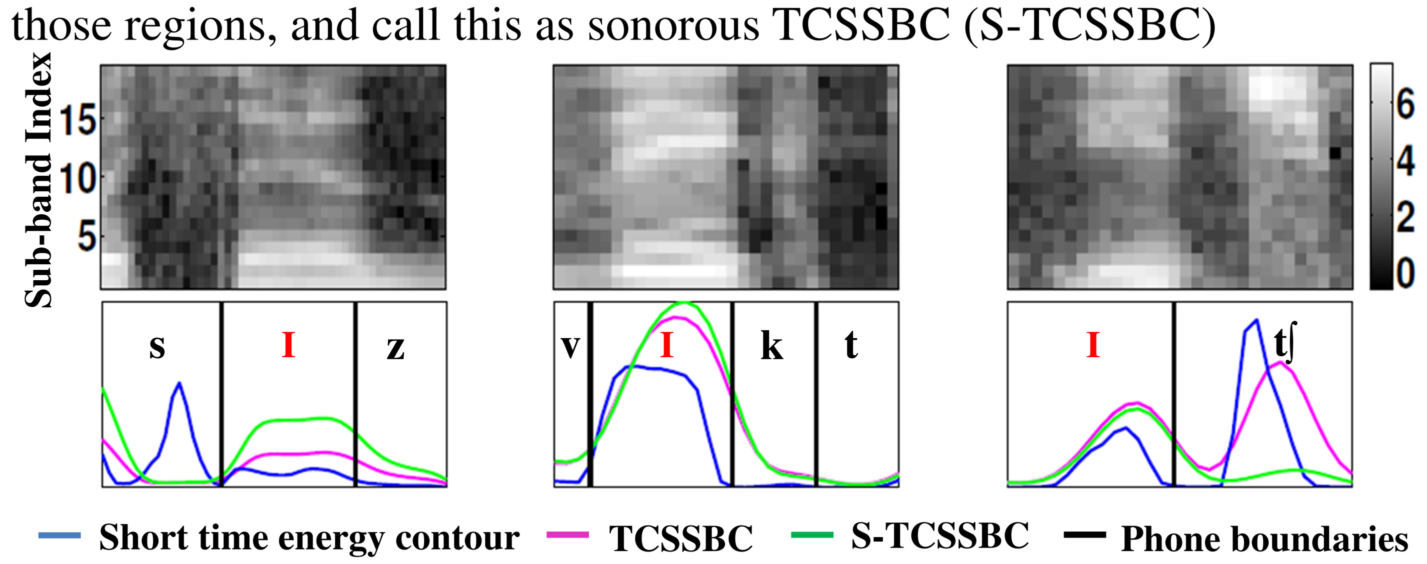
Works on measuring sonority

We assume that the sonority is related to the consistent temporal pattern in sub-band energies captured by spectro-temporal correlation (STC) [2], which has been used to compute Temporal Correlation and Selected Sub-Band Correlation (TCSSBC) [3] contour.

STC has been shown to be effective in exploiting the formant like structures in the spectral domain with the help of short-time energy contours of 19 sub-bands.

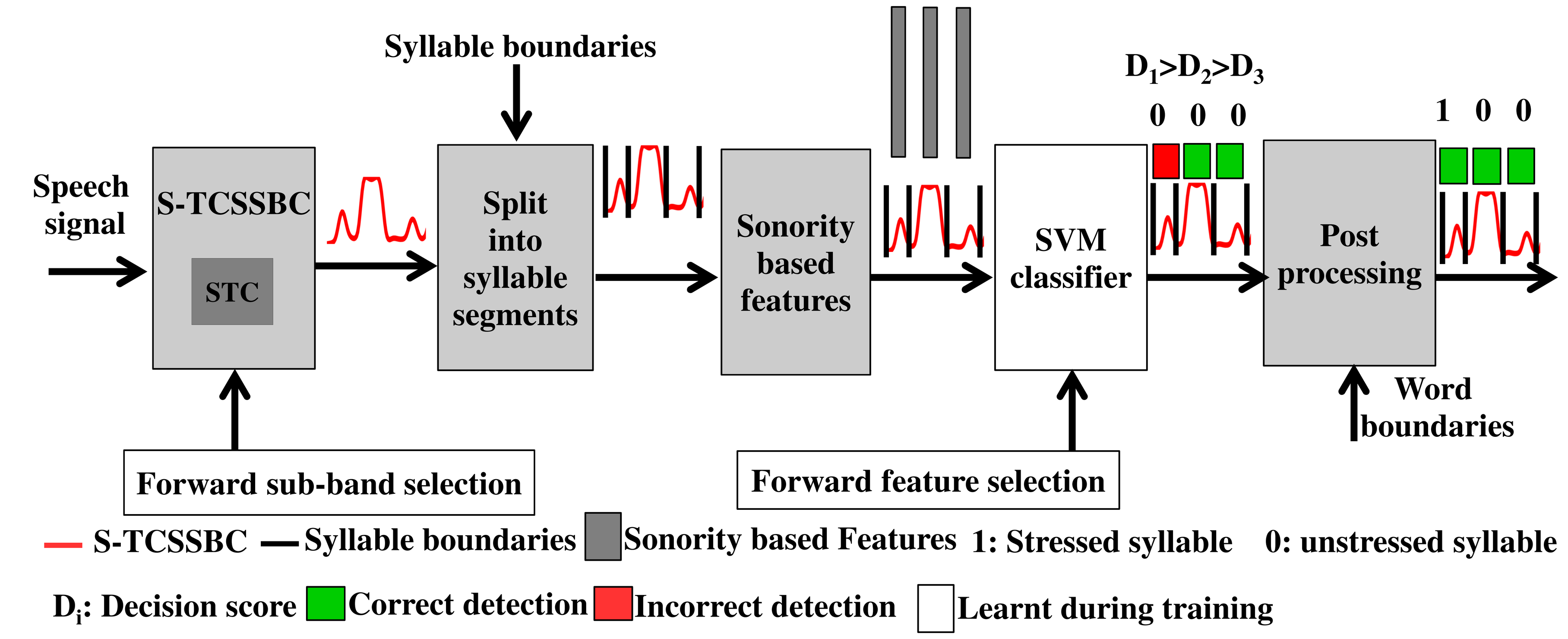
However, TCSSBC introduces peaks in the less sonorous regions (as shown below) due to using all 19 sub-bands.

We modify TCSSBC by selecting a few sub-bands to reduce its peaky nature in those regions, and call this as sonorous TCSSBC (S-TCSSBC)



Proposed approach

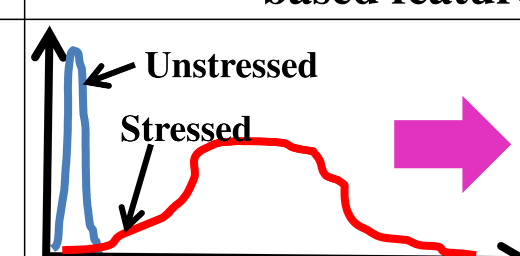
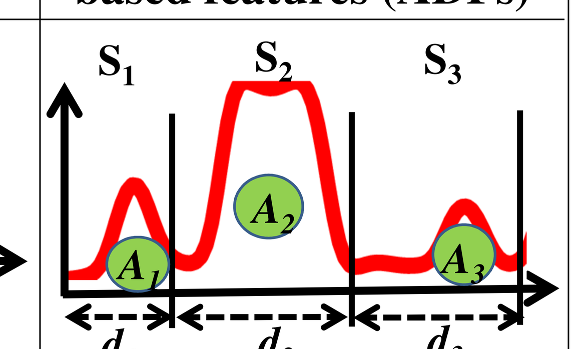
Block diagram representing the proposed approach



Sonority based feature computation:

20-dim features for each syllable S_i

- 10-dim syllable level features using x {x is S-TCSSBC within S_i}
  - 5-dim SFs, 3-dim TFs, 2-dim ADFs
- 10-dim syllable nuclei level features using x_l {x_l is S-TCSSBC within the syllable nuclei of S_i}
  - 5-dim SFs, 3-dim TFs, 2-dim ADFs

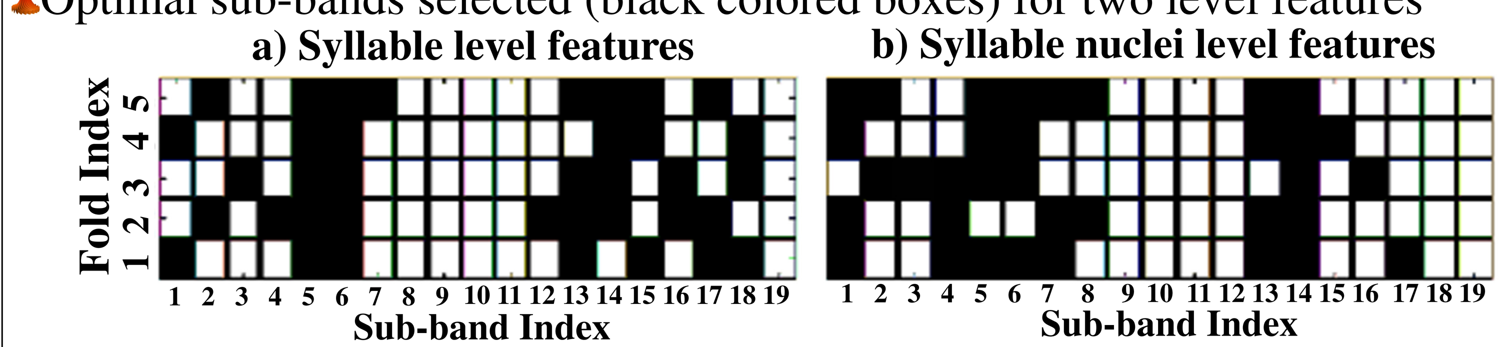
Strength based features (SFs)	Temporal variability based features (TFs)	Area & duration based features (ADFs)
Let z is equal to either x or x_l of length N 1. *Mean* ($\hat{z} = \frac{1}{N}\sum z$) 2. *Standard deviation* ($\frac{1}{N}\sqrt{\sum(z-\hat{z})^2}$) 3. *Geometric mean* ($\sqrt[N]{\prod z}$) 4. *Range* (max($z$)-min($z$)) 5. *Median of z*	 $p(m) = \frac{y(m)}{\sum y(m)}, \mu = \sum m p(m)$; y is resampled values of x or x_l to a fixed length. 1. $\sigma = \sqrt{\sum(m-\mu)^2 p(m)}$ 2. $\gamma = \frac{1}{\sigma^3}\sum(m-\mu)^3 p(m)$ 3. $\kappa = \frac{1}{\sigma^4}\sum(m-\mu)^4 p(m)$	 A_i, d_i are area & duration under x or x_l of S_i. 1. $\widehat{A_i} = \frac{A_i}{A_1+A_2+A_3}$ 2. $\widehat{d_i} = \frac{d_i}{d_1+d_2+d_3}$

Experimental set-up

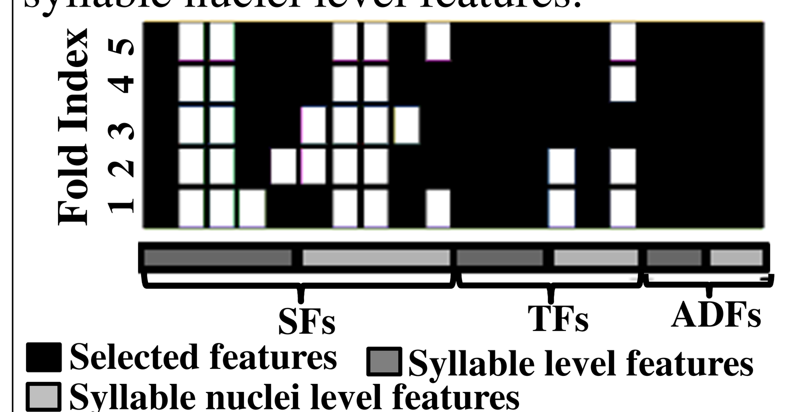
- We consider unweighted accuracy (UA) and weighted accuracy (WA) as objective measures.
- We consider work by Tepperman et al. [4] as the baseline method.
- Experiments are conducted on ISLE corpus containing 7834 sentences.
- We perform the experiments under two setups -- 1) five fold cross validation 2) as in baseline.
- In the cross validation, we use three fold for training, one fold for feature selection and one fold for testing. We find the optimal sub-bands using one fold selected randomly from training set, in which half of the data is selected for SVM training and remaining for selecting the sub-bands.
- We select STC parameters identical to work by Wang et al. [2].
- We use SVM classifier with RBF kernel for the classification task with the complexity parameter (C) equal to 1.0 and gamma (γ) equal to 1/number of features.
- In the post processing, we use estimated labels and decision scores from SVM classifier.

Results

Optimal sub-bands selected (black colored boxes) for two level features



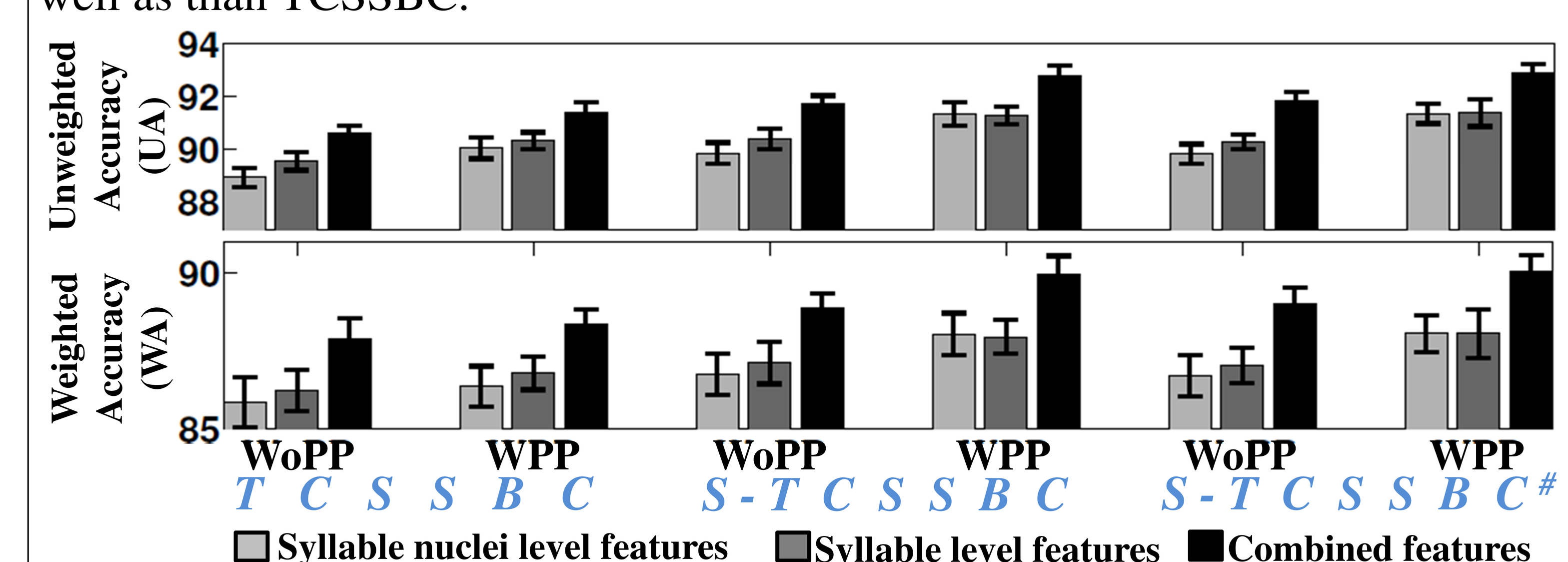
ADFs are selected in both syllable and syllable nuclei level features.



UAs estimated from the baseline as well as from the proposed approach without & with post processing (WPP & WoPP).

	Baseline		S-TCSSBC#	
	WoPP	WPP	WoPP	WPP
GER	85.57	85.81	84.29	87.53
ITA	82.57	83.17	83.73	86.26

S-TCSSBC with optimal features performs better than with all features as well as than TCSSBC.



Conclusion & future work

- Sonority based feature contour is proposed for automatic syllable stress detection task unlike traditional short-time energy contour.
- The contour is computed by combining the sonority motivated cues with sub-band short-time energy contours reflecting prominence measures.
- Experiments with ISLE corpus reveal that the proposed method improves the stress detection performance compared to baseline scheme.
- Future work includes the use of the proposed features for the stress detection task in the native English speech as well as non-native English speech from the nativities other than German and Italian.

References

1. Alan Cruttenden, Gimson's pronunciation of English, Routledge, 2014.
2. Supriya Nagesh, Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5400–5404, 2016.
3. Dagen Wang and Shrikanth S Narayanan, "Robust speech rate estimation for spontaneous speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2190–2201, 2007.
4. Joseph Tepperman and Shrikanth 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.

This work is supported by "Prime Minister Fellowship for Doctoral Research", and we are thankful to "Xerox Research Center India" for the fellowship.