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

Chiranjeevi Yarra1, Om D. Deshmukh2, Prasanta Kumar Ghosh1
1Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India; 2Xerox Research Center India, Bangalore, India; chiranjeeviiyarra@iiisc.ernet.in, om.d.deshmukh@xerox.com, prasanta@iiisc.ernet.in

Problem definition

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

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

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 variations, but not the proposed hypothetical sonority contour.

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

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

Proposed approach

1. Block diagram representing the proposed approach

Syllable boundaries

Speech signal

Split into syllable segments

S-TCSBBC

STC

1. Intensity 2. Duration 3. Pitch

Features Classifier

Forward sub-band selection

S-TCSBBC = Syllable boundaries

S-TCSBBC

Sonority based Features

D1: Decision score

Correct detection Incorrect detection Learnt during training

Sonority based feature computation:

- 10-dim syllable level features using x (z, v i j k t I A. 67 v i j k t I A. 67 z v i j k t I A. 67)
- 10-dim syllable level features using x (w, w i j k t I A. 67 w v i j k t I A. 67 w v i j k t I A. 67 w v i j k t I A. 67)

Strength based features (SFs)

Let z is equal to either x or x0 of length N

1. Mean (x = z / N)

N

2. Standard deviation (σ = 1 / N ∑(x - x̄)²)

3. Geometric mean (g = 1 / N ∏x)

4. Range (max(z) - min(z))

5. Medium of z

Temporal variability based features (TFs)

1. Unstressed

2. Stressed

3. Area & duration under or x or x0 of length N

4. äk = x̂k - x̂k-1

5. äd = x̂d - x̂d-1

Experimental set-up

We consider unbiased accuracy (UA) and weighted accuracy (WA) as objective measures.

We consider work by Tetterman 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 and 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 10 and gamma (gamma) 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

<table>
<thead>
<tr>
<th>Sub-band index</th>
<th>Syllable features</th>
<th>Syllable stress detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-TCSBBC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 native languages other than German and Italian.

References


