Acoustic and articulatory feature based speech rate estimation using a convolutional dense neural network

†Renuka Mannem, †Jhansi Mallela, †Aravind Ila and †Prasanta Kumar Ghosh

†Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India-560 012
*Electronics and Communication Engineering, RGUKT, APIIT, RK Valley, Kadapa, India-516 320

Introduction

- **Speech rate estimation:** To estimate the number of speech units (syllables) in a given test speech using both acoustic and articulatory features.
- **Challenge:** Articulatory features are not directly available, unlike acoustic features from speech recording.
- **Data set:** USC-TIMIT corpus
  - Stimuli: 460 sentences from MOCHA-TIMIT

Proposed Approach

- **Feature Extraction:**
  1. **Acoustic features:** MFCC features computed with window length of 20 msec and a shift of 10 msec.
  2. **Articulatory features:** The articulogram is a time-varying sequence of vocal tract shape during articulation. The steps involved in this are:
    - i) Estimate air tissue boundaries (ATB’s) from rtMRI videos using SegNet [2].
    - ii) Generate 30-dimensional VTTP from ATB [3].
  3. **Estimated articulatory features:** An acoustic-to-articulatory inversion (AAI) method [4] is deployed to estimate articulogram from MFCC using BLSTM (2 hidden layers with 128 units).

CDNN-based speech rate estimation

- **Illustration of feature representations for the given speech waveform:**
- **Illustration of CDNN architecture:**
  - Based on the type of features used for CDNN, they are denoted as, $CDNN_{mfcc}$, $CDNN_{arti}$.
  - Articulogram from rtMRI, $CDNN_{arti}$ : Interpolated Articulogram, $CDNN_{arti}$ : Estimated Articulogram.
  - The syllable rate estimation is formulated as a regression problem; hence, mean squared error (MSE) loss is optimized to train the CDNNs.

Experiments & Results

- **Baseline Approach (TCSSBC):** Peak detection strategy that uses temporal and selected sub-band correlation based feature contour.
- **In our work,** experiments are performed in two conditions using both acoustic and articulatory features: 1) seen subject condition and 2) unseen subject condition.

<table>
<thead>
<tr>
<th></th>
<th>Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub</td>
<td>F1</td>
<td>F1</td>
</tr>
<tr>
<td>TCSSBC</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>$CDNN_{mfcc}$</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>$CDNN_{arti}$</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>$CDNN_{arti}$</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>$CDNN_{arti}$</td>
<td>0.68</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The $CDNN_{arti}$ performs better in all and yields 81.58% and 73.68% improvement over the baseline TCSSBC approach for seen and unseen subject conditions respectively.

- **The proposed $CDNN_{arti}$ approach requires both acoustic and articulatory features during training but only acoustic data during testing.**
- **Results are not consistent across all subjects due to limited number of subjects (four) that affected the generalizability of the network for a new test subject.**

Conclusion

- **We proposed CDNN based approach using acoustic and articulatory features for speech rate estimation.**
- **Future work:**
  1. To implement an end-to-end model which extracts both acoustic and articulatory features from the speech data and estimates the speech rate.
  2. To increase number of speakers in training to acquire generalizability on test.
  3. Experimenting with different noisy conditions to obtain a robust speech rate estimation model.

Acknowledgement: Authors thank the Pratiksha Trust for their support.

References


Author Details:

†Authors contributed equally to this manuscript.

**Data availability:** The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.