Relating articulatory motions in different speaking rates

Asthya Singh¹, G. Nisha Meenakshi² and Prasanta Kumar Ghosh²

¹Electrical Communication Engineering, ²Electrical Engineering, Indian Institute of Science (IISc) Bangalore, India-560012

Introduction

What is known: Speaking rate (phonemes/second) affects the range of articulatory movements [1]

Aim: To gain some insights about the differences in articulation corresponding to speech in different rates by quantifying the relationship between them

Problem Statement: To learn and examine the functions that transform the articulatory movements corresponding to speech in Neutral (N) rate to that in Slow (S) or Fast (F) rate,

Introduction

1. Full affine transformation matrix, \( g_r \)
2. Diagonal affine transformation matrix, \( g_d \)
3. Non linear transformation function, \( DNN \)

Candidate Transformation Functions (TF)

| 1. Articulatory data: Electromagnetic Articulography (EMA) - AG501 |
| 2. Subjects: 2 females (F1, F2) and 3 males (M1, M2 and M3) |
| 3. Sensors: 9, along the mid-sagittal X-Z plane (9 x 2 = 18 trajectories, Eq. UL and IL) |

Approach

Target articulatory trajectory (F or S) denoted as \( R \)

Training utterances: \( i = 1, 2, \ldots, M \)

Objective function:

\[
[g', [w_r]] = \arg \min_{g, [w_r]} \Delta(g, [w_r]), \quad (1)
\]

\[
\Delta(g, [w_r]) = \sum_{i=1}^{n} D_N(w_i, N), \quad (2)
\]

where, \( D_N(x, y) \) : Euclidean distance between \( x \) and \( y \) along warping path \( w \).

Requirement to learn TF: Aligned N and R data, \( \{N_i, R_i\} \), via warping paths \( [w_r] \)

\[
g' = \arg \min_{g} \|N_i - R_i\|^2 \quad (3)
\]

Requirement to compute warping paths: Known \( g \)

\[
[w_r] = \arg \min_{w_r} D_N(g(N_i), R_i) \quad i = 1, 2, \ldots, M\]

Iterative algorithm to alternately optimize for \( g_r \) and \( [w_r] \) [3]:

Result: Optimal TF, \( g_r = g_r(1) \) and \( [w_r] = [w_r(1)] \)

Algorithm: Initialization: \( D^{(0)} = \infty \), Iteration \( j = 1 \), \( g_r(1) = g_r(j); \)

Optimize for \( [w_r] \), \( i = 1, 2, \ldots, M \); \n
Compute \( D_k(0), [w_r(0)] \); \n
while \( D^{(j-1)} < D^{(j)} \) do

\[ j = j + 1; \]

Optimize for \( g_r(j) \) using \( [w_r(j-1)] \), Eq. (3); \n
Optimize for \( [w_r] \) using \( g_r(j) \), Eq. (4); \n
Compute \( D_k(j), [w_r(j)] \), Eq. (2); \n
end

Experimental Setup

Baseline Function: Identity, \( g \)

4 Fold setup

Train - Test Split: \( 3 : 1 \)

Validation set for DNN: 15% of the training set

The TFs are learnt subject wise for N2F and N2S separately

Evaluation Metric: Average DTW distance (in mm) between the transformed N and target R trajectories of the test set

Results

<table>
<thead>
<tr>
<th></th>
<th>( N2F )</th>
<th>( F1 )</th>
<th>( F2 )</th>
<th>( M1 )</th>
<th>( M2 )</th>
<th>( M3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_r(0) )</td>
<td>6.51</td>
<td>6.40</td>
<td>6.81</td>
<td>6.00</td>
<td>5.64</td>
<td>5.64</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>5.32</td>
<td>5.21</td>
<td>5.38</td>
<td>5.24</td>
<td>4.90</td>
<td>4.90</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>4.82</td>
<td>5.04</td>
<td>4.74</td>
<td>4.88</td>
<td>4.59</td>
<td>4.59</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Summary of Key Findings

<table>
<thead>
<tr>
<th>( N2S )</th>
<th>( F1 )</th>
<th>( F2 )</th>
<th>( M1 )</th>
<th>( M2 )</th>
<th>( M3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_r(0) )</td>
<td>5.46</td>
<td>8.27</td>
<td>7.52</td>
<td>6.03</td>
<td>6.30</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>1.27</td>
<td>1.07</td>
<td>0.77</td>
<td>0.89</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>5.93</td>
<td>8.14</td>
<td>6.90</td>
<td>5.76</td>
<td>6.12</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>1.24</td>
<td>1.02</td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>5.16</td>
<td>7.46</td>
<td>6.35</td>
<td>5.35</td>
<td>5.65</td>
</tr>
<tr>
<td>( g_r(0) )</td>
<td>0.85</td>
<td>1.13</td>
<td>1.02</td>
<td>0.71</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Optimal TF

\( DNN \)

Hypothesis

Hyper articulation in S

Complex mapping

Gross articulatory motion in F

Less complex mapping

Common Attributes

Several N trajectories contribute to reconstruct 1 R trajectory

TFs are subject dependent

\( g_r(0) \) TF learnt from one fold for the 5 subjects:

Acknowledgements

We thank the 5 subjects and Mr. Aravind Iilla for the data collection. We also thank the Pratiksha Trust for their support.

References

2. A. Wrench, MOCHA-TIMIT speech database, The 18th International Conference on Pattern Recognition, 1999
3. G. Nisha Meenakshi, P. Ghosh Reconstruction of articulatory movements during neutral speech from those during whispered speech, JASA, 2018