The X-ray Microbeam database

- The X-ray microbeam (XRMB) database [1] consists of continuous speech along with time aligned coordinates of pellets placed at various points along the vocal tract as shown in Fig. 1.

- The database consists of 2500 speech utterances with pellet trajectories across 57 different speakers.
- The database provides X and Y coordinates of the pellets as the subject produces the desired utterance.
- However, these pellet positions are closely connected to speaker anatomy and head positions leading to cross-speaker variability.
- A total of 1720 utterances across 46 speakers were successfully transformed to TVs. This formed the natural XRMB database.

XRMB pellet trajectories to
Tract Variables

- Tract variables (TVs) are continuous time functions that specify the shape of the vocal tract in terms of constriction degree and location of the constriction.
- The pellet trajectories from XRMB data are geometrically transformed into relative TV measures.
- The XRMB data after transformation is represented in terms of six TVs as shown in Fig. 2.

Synthetic XRMB database with synthetic TVs from TADA

- The Task Dynamics and Applications (TADA) system [2] from Haskins laboratories along with HLSyn [3] was used to generate synthetic speech along with time aligned TVs for the XRMB conditions.
- The synthetic speech was then warped to align with the natural XRMB utterances. The same warping function was used to warp the synthetic TVs.
- The details of this synthetic XRMB data generation are given in [4].

The objective of this research is to train efficient acoustic to articulatory speech inversion systems on natural speech that provide reliable articulatory features for unseen test speakers.

We constructed two such systems using feedforward neural networks. One was trained using natural speech data from the XRMB database and the second using synthetic data generated by the Haskins Laboratories TADA model that approximated the XRMB data.

XRMB pellet trajectories were first converted into vocal tract constriction variables (TVs), providing a relative measure of constriction kinematics (location and degree).

TV estimators were tested using previously collected acoustic data on the utterance ‘perfect memory’ spoken at slow, normal, and fast rates.

The TV estimator trained on XRMB data (but not on TADA data) was able to recover the tongue tip gesture for /r/ in the fast utterance despite the gesture occurring partly during the acoustic silence of the closure.

The XRMB system (but not the TADA system) could distinguish between bunched and retroflexed /r/.

Speaker dependent TV estimators were trained an tested on matched and mismatched speaker conditions.

Results

- The multi layer feed forward neural networks were trained to estimated the TVs from contextualized MFCCs.

- The correlation between estimated and ground truth TVs for the natural and synthetic TV estimators are shown in Table 1.

- The correlation between estimated and ground truth TVs for the synthetic TV estimator are much better than that of the natural TV estimator because the synthetic speech is single speaker with very less variability in production.

Table 1: Test set correlation values for TV estimators trained on natural and synthetic XRMB database

<table>
<thead>
<tr>
<th>TV</th>
<th>Synthetic TV estimator</th>
<th>Natural speech TV estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.8881</td>
<td>0.7101</td>
</tr>
<tr>
<td>LP</td>
<td>0.9150</td>
<td>0.5721</td>
</tr>
<tr>
<td>TBCD</td>
<td>0.9409</td>
<td>0.5699</td>
</tr>
<tr>
<td>TCD</td>
<td>0.9106</td>
<td>0.5630</td>
</tr>
<tr>
<td>TLCD</td>
<td>0.8876</td>
<td>0.5944</td>
</tr>
<tr>
<td>TTD</td>
<td>0.9223</td>
<td>0.7303</td>
</tr>
</tbody>
</table>

Uncovering coarticulation using TV estimators

- The speaker dependent TV estimator was trained using one of the speakers’ data from the XRMB database.
- The speaker dependent TV estimator was used to obtain the TVs for the utterance ‘perfect memory’ for the two cases – (1) clearly spoken (2) fast spoken.
- Figure 1 also shows the estimated TVs (T, TBCD, TCD) for clearly spoken utterance and Figure 6 shows the same TVs for fast spoken utterance.

In [5], the authors show that the spacing between F5 and F4 is an acoustic signature for tongue shape. F5-F4 for retroflex /r/ is around 1340Hz And for bunched /r/ is around 740Hz.

We low pass filtered the speech to remove the F4 and F5 frequencies from the spectrum. We then passed the filtered speech through the TV estimator.

We can see from figure 8 that the TV estimator does not show the correct tongue constrictions for the bunched and retroflexed /r/, providing further evidence that F5 and F4 information provides the distinction between bunched and retroflexed /r/.