Introduction
The goal of our research is to develop a gesture and landmark-based speech recognition system. This work presents the initial step to achieve such a system, where the mapping between the speech signal and the various types of time functions (VTTF) is considered.

VTTFs are time-varying physical realizations of articulatory gestures at distinct vocal tract sites for a given utterance.

VTTFs describe the geometric features of the vocal tract shape in terms of constriction degree and location.

VTTFs would help to obtain gestural information from the acoustics.

The proposed mapping is based on a hierarchical support vector regression (SVR) followed by Kalman smoothing. The smoothed VTTFs will be used to recover gestural information from the speech signal.

Motivation
• Automatic Speech Recognition (ASR) suffers from poor performance in casual speech due to acoustic variations
• Phone-based ASR systems suffer due to co-articulation
• Phone units are distinctive in the cognitive domain but are not invariant in the physical domain.
• Phone-based ASR systems do not adequately model the temporal overlap that occurs in casual speech.

To address co-articulation, diphone and triphone models are used.

They limit contextual influence to only immediately close neighbors.

• Articulatory phonology proposes the articulatory constriction gesture as an invariant action unit and argues that human speech can be decomposed into a constellation of articulatory gestures.

• This representation allows for temporal overlap between neighboring gestures.

• Acoustic variations are accounted for by gestural co-articulation and reduction.

How can Gestures address speech variability?

The estimated VTTFs were found to be noisy so that we used a smoother to help reduce the root-mean-square error (rmse).

Two types of smoothing explored
(1) Running average and (2) Kalman smoothing

rmse for the different VTTFs

<table>
<thead>
<tr>
<th>VTTF</th>
<th>e-SVR</th>
<th>After averaging</th>
<th>After kalman smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLO</td>
<td>0.039</td>
<td>0.040</td>
<td>0.036</td>
</tr>
<tr>
<td>VEL</td>
<td>0.025</td>
<td>0.025</td>
<td>0.023</td>
</tr>
<tr>
<td>LP</td>
<td>0.566</td>
<td>0.536</td>
<td>0.508</td>
</tr>
<tr>
<td>LA</td>
<td>2.316</td>
<td>2.227</td>
<td>2.178</td>
</tr>
<tr>
<td>TTD</td>
<td>3.537</td>
<td>3.345</td>
<td>3.293</td>
</tr>
<tr>
<td>TBCD</td>
<td>1.878</td>
<td>1.749</td>
<td>1.681</td>
</tr>
<tr>
<td>TTCL</td>
<td>8.372</td>
<td>8.037</td>
<td>7.485</td>
</tr>
</tbody>
</table>

Post-processing

Hierarchical SVR structure

Certain VTTFs (TTCL, TBCD, TDCD and TBCD) are known to be functionally dependent upon other VTTFs (FVY).

The remaining VTTFs (GLO, VEL, LA and LP) are relatively independent and can be obtained directly from the APs.

The c-SVR [20] (a generalization of the SVM) works for only single output.

• c-SVR uses the parameter ε (the unsatisfactory coefficient) to control the number of support vectors.

• SVR projects input data into a high dimensional space via non-linear mapping and performs linear regression in that space.

For 8 VTTFs, 8 different c-SVRs were created, the hierarchical structure was based upon the correlation amongst the VTTFs.

The hierarchical c-SVR architecture for generating VTTFs.

Conclusion

• Proposed a hierarchical SVR for VTTF estimation from acoustics.
• Kalman smoothing helped to reduce rmse by 9.44%.
• Contextual information helped to reduce reconstruction error.

Future Directions
• Explore other machine learning approaches to perform the same task.
• Address the issue of non-uniqueness in speech inversion in a probabilistic manner.
• Explore other acoustic features.
• Evaluate the system performance when speech is corrupted with noise.

References


Acknowledgement

This research was supported by NSF grant IIS0703389 and NIH grant DC-02717. We would also like to acknowledge the helpful advice from Elliot Saltzman (BS), Mark Hasegawa-Johnson (UIUC) and Louis Goldstein (USC).