Towards articulatory-based adaptation in recognition of dysarthric speech

Frank Rudzicz
PhD candidate
Department of Computer Science
University of Toronto
Automatic speech recognition (ASR)

“open the pod bay doors”

Acoustic model

Language model

open(podBay.doors);
Dysarthria

Articulatory disabilities resulting in unintelligible speech.

Causes (e.g.)
- Cerebral palsy.
- Parkinson’s disease.
- Amyotrophic lateral sclerosis.

Effects (e.g.)
- Poor respiration, phonation, and resonance.
- Inaccurate timing and speed.
- Imprecise consonants.
- Distorted or indistinguishable vowels.
- Atypical control of volume and pitch.
Dysarthria and speech recognition

Word-recognition accuracy given traditional ASR (red):
- **84.9%** with typical speech.
- **3.1%** with severe dysarthria.

We can improve accuracy by
- Increasing the complexity of our models (# Gaussians).
- Adapting existing models to new data (green).
- Learning from scratch (blue).

Still, fewer than 10% of words will be recognized. *Why?*
ASR has trouble with dysarthric speech because

- it is inconsistent.
- it has indistinguishable targets.
- it is interrupted by disfluencies (e.g., stutter).

E.g., given repetitions of the sequence /iy p ae/ (right)
Non-dysarthric speakers can produce very distinct vowels.

It is difficult to correctly associate an observed sound with a cluster if those clusters are not clearly delineated.
Towards articulatory models

Dysarthria is characterized by aberrant neuro-motor signals. The result is atypical articulation.

Can we identify how dysarthric articulation differs from the general public?

Can we use these articulatory models within ASR?
"Torgo"
A database of aligned acoustics and articulation.

Population
7 individuals with CP or ALS, + matched controls.

Data
3 hrs of sentences and words each.

Methods
Electromagnetic articulography.
3D video reconstruction.
3D position, velocity, and acceleration of coils on tongue and lips are measured at 200Hz. Dynamic patterns of articulation are aligned with their acoustic consequences, and analyzed.
Articulatory models for ASR

Build models that combine acoustics, articulation, and phonemes.

Dynamic Bayesian networks model the probabilistic behaviour of speech over time, and condition observed acoustics on articulatory sources.
## Success of articulatory models

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Dysarthric</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong> (Hidden Markov Model)</td>
<td>-</td>
<td>14.1</td>
<td>72.8</td>
</tr>
<tr>
<td>Neural network</td>
<td>Phonological</td>
<td>14.4</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>EMA</td>
<td>16.1</td>
<td>72.7</td>
</tr>
<tr>
<td>Dynamic Bayesian network</td>
<td>Phonological</td>
<td>15.0</td>
<td>73.3</td>
</tr>
<tr>
<td></td>
<td>EMA</td>
<td>16.5</td>
<td>73.8</td>
</tr>
<tr>
<td>Conditional random field</td>
<td>Phonological</td>
<td>16.5</td>
<td>73.4</td>
</tr>
<tr>
<td></td>
<td>EMA</td>
<td>16.8</td>
<td>73.5</td>
</tr>
</tbody>
</table>

Statistically significant improvement in phoneme recognition.
Current work

Estimate articulator positions given only acoustics.

More sophisticated models of speech production and recognition.
Future work

Telephony

Buy ticket... AC490... yes...

AAC

Dictation
Appendices
## Nosology of the dysarthrias

<table>
<thead>
<tr>
<th>Type</th>
<th>Primary lesion site</th>
<th>Perceptual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ataxic</td>
<td>cerebellum</td>
<td>Articulatory inaccuracy, prosodic excess, phonatory insufficiency</td>
</tr>
<tr>
<td>Flaccid</td>
<td>lower motor neurons</td>
<td>Phonatory incompetence, resonatory incompetence</td>
</tr>
<tr>
<td>Spastic</td>
<td>upper motor neurons</td>
<td>Prosodic excess, articulatory-resonatory incompetence</td>
</tr>
<tr>
<td>Spastic-Flaccid</td>
<td>upper/lower motor neurons</td>
<td>Prosodic excess, articulatory-resonatory incompetence, phonatory stenosis</td>
</tr>
<tr>
<td>Hyperkinetic (chorea)</td>
<td>Basal ganglia (especially putamen or caudate)</td>
<td>articulatory-resonatory incompetence, phonatory stenosis</td>
</tr>
<tr>
<td>Hyperkinetic (distonia)</td>
<td>ibid</td>
<td>articulatory inaccuracy, prosodic excess, phonatory stenosis</td>
</tr>
<tr>
<td>Hypokinetic</td>
<td>Basal ganglia (substantia nigra)</td>
<td>Prosodic insufficiency, phonatory incompetence</td>
</tr>
</tbody>
</table>
Discriminative classification

- **Neural networks**
  - Multi-layer perceptron.
  - Recurrent Elman network.
- **Support vector machines**
  - Radial-basis function kernel.
  - Dynamic-time warped kernel.
- We discriminatively identify PFs, and combine these classifiers to identify triphones.

![Diagram of neural network and SVM](image)
### Phonological features

<table>
<thead>
<tr>
<th>Feature (PF)</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manner</strong></td>
<td>approximant, fricative, nasal, retroflex, silence, stop, vowel</td>
</tr>
<tr>
<td><strong>Place</strong></td>
<td>alveolar, bilabial, dental, labiodental, silence, velar, nil</td>
</tr>
<tr>
<td><strong>High/Low</strong></td>
<td>high, mid, low, silence, nil</td>
</tr>
<tr>
<td><strong>Voice</strong></td>
<td>voiced, unvoiced</td>
</tr>
<tr>
<td><strong>Front/Back</strong></td>
<td>front, central, back, nil</td>
</tr>
<tr>
<td><strong>Round</strong></td>
<td>round, non-round, nil</td>
</tr>
<tr>
<td><strong>Static</strong></td>
<td>static, dynamic</td>
</tr>
</tbody>
</table>