Improving the Intelligibility of Dysarthric Speech

Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken b, Janice Staehelyb

Speech Communication 2007
Motivation:
Various research has proven that vowel articulation is a key factor in dysarthric speech intelligibility.

Goal:
Propose a method to transform unintelligible vowels into intelligible ones to improve the overall intelligibility of dysarthric speech
Method Outline

1. Recording database
2. Training input/output features
3. Training
4. Analysis
5. Transformation
6. Synthesis
7. Testing
Method
Overview
Recording Data

Participants
1 dysarthric speaker
- Female native American English speaker
- Friedreich’s ataxia
- Clinically judged to be 70% intelligible

1 non-dysarthric speaker
- Male native American English speaker
Recording Data

- 278 isolated monosyllabic CVC words
- Recorded in 16 kHz, 16-bit PCM format using headset

Vowels
- Front: /i/, /ɪ/, /E/, /@/
- Back: /u/, /U/, /ʌ/, /A/

Consonants
- Stops: /p/, /b/, /t/, /d/, /k/, /g/
- Fricatives: /v/, /s/, /z/, /S/
- Approximates: /l/, /j/, /w/

Omissions
- Nasal consonants, diphthongs (gliding vowel), /⟩/ vowel
Recording Data

Procedure
1. CVC word presented on screen
2. Rhyming word to CVC word then shown
3. Played recording of CVC from non-dysarthric speaker
4. Tone prompt to say CVC word

All 278 words were recorded over a single 1.25 hour session with several rest breaks
One More Addition

Vowel-target database

9 words recorded several times over some months:


**Purpose:** draw out formant values reaching their intended target without influence of other another speech sounds
Recording Data

Recording data was segmented into training & testing sets:

- Separate sets for dysarthric/non-dysarthric speakers
- Training sets = 214 feature vectors
- Testing sets = 64 feature vectors
  - Uniform distribution of the vowels (8 occurrences of all 8 vowels)
Analysis

Energy, formant and voicing features derived from ESPS Waves+ software package

1. Find F1 & F2 **stable points**
2. Measure F1 & F2 at their stable points to **estimate the formant targets**
3. F3 stable points calculated the same way as F2
F1 & F2 Stable Points

Diagram (a) shows a scatter plot of F1 vs. F2 with various symbols and colors indicating different points. Diagram (b) is similar with a different set of data points.
Transformation

Output Features:

- F1 and F2 (previously shown to be useful)
- F3 and vowel duration (new features in this paper)
- These with energy and pitch trajectories specify how vowel portions are made up

Input Features:

- F1, F2 and ?
# Input Feature Configurations

<table>
<thead>
<tr>
<th>Set</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F1median + F2median</td>
</tr>
<tr>
<td>2</td>
<td>F1stable + F2stable</td>
</tr>
<tr>
<td>3</td>
<td>F1median + F2median + duration</td>
</tr>
<tr>
<td>4</td>
<td>F1stable + F2stable + duration</td>
</tr>
<tr>
<td>5</td>
<td>F1median + F2median + F3median</td>
</tr>
<tr>
<td>6</td>
<td>F1stable + F2stable + F3stable</td>
</tr>
<tr>
<td>7</td>
<td>F1median + F2median + F3median + duration</td>
</tr>
<tr>
<td>8</td>
<td>F1stable + F2stable + F3stable + duration</td>
</tr>
<tr>
<td>9</td>
<td>F1stable + F2stable + F3stable + duration + F1slopleft + F1slopeRight</td>
</tr>
<tr>
<td>10</td>
<td>F1stable + F2stable + F3stable + duration + F2slopleft + F2slopeRight</td>
</tr>
<tr>
<td>11</td>
<td>F1stable + F2stable + F3stable + duration + F1slopleft + F1slopeRight + F2slopleft + F2slopeRight</td>
</tr>
<tr>
<td>12</td>
<td>F1stable + F2stable + F3stable + duration + F2slopeRight</td>
</tr>
<tr>
<td>13</td>
<td>F1stable + F2stable + F2rms</td>
</tr>
<tr>
<td>14</td>
<td>F1stable + F2stable + duration + F2rms</td>
</tr>
<tr>
<td>15</td>
<td>F1stable + F2stable + F3stable + F2rms</td>
</tr>
<tr>
<td>16</td>
<td>F1stable + F2stable + F3stable + duration + F2rms</td>
</tr>
<tr>
<td>17</td>
<td>F1stable + F2stable + F2poly</td>
</tr>
<tr>
<td>18</td>
<td>F1stable + F2stable + duration + F2poly</td>
</tr>
<tr>
<td>19</td>
<td>F1stable + F2stable + F3stable + F2poly</td>
</tr>
<tr>
<td>20</td>
<td>F1stable + F2stable + F3stable + duration + F2poly</td>
</tr>
<tr>
<td>21</td>
<td>F1stable + F2stable + F3stable + duration + energy</td>
</tr>
</tbody>
</table>
Best Input Feature Set?

Configuration Scores:
- Number of times correct vowel recognized by Eq. (1)
- Normalize by number of samples in the test set

Configuration 8: best average score = 0.62
- Consisted of F1, F2, F3 stable points and the duration

Rotation 6: closest to the average performance of all rotations averaged by all configurations
Config. Scores

![Box plot showing values for different configurations.](image)
Rotation Scores
Transformation Input & Output Data

Input: training data set of dysarthric speaker

Output: context-independent vowel-specific target values
  ○ Generic values from Peterson and Barney’s work
  ○ Individual values derived from vowel-target DB

Context-independent vowel-specific target values:
  ● Doesn’t need formant matching, stable point mapping
## Output Target Features

### Table 3
Output target feature values

<table>
<thead>
<tr>
<th>Vowel (word)</th>
<th>F1 (Hz)</th>
<th>F2 (Hz)</th>
<th>F3 (Hz)</th>
<th>Duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Generic</td>
<td>Individual</td>
<td>Generic</td>
<td>Individual</td>
</tr>
<tr>
<td>/i/ (he)</td>
<td>310</td>
<td>300</td>
<td>2790</td>
<td>2300</td>
</tr>
<tr>
<td>/I/ (hit)</td>
<td>430</td>
<td>400</td>
<td>2480</td>
<td>1900</td>
</tr>
<tr>
<td>/E/ (heck)</td>
<td>610</td>
<td>600</td>
<td>2330</td>
<td>1850</td>
</tr>
<tr>
<td>/@/ (hack)</td>
<td>860</td>
<td>750</td>
<td>2050</td>
<td>1800</td>
</tr>
<tr>
<td>/u/ (who)</td>
<td>370</td>
<td>350</td>
<td>950</td>
<td>1150</td>
</tr>
<tr>
<td>/U/ (hook)</td>
<td>470</td>
<td>500</td>
<td>1160</td>
<td>1100</td>
</tr>
<tr>
<td>// (huff)</td>
<td>760</td>
<td>700</td>
<td>1400</td>
<td>1500</td>
</tr>
<tr>
<td>/A/ (ha)</td>
<td>850</td>
<td>750</td>
<td>1220</td>
<td>1300</td>
</tr>
</tbody>
</table>
Training the Analyzed Data

**Gaussian Mixture Model:** maps dysarthric to non-dysarthric speech data relationship

\[
\hat{y} = \mathcal{F}(x|t, \theta) = \sum_{q=1}^{Q} t_q \cdot p(c_q|x, \theta) \tag{1}
\]

\[
p(c_q|x, \theta) = \frac{\alpha_q \cdot \mathcal{N}(x, \mu_q, \Sigma_q)}{\sum_{i=1}^{Q} \alpha_i \cdot \mathcal{N}(x, \mu_i, \Sigma_i)} \tag{2}
\]

\[
\mathcal{N}(x, \mu, \Sigma) = \frac{e^{-0.5(x-\mu)\Sigma^{-1}(x-\mu)}}{(2\pi)^{N/2} \sqrt{\det(\Sigma)}} \tag{3}
\]
Equation 1 (GMM)

Finds the class yielding the maximum posterior probability

Advantages
- Covariance modelled strictly from dysarthric speaker
- Led to large reduction in modelling parameters

Drawbacks
- Cannot map coarticulation patterns
Transformed Vowels

Green dotted ellipses: Dysarthric
Blue dashed-dotted ellipses: Non-dysarthric
Red solid ellipses: Transformed
dysarthric
Synthesis

1. Calculate stable-point vector of formant trajectory
2. Apply transformation function to stable-point vector

Crossfade Trajectory VS. Straight-line Trajectory

- Crossfade avoids discontinuities in trajectory, but may identify vowels incorrectly at the CVC boundaries
- Straight-line vulnerable to discontinuities at CVC boundaries, but maintains the constant transformed formant value throughout the vowel duration
Synthesis Trajectories
Transform Results
Transform Results
Testing Setup

- 64 CVCs / 8 conditions => 512 stimuli for evaluation
- 24 listeners, each evaluated 128 stimuli
- Randomized order of CVCs
- Each listener heard CVC twice in 2 different conditions
- Intelligibility was computed as number of correctly identified vowels divided by the total number of vowels
- Testing done over speakers, but not a sound-isolated room
Test Stimuli

B & C = true tests of improving the intelligibility
A - B & C = tests the transformation method
B - D & C - E = tests the mapping function
B - C & D - E = tests speaker-indep/dep formant targets

<table>
<thead>
<tr>
<th>Stimulus condition</th>
<th>Duration</th>
<th>Pitch</th>
<th>Energy</th>
<th>Unvoiced</th>
<th>Voiced</th>
<th>Formants</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – dysarthric</td>
<td>dysarthric</td>
<td>dysarthric</td>
<td>dysarthric</td>
<td>dysarthric</td>
<td>dysarthric</td>
<td>n/a</td>
</tr>
<tr>
<td>B – dysarthric-map-generic</td>
<td>mapping</td>
<td>synthetic</td>
<td>smoothed</td>
<td>dysarthric</td>
<td>synthetic</td>
<td>map-gen.</td>
</tr>
<tr>
<td>C – dysarthric-map-individual</td>
<td>mapping</td>
<td>synthetic</td>
<td>smoothed</td>
<td>dysarthric</td>
<td>synthetic</td>
<td>map-ind.</td>
</tr>
<tr>
<td>D – dysarthric-oracle-generic</td>
<td>oracle</td>
<td>synthetic</td>
<td>smoothed</td>
<td>dysarthric</td>
<td>synthetic</td>
<td>oracle-generic</td>
</tr>
<tr>
<td>E – dysarthric-oracle-individual</td>
<td>oracle</td>
<td>synthetic</td>
<td>smoothed</td>
<td>dysarthric</td>
<td>synthetic</td>
<td>oracle-individual</td>
</tr>
<tr>
<td>F – normal-synth-individual</td>
<td>normal</td>
<td>synthetic</td>
<td>smoothed</td>
<td>normal</td>
<td>synthetic</td>
<td>normal-individual</td>
</tr>
<tr>
<td>G – normal-synth-contextdependent</td>
<td>normal</td>
<td>synthetic</td>
<td>smoothed</td>
<td>normal</td>
<td>synthetic</td>
<td>normal-contextdependent</td>
</tr>
<tr>
<td>H – normal</td>
<td>normal</td>
<td>normal</td>
<td>normal</td>
<td>normal</td>
<td>normal</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Testing Procedures

● Testing done on specifically designed graphical UI
● 3 preliminary stages to familiarize participants
  ○ (1) words representing each vowel shown on screen; vowel sound played if word clicked; each word had to be clicked before next stage
  ○ (2) CVC words played; 10 correctly identified to go to next stage
  ○ (3) Same as stage 2 except dysarthric CVC words included now
Testing Participants

Listeners

- Reported to have normal hearing
- Native American English speakers
- No clinical or research work in dysarthria
- Paid to participate

Qualifications

- Achieve a minimum of 90% correct identification rate of the non-dysarthric speaker
Results

- **B & C** higher than **A** by increase of 6%
- **D & E** better than **B & C**
- Mapping function: /@/ and /A/ **good**; /E/ and /i/ **okay**; /I/, /u/, /U/ and /^/ **poor**

Table 5
Intelligibility of stimulus conditions in percent

<table>
<thead>
<tr>
<th>Stimulus condition</th>
<th>/i:/</th>
<th>/I/</th>
<th>/E/</th>
<th>/@/</th>
<th>/u/</th>
<th>/U/</th>
<th>/^/</th>
<th>/A/</th>
<th>Average</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – dysarthric</td>
<td>73</td>
<td>63</td>
<td>40</td>
<td>10</td>
<td>92</td>
<td>73</td>
<td>27</td>
<td>6</td>
<td>48 (13)</td>
<td>69</td>
</tr>
<tr>
<td>B – dysarthric-map-generic</td>
<td>67</td>
<td>42</td>
<td>54</td>
<td>83</td>
<td>46</td>
<td>52</td>
<td>19</td>
<td>73</td>
<td>54 (10)</td>
<td>56</td>
</tr>
<tr>
<td>C – dysarthric-map-individual</td>
<td>50</td>
<td>65</td>
<td>56</td>
<td>83</td>
<td>56</td>
<td>56</td>
<td>21</td>
<td>46</td>
<td>54 (10)</td>
<td>75</td>
</tr>
<tr>
<td>D – dysarthric-oracle-generic</td>
<td>96</td>
<td>42</td>
<td>77</td>
<td>94</td>
<td>81</td>
<td>63</td>
<td>71</td>
<td>92</td>
<td>77 (11)</td>
<td>81</td>
</tr>
<tr>
<td>E – dysarthric-oracle-individual</td>
<td>94</td>
<td>83</td>
<td>88</td>
<td>96</td>
<td>88</td>
<td>63</td>
<td>54</td>
<td>71</td>
<td>79 (9)</td>
<td>100</td>
</tr>
<tr>
<td>F – normal-synth-individual</td>
<td>96</td>
<td>88</td>
<td>92</td>
<td>98</td>
<td>98</td>
<td>73</td>
<td>96</td>
<td>94</td>
<td>92 (8)</td>
<td>88</td>
</tr>
<tr>
<td>G – normal-synth-context-dependent</td>
<td>96</td>
<td>83</td>
<td>83</td>
<td>98</td>
<td>71</td>
<td>79</td>
<td>94</td>
<td>98</td>
<td>88 (9)</td>
<td>100</td>
</tr>
<tr>
<td>H – normal</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>92</td>
<td>100</td>
<td>100</td>
<td>98 (3)</td>
<td>100</td>
</tr>
</tbody>
</table>
Summary

- Proposed mapping statistically a lot more intelligible than the original dysarthric speech, but not near as intelligible as the oracle condition
- **No difference** between speaker-dep VS speaker-indep
- Difference between the dysarthria-oracle condition and the normal-synth-individual conditions (*intelligibility relies on consonant*)
- Synthesis framework decreased the intelligibility
Summary (Expert Listener)

- Expert listener on original dysarthric speech was still better than the average listener on the mapped dysarthric speech

- Expert intelligibility increased for the transformed system with the dysarthric individual formant targets
Conclusions

● Intelligibility improved from 48% to 54%
● Results very preliminary

Future Work: Sentence level processing
  ○ Consonant-vowel boundary detector
  ○ Diphthongs - more than one stable point
  ○ F0 predictions more complex
Conclusions

Possibilities to increase intelligibility:

- Use formant frequencies “de-coarticulated” from surroundings in transformation
- More sophisticated formant trajectory model
- Choose a more naturally wide vowel space
- Transforming consonants
Questions?