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.

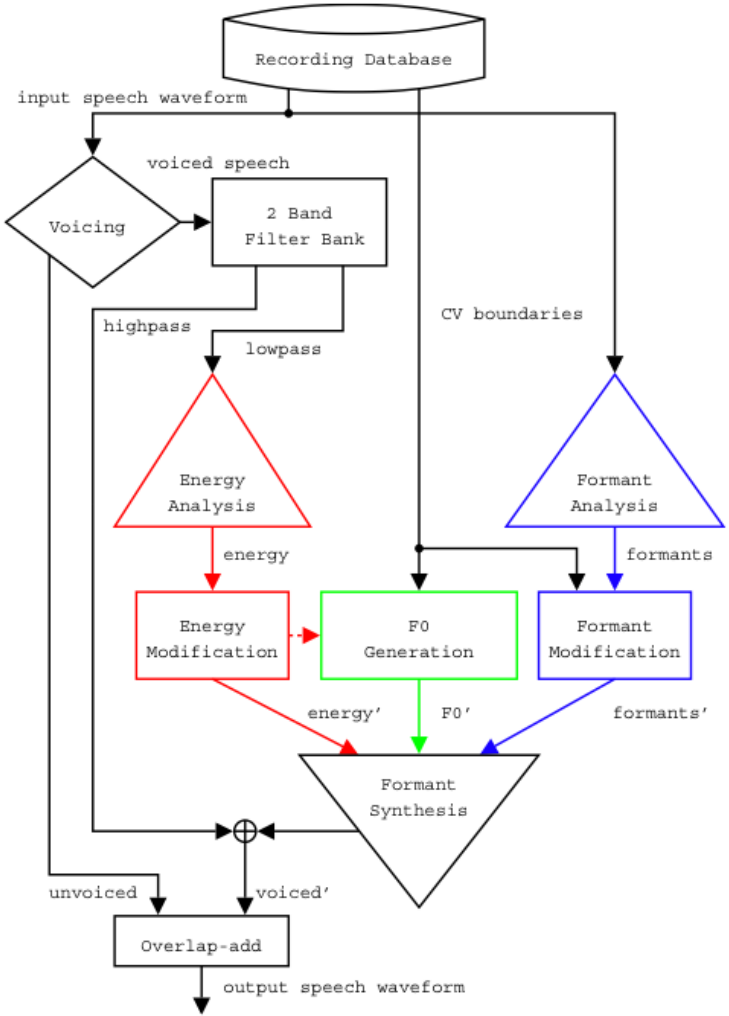
Goal:

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/, /I/, /E/, /@/
- Back: /u/, /U/, /^/, /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:

“he”, “hit”, “heck”, “hack”, “who”, “hook”, “huff”, “hoe”, “ha”

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:

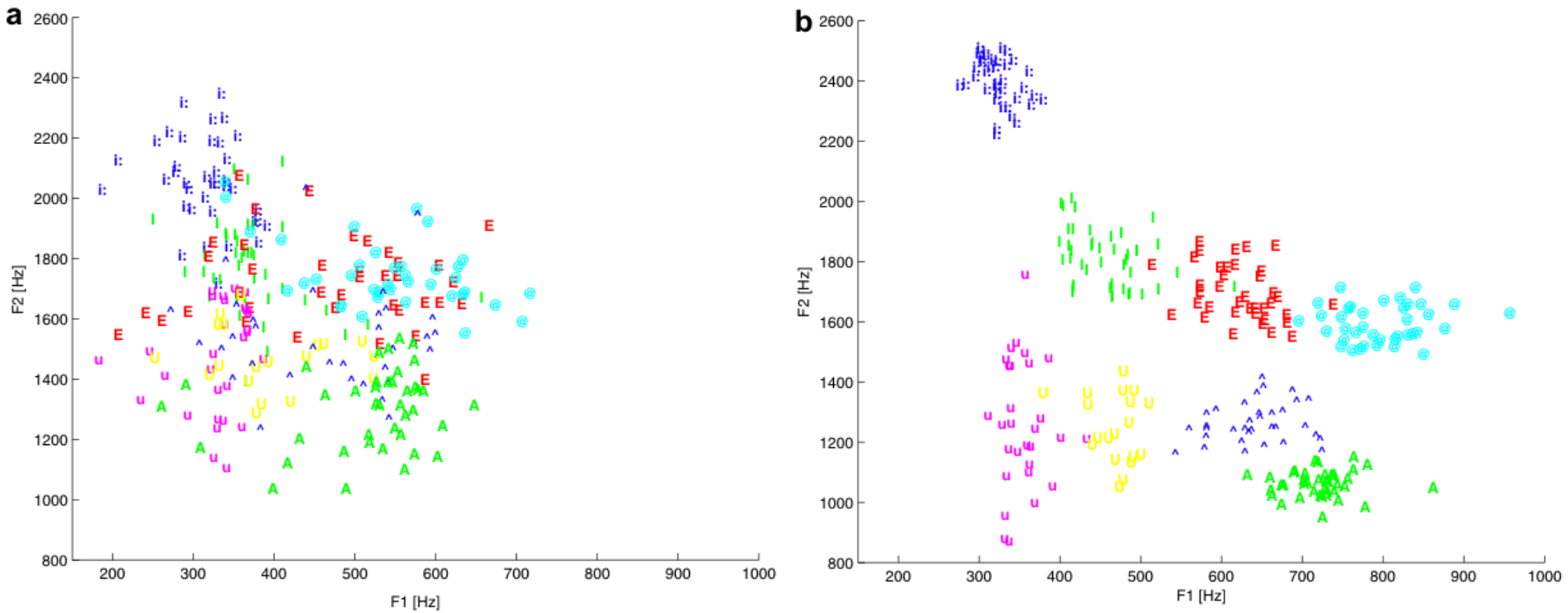
- Separate sets for dysarthric/non-dysarthric speakers
- Training sets = **214** feature vectors
- Testing sets = **64** feature vectors
  - Uniform distribution of the vowels (**8 occurrences of all 8 vowels**)

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



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 2
Sets of features used as input to the transformation function

Set	Features
1	F1median + F2median
2	F1stable + F2stable
3	F1median + F2median + duration
4	F1stable + F2stable + duration
5	F1median + F2median + F3median
6	F1stable + F2stable + F3stable
7	F1median + F2median + F3median + duration
8	F1stable + F2stable + F3stable + duration
9	F1stable + F2stable + F3stable + duration + F1slopeLeft + F1slopeRight
10	F1stable + F2stable + F3stable + duration + F2slopeLeft + F2slopeRight
11	F1stable + F2stable + F3stable + duration + F1slopeLeft + F1slopeRight + F2slopeLeft + F2slopeRight
12	F1stable + F2stable + F3stable + duration + F2slopeRight
13	F1stable + F2stable + F2rms
14	F1stable + F2stable + duration + F2rms
15	F1stable + F2stable + F3stable + F2rms
16	F1stable + F2stable + F3stable + duration + F2rms
17	F1stable + F2stable + F2poly
18	F1stable + F2stable + duration + F2poly
19	F1stable + F2stable + F3stable + F2poly
20	F1stable + F2stable + F3stable + duration + F2poly
21	F1stable + F2stable + F3stable + duration + energy

Best Input Feature Set?

Configuration Scores:

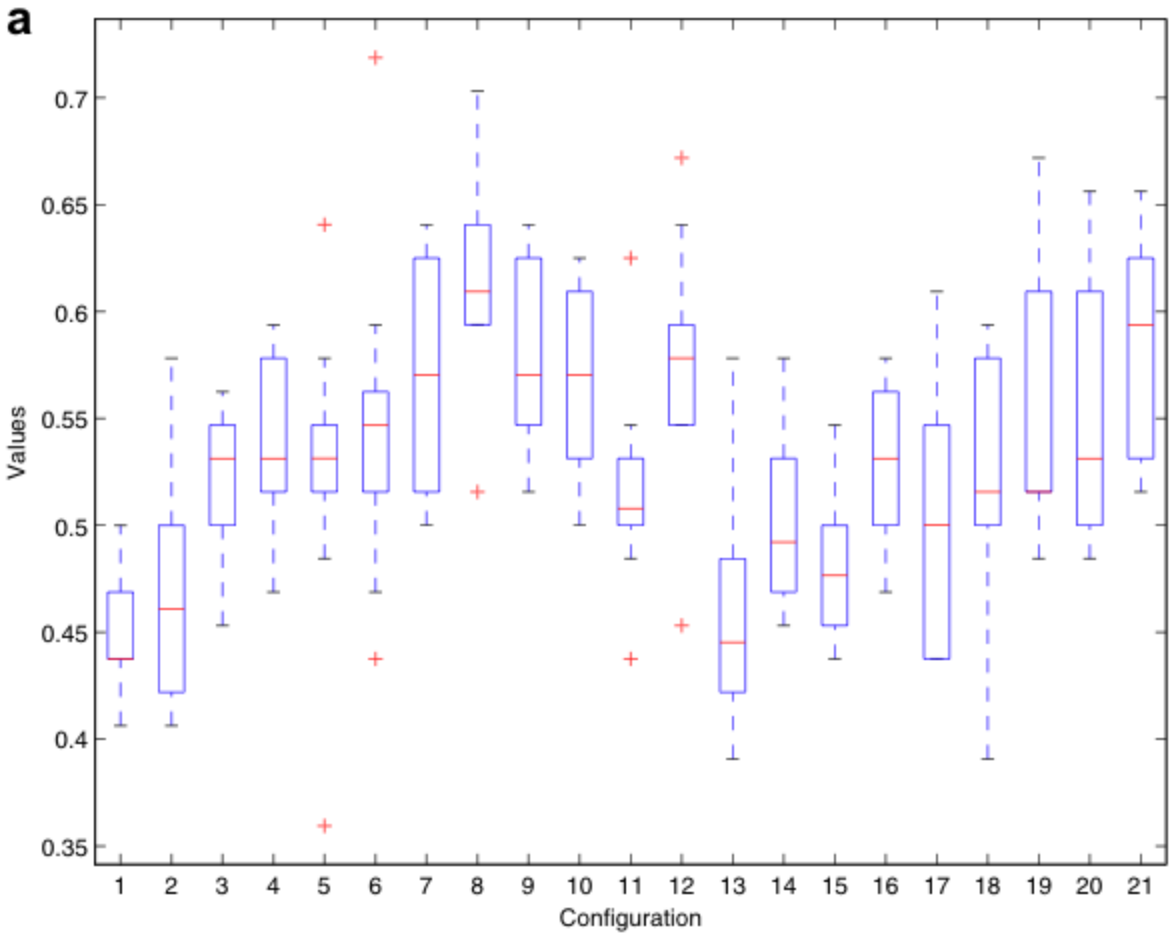
- Number of times correct vowel recognized by Eq. (1)
- Normalize by number of samples in the test set

Configuration 8: best average score = 0.62

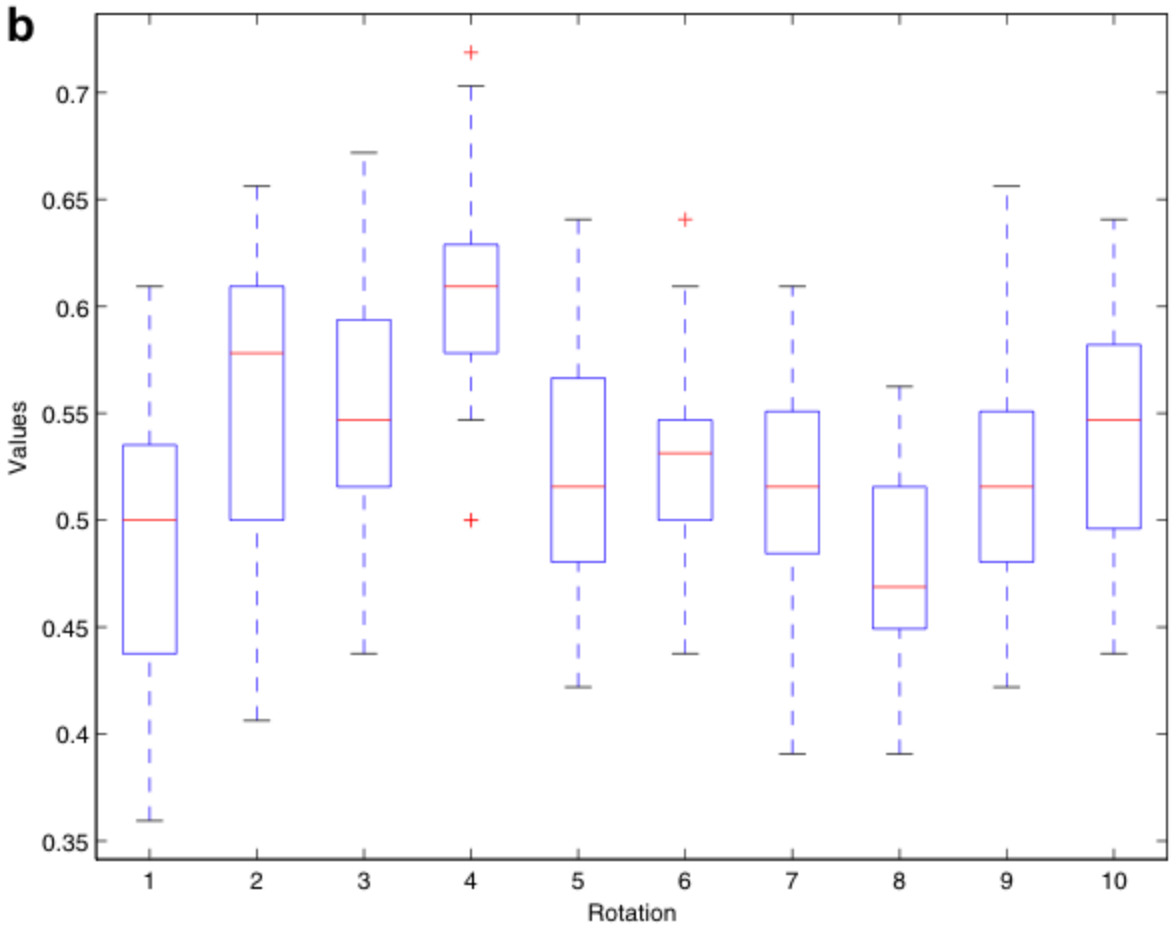
- Consisted of F1, F2, F3 stable points and the duration

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

Config. Scores



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

Vowel (word)	F1 (Hz)		F2 (Hz)		F3 (Hz)		Duration (ms)
	Generic	Individual	Generic	Individual	Generic	Individual	
/i:/ (he)	310	300	2790	2300	3310	2900	212
/I/ (hit)	430	400	2480	1900	3070	2650	138
/E/ (heck)	610	600	2330	1850	2990	2750	167
/@/ (hack)	860	750	2050	1800	2850	2850	257
/u/ (who)	370	350	950	1150	2670	2400	179
/U/ (hook)	470	500	1160	1100	2680	2700	120
/^/ (huff)	760	700	1400	1500	2780	2800	150
/A/ (ha)	850	750	1220	1300	2810	2750	224

Training the Analyzed Data

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

$$\hat{\boldsymbol{y}} = \mathscr{F}(\boldsymbol{x}|\boldsymbol{t}, \theta) = \sum_{q=1}^{Q} \boldsymbol{t}_q \cdot p(c_q|\boldsymbol{x}, \theta) \tag{1}$$

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

$$\mathcal{N}(\boldsymbol{x}, \mu, \Sigma) = \frac{\mathrm{e}^{-0.5(\boldsymbol{x}-\mu)\Sigma^{-1}(\boldsymbol{x}-\mu)}}{(2\pi)^{N/2}\sqrt{\det(\Sigma)}} \tag{3}$$

Equation 1 (GMM)

Finds the class yielding the maximum posterior probability

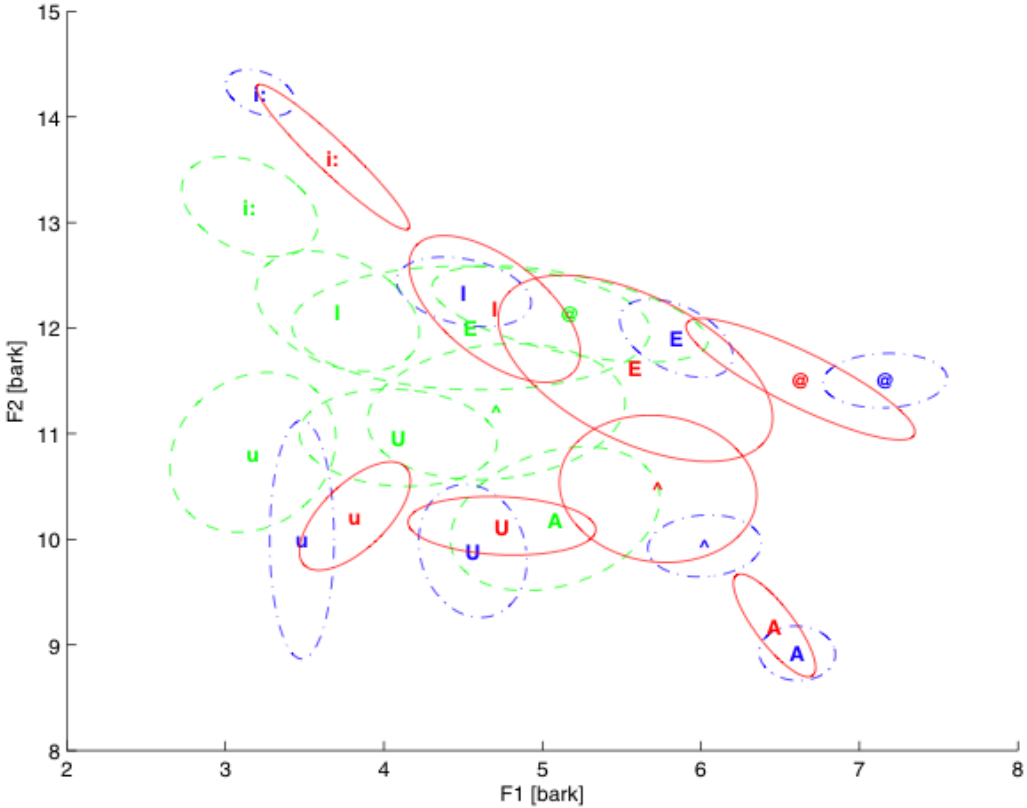
Advantages

- 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

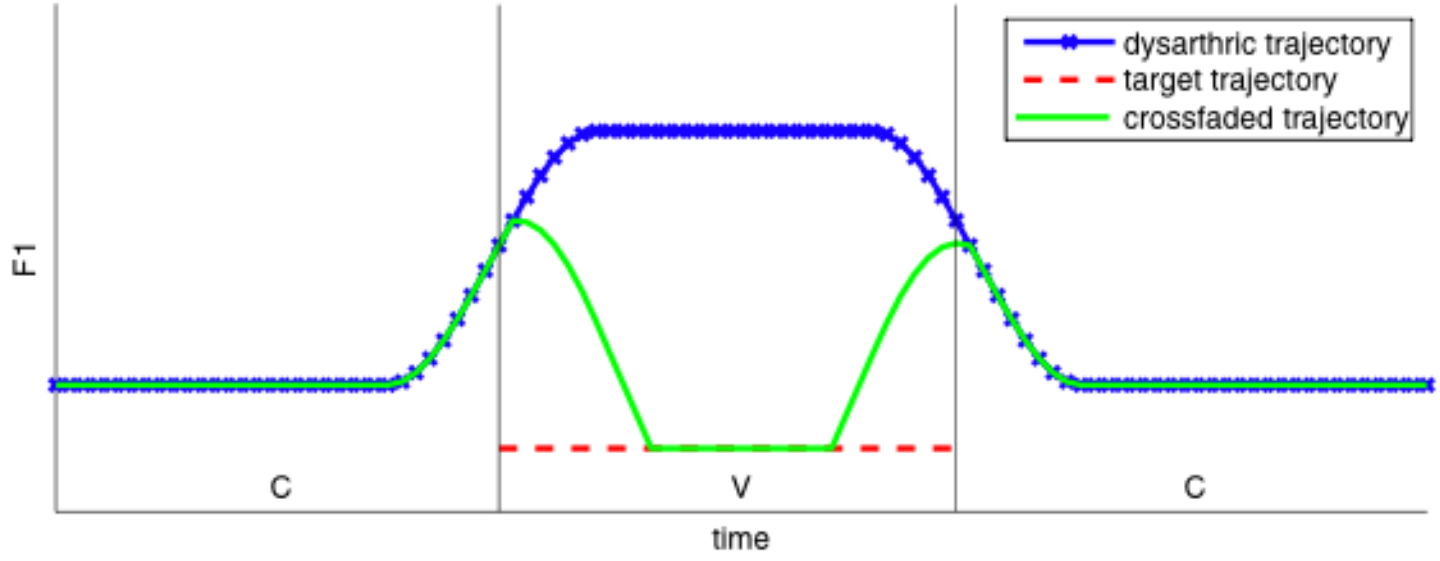
Synthesis

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

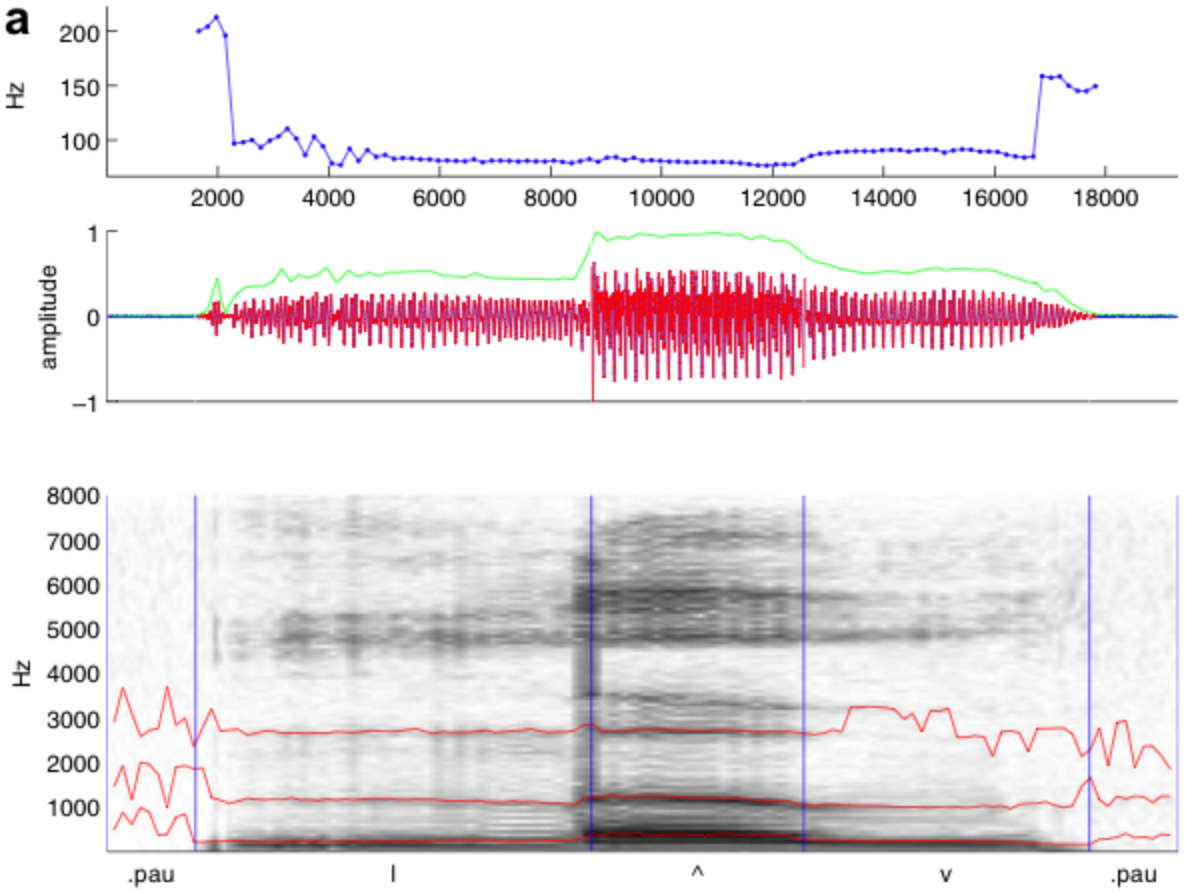
Crossfade Trajectory VS. Straight-line Trajectory

- 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

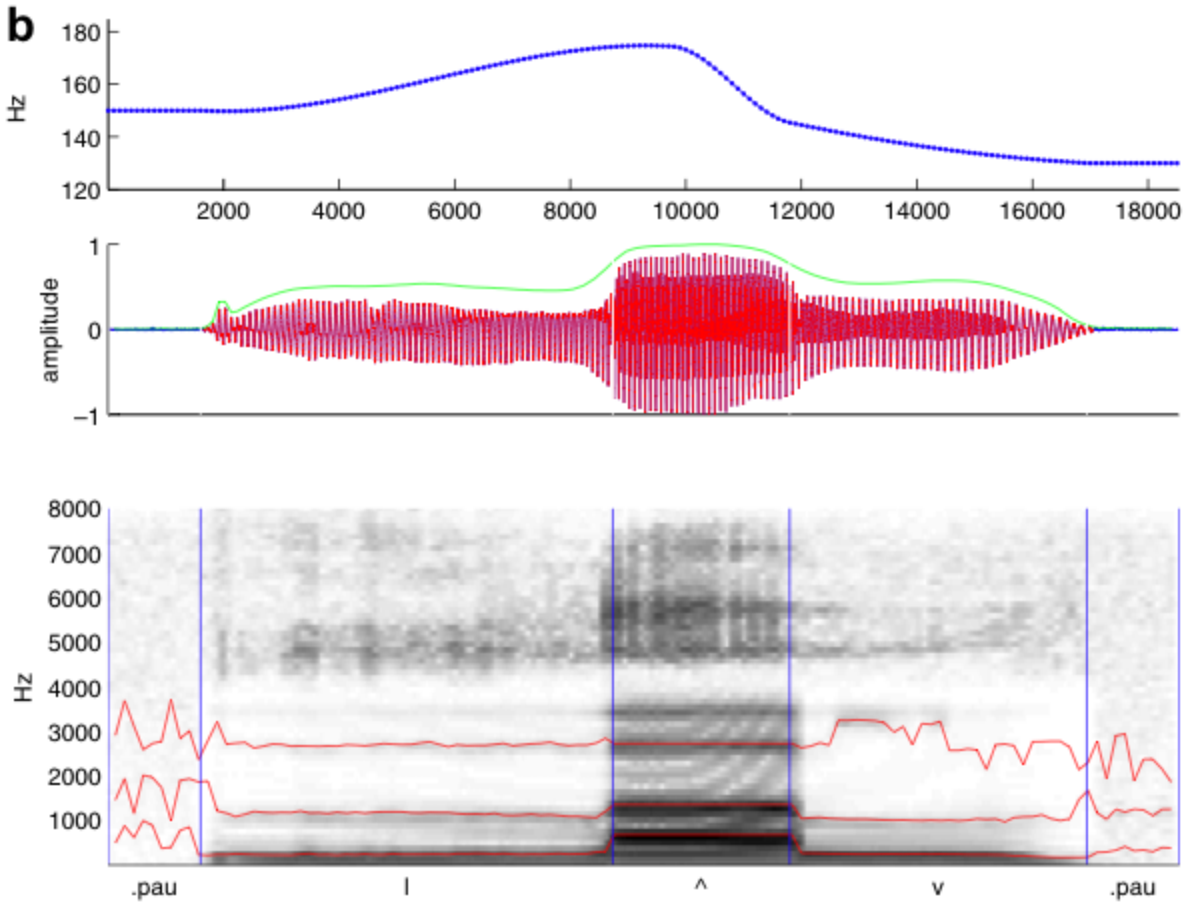
Synthesis Trajectories



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 4
Stimulus conditions and their associated configurations

Stimulus condition	Duration	Pitch	Energy	Unvoiced	Voiced	Formants
A – dysarthric	dysarthric	dysarthric	dysarthric	dysarthric	dysarthric	n/a
B – dysarthric-map-generic	mapping	synthetic	smoothed	dysarthric	synthetic	map-gen.
C – dysarthric-map-individual	mapping	synthetic	smoothed	dysarthric	synthetic	map-ind.
D – dysarthric-oracle-generic	oracle	synthetic	smoothed	dysarthric	synthetic	oracle-generic
E – dysarthric-oracle-individual	oracle	synthetic	smoothed	dysarthric	synthetic	oracle-individual
F – normal-synth-individual	normal	synthetic	smoothed	normal	synthetic	normal-individual
G – normal-synth-contextdependent	normal	synthetic	smoothed	normal	synthetic	normal-contextdependent
H – normal	normal	normal	normal	normal	normal	n/a

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

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

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?