Dysarthric Speech Recognition Using Listen-Attend-Spell

Jeremy Tate Campbell (jcamp12@stanford.edu) Stanford University – CS 229

Introduction

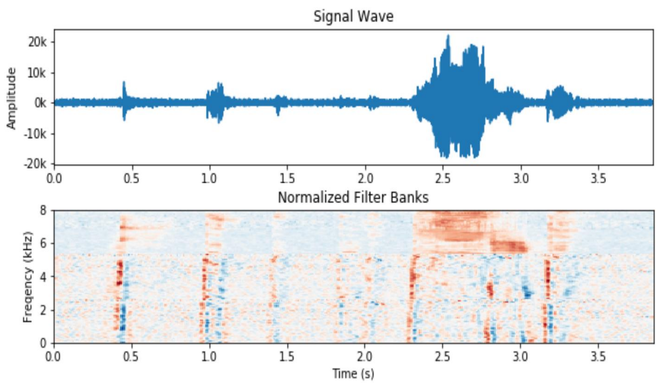
- **Goal:** Build a speech recognition system for people with dysarthria, a motor speech disorder caused by muscle weakness in the face, lips, tongue, or throat.
- **Input:** A single, isolated word from a speaker with dysarthria
- **Output:** The phonetic transcription of the word

Dataset & Features

- From UA-Speech Database (~120 hrs)
- Each file was a one-word utterance
- Words chosen using greedy algorithm to maximize uncommon phonemes
- Transcriptions → phonemes

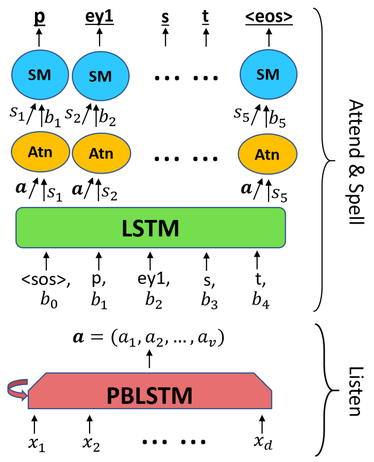
command	k ah0 m ae1 n d
pajamas	p ah0 jh aa1 m ah0 z
observation	aa2 b z er0 v ey1 sh ah0 n

- Audio (WAV) file → filter banks



- Each input feature vector x_i had 123 features from a 10ms window of the normalized filter bank

Model



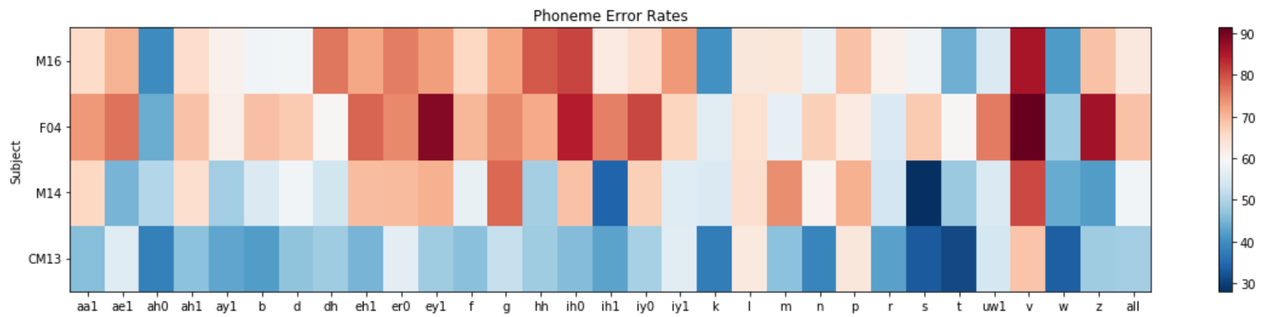
- **Listener:** Pyramidal Bidirectional LSTM – $\boldsymbol{a} = \text{PBLSTM}(\boldsymbol{x})$
- **Speller:** $b_i = \text{Attention}(\boldsymbol{a}, s_i) \quad s_i = \text{LSTM}(y_{i-1}, s_{i-1}, b_{i-1})$
  $p(y_i \mid \boldsymbol{x}, y_{<i}) = \text{Softmax}(s_i, b_i)$
- **Loss:** Average cross-entropy across each phoneme
- **Decoder:** Beam search decoder

Model	Train Set PER (%)	Dev Set PER (%)
Listener: 128 HU, 2 layers, keep_prob = 0.5 Speller: 128 HU, 2 layers, keep_prob = 0.5	48.66	62.56
Listener: 256 HU, 3 layers, keep_prob = 0.7 Speller: 256 HU, 3 layers, keep_prob = 0.7	41.19	64.13
Listener: 256 HU, 3 layers, keep_prob = 0.5 Speller: 128 HU, 2 layers, keep_prob = 0.5	44.20	60.02

LR: 1E-3 w/ decay
epochs: 30
Batch size: 64
Beam width: 16

Results

- **Test Set PER:** 60.15% (~10,000 words)



Best

Control			Dysarthric			Difference		
Phoneme	Example	PER (%)	Phoneme	Example	PER (%)	Phoneme	Example	PER (%)
t	paste – p ey1 s *t*	31.25	w	with – *w* ih1 dh	44.58	l	l – *l* ay1 n	1.71
s	so – *s* ow1	33.17	ah0	the – dh *ah0*	44.99	p	paste – *p* ey1 s t	4.28
w	with – *w* ih1 dh	33.50	k	can – *k* ae1 n	50.35	ah0	the – dh *ah0*	6.82
k	can – *k* ae1 n	37.45	t	paste – p ey1 s *t*	50.50	uw1	do – d *uw1*	7.68
ah0	the – dh *ah0*	38.16	s	so – *s* ow1	51.29	ae1	tab – t *ae1* b	8.06

Worst

Phoneme	Example	PER (%)	Phoneme	Example	PER (%)	Phoneme	Example	PER (%)
iy1	he – hh *iy1*	56.08	er0	her – hh *er0*	73.50	g	go – *g* ow1	22.45
er0	her – hh *er0*	56.41	g	go – *g* ow1	74.90	n	in – ih0 *n*	22.97
l	l – *l* ay1 n	62.69	ey1	way – w *ey1*	77.51	eh1	then – dh *eh1* n	28.15
p	paste – *p* ey1 s t	63.24	ih0	in – *ih0* n	77.96	ey1	way – w *ey1*	28.94
v	of – ah1 *v*	68.69	v	of – ah1 *v*	85.89	ih0	in – *ih0* n	31.96

Discussion

A PER of 60% is somewhat reasonable given the challenging nature of the task, though slightly disappointing. A deeper bidirectional LSTM (trained for CS 230) achieved a PER of 47%. Other state-of-the-art models have PERs of ~35%, but our test subjects had below average speech intelligibility scores. Interestingly, the LAS word error rate was only 5% worse than the CS 230 model, indicating that the beam search did well when the first phonemes were predicted correctly. There was also more randomness in the phonemes that were successful, unlike in the CS 230 model, where phonemes that were stressful on speech muscles generally did worse. We think this LAS model simply requires more data to prevent the speller from overfitting to the limited number of phoneme combinations.

Future Work

- Incorporate prior knowledge of phonetic relationships
- Thorough hyperparameter search
- Transfer learning using large corpus of non-dysarthric speech

References

Special thanks to Professor Mark Hasegawa-Johnson of the University of Illinois for allowing us access to the UA-Speech database he helped to create.

[1] Kim, Heejin, et al. "Dysarthric speech database for universal access research." *Ninth Annual Conference of the International Speech Communication Association*. 2008.