Dysarthric Speech Recognition Using a Deep Bidirectional LSTM

Jeremy Tate Campbell (jcamp12@stanford.edu)   Stanford University – CS 230

Introduction

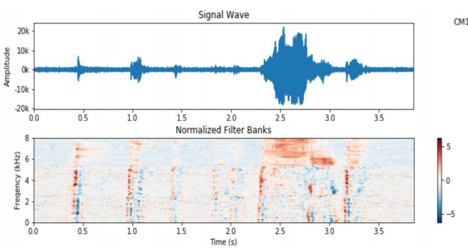
- **Goal:** Build a speech recognition system for people with dysarthria, a motor speech disorder caused by muscle weakness in the face, lips, tongue, or throat.
- **Input:** A single, isolated word from a speaker with dysarthria
- **Output:** The phonetic transcription of the word

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

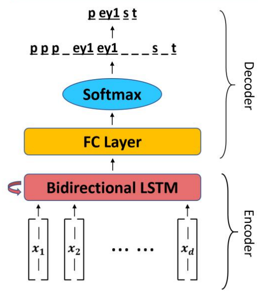
command	k ah0 m ae1 n d
pajamas	p ah0 jh aa1 m ah0 z
observation	aa2 b z er0 v ey1 sh ah0 n

- Audio (WAV) file → filter banks



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

Model



- **Encoder:** Bidirectional LSTM
- **Decoder:** 1 fully-connected layer to a CTC decoder that condenses phonemes that aren't separated by a "blank"
- **Loss:** Negative log-likelihood of true phoneme sequence given the softmax probabilities.

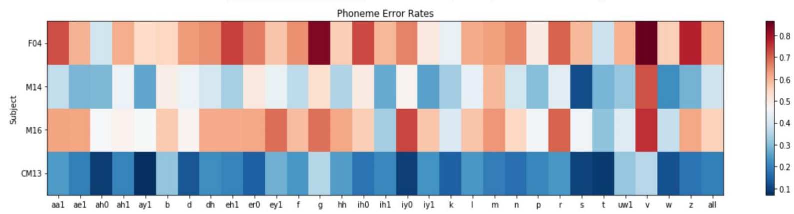
$$\text{Loss} = \sum_{(x,y) \in B} -\log p(y \mid x)$$

Model	Dev Set PER (%)
128 hidden units, 3 layers, keep_prob = 0.5	48.12
128 hidden units, 4 layers, keep_prob = 0.5	47.03
256 hidden units, 5 layers, keep_prob = 0.7	44.50

- Learning rate: 0.001, # epochs: 50, batch size: 256

Results

- **Test Set PER:** 44.68% (~10,000 words)



Best

Control

Phoneme	Example	PER (%)
ay1	line – l *ay1* n	7.03
t	paste – p ey1 s *t*	8.63
ah0	the – dh *ah0*	9.08
iy0	many – m eh1 n *iy0*	9.25
s	so – *s* ow1	10.08

Dysarthric

Phoneme	Example	PER (%)
t	paste – p ey1 s *t*	32.17
ah0	the – dh *ah0*	37.75
s	so – *s* ow1	38.50
w	with – *w* ih1 dh	38.65
k	can – *k* ae1 n	38.94

Difference

Phoneme	Example	PER (%)
uw1	to – t *uw1*	13.34
ih1	it – *ih1* t	19.23
p	paste – *p* ey1 s t	21.01
iy1	he – hh *iy1*	21.24
b	tab – t ae1 *b*	22.69

Worst

Control

Phoneme	Example	PER (%)
ey1	paste – p *ey1* s t	27.75
b	tab – t ae1 *b*	30.92
uw1	to – t *uw1*	31.37
g	go – *g* ow1	35.55
v	of – ah1 *v*	35.65

Dysarthric

Phoneme	Example	PER (%)
r	their – dh eh1 *r*	60.30
m	some – s ah1 *m*	62.35
iy0	many – m eh1 n *iy0*	62.72
g	go – *g* ow1	68.52
v	of – ah1 *v*	78.00

Difference

Phoneme	Example	PER (%)
v	of – ah1 *v*	42.35
ih0	in – *ih0* n	42.40
m	many – *m* eh1 n iy0	42.96
er0	her – hh *er0*	45.77
iy0	many – m eh1 n *iy0*	53.46

Discussion

A PER of 45% is about what we expect – other state-of-the-art models on dysarthric speech have PERs of ~35%, and the subjects in our test set had below average speech intelligibility scores. Phonemes that were similar to another, such as "r" (as in their) and "er0" (as in her), generally were less successful. Certain phonemes that are more stressful on speech muscles, such as "m" and "er0" did particularly worse on dysarthric speakers. Overall, we believe our DBLSTM sufficiently fits the training data and successfully learns the inconsistent temporal acoustic cues present in dysarthric speech.

Future Work

- Incorporate prior knowledge of phonetic relationships
- Conduct a more thorough search of hyperparameter space
- Combine DBLSTM with beam search decoder
- 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.