Automatic Speech Recognition of Scripted Productions from PWAs

Brian MacWhinney, Davida Fromm, Eric Riebling, Florian Metze

Carnegie Mellon University

Objective and Rationale

Transcription of speech productions from persons with aphasia (PWA) and apraxia of speech (AoS) requires painstaking transcription work and analysis.

Our aim was to explore how SpeechKitchen methodology (Metze et al., 2015) could be used for transcription, alignment, and analysis of:

- single monosyllabic words from the Chapel Hill Multilingual Intelligibility Test (CHMIT; Haley, 2011) designed to estimate overall speech production ability in adults with speech difficulties
- scripts produced by persons with aphasia (PWA) receiving aphasia therapy (Fridriksson et al., 2012) and a normal speaker.

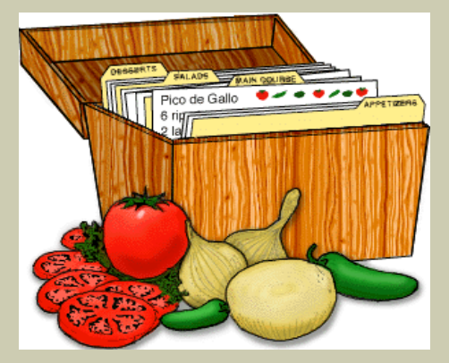
Background

Systems like Praat can produce excellent results but they cannot provide PWAs with immediate feedback regarding the correctness of their productions. Moreover, alignment in Praat is very tedious.

Feedback about correct productions and errors is important for therapy practice in AoS and aphasia. It is also important for interaction with conversational agents and other computerized facilities.

According to Metze at al. (2015), SpeechKitchen provides:

APPLIANCES
speech recognition tool-kits



RECIPES
scripts for creating
state-of the art systems

INGREDIENTS
language data

Methods

CHMIT word list examples:

Set 1	Set 2
lease	lamb
knees	glad
tea	slam
free	swam
bee	bad
trees	cab
flee	track
need	dad
peace	trap
three	grab
key	black
freeze	lap

Script example from PWA therapy :

EGGS
I like to eat scrambled eggs for breakfast.
I like them because they are fast and easy.
To make eggs, I get out a pan and melt some butter over medium heat.
I crack the eggs into the pan and stir.
I like scrambled eggs best so I stir until they are done.

Script example from normal adult speaker:

CLIMATE
Things will change in ways that their fragile environment simply can't support. And that leads to starvation, it leads to uncertainty, it leads to unrest. So that climate changes will be terrible for them.

SpeechKitchen uses 2 possible models for analysis:

- Language – looks for words, based on recognizing phones and comparing to a training set
- Acoustic –looks for phones in a given time segment, moves borders of each segment to maximize the chance of getting the phone correct vis-à-vis its neighboring phones

How to run the Eesen Transcriber at SpeechKitchen – http://speechkitchen.org/

- Download by cloning from git
- Install vagrant
- Upload input audio and script

Results

ASR results using the CHMIT monosyllabic single word list spoken word-by-word by normal speakers were poor. In reality, real-life perception of monosyllables out of context (considering dialect variation) is not close to perfect.

However, ASR recognition (WER) is much better when the list is read as a single sentence.

ASR results using the scripts spoken by normals and PWA yielded more promising results.

Here are examples of the types of outputs available from audio (or video) input:

1. *.ali files – Alignment of word production and time

Speaker is PWA with AoS, Production is from EGGS script

Command: run_align.sh

```
P8_P2_SE_C4_PAR___0000.000_0012.450 1  0      6.36           like
P8_P2_SE_C4_PAR___0000.000_0012.450 1  6.36  0.63            to
P8_P2_SE_C4_PAR___0000.000_0012.450 1  6.99 1.08             eat
P8_P2_SE_C4_PAR___0000.000_0012.450 1  8.07  0.33            <unk>
P8_P2_SE_C4_PAR___0000.000_0012.450 1  8.40  1.41            eggs
P8_P2_SE_C4_PAR___0000.000_0012.450 1  9.81  0.51            for
P8_P2_SE_C4_PAR___0000.000_0012.450 1 10.32  2.13            breakfast
P8_P2_SE_C4_PAR___0012.450_0019.240 1 12.45 0.54             i
P8_P2_SE_C4_PAR___0012.450_0019.240 1  12.99 0.30            like
P8_P2_SE_C4_PAR___0012.450_0019.240 1  13.29 1.23            <unk>
P8_P2_SE_C4_PAR___0012.450_0019.240 1  14.52 1.29            cause
P8_P2_SE_C4_PAR___0012.450_0019.240 1  15.81 0.69            they
P8_P2_SE_C4_PAR___0012.450_0019.240 1  16.50 0.12            are
P8_P2_SE_C4_PAR___0012.450_0019.240 1  16.62 0.09            <unk>
P8_P2_SE_C4_PAR___0012.450_0019.240 1  16.71 0.54            and
P8_P2_SE_C4_PAR___0012.450_0019.240 1  17.25 1.98            <unk>
```

Compare *.ali file with manual transcription in CHAT format

*PAR: like to eat sæmbəld@u [: scrambled] eggs for breakfast .
*PAR: I like tɪm@u [: them] (be)cause they are sfæst@u [: fast] and sizɪ@u [: easy].

Note:

ASR marked <unk> for the words that were unknown (out of vocabulary). All other words were recognized accurately.

2. *.ctm files – Phonemic transcription with timing information

Speaker is non-aphasic control, Production is Climate script

Command: speech2phonectm.sh test2.mp3

```
test2_S0___0000.090_0006.460 S0    θ              0.0   0.45
test2_S0___0000.090_0006.460 S0    ɪ              0.45  0.09
test2_S0___0000.090_0006.460 S0    ŋ              0.54  0.06
test2_S0___0000.090_0006.460 S0    z              0.6   0.09
test2_S0___0000.090_0006.460 S0    w              0.69  0.09
test2_S0___0000.090_0006.460 S0    ɪ              0.78  0.06
test2_S0___0000.090_0006.460 S0    ɫ              0.84  0.03
test2_S0___0000.090_0006.460 S0    tʃ             0.87  0.12
test2_S0___0000.090_0006.460 S0    eɪ             0.99  0.15
test2_S0___0000.090_0006.460 S0    n              1.14  0.12
test2_S0___0000.090_0006.460 S0    dʒ             1.26  0.09
test2_S0___0000.090_0006.460 S0    ɪ              1.35  0.12
test2_S0___0000.090_0006.460 S0    m              1.47  0.03
test2_S0___0000.090_0006.460 S0    w              1.5   0.09
test2_S0___0000.090_0006.460 S0    eɪ             1.59  0.12
test2_S0___0000.090_0006.460 S0    z              1.71  0.27
test2_S0___0000.090_0006.460 S0    ð              1.98  0.48
test2_S0___0000.090_0006.460 S0    ʌ              2.46  0.03
test2_S0___0000.090_0006.460 S0    t              2.49  0.06
```

Results, cont.

3. *.phones files – Phonemic transcription

Speaker is non-aphasic control, Production is Climate script

Command: speech2phonectm.sh test2.mp3

```
utterance ID: test2-S0---0000.000-0006.360 θ ɪ ŋ z w ɪ ɫ tʃ eɪ n dʒ ɪ m w eɪ z ð ʌ t
ð ɛ r f r æ dʒ ʌ ɫ ɪ n v aɪ r ʌ n m ʌ n t s ɪ m p ɫ i k æ n t s ʌ p ɔ r t
utterance ID: test2-S1---0006.360-0009.240 ʌ n d ð ʌ t ɫ i d z t u s ɑ r v eɪ ʃ ʌ n ð ʌ
t ɫ i d z d u ʌ n s ɝ t ʌ n t i ɪ t ɫ i d z [SMK]
utterance ID: test2-S2---0009.240-0014.250 [UM] s t ʌ n r ɛ s t ɫ s oʊ ʌ ð i k ɫ aɪ
m ɪ t tʃ eɪ n dʒ ɪ z w ɪ ɫ b i t ɛ r ʌ b ʌ ɫ f r ɝ ð ɛ m
```

4. *.phon.sys -- Phonemic Error Rate (PER), Sentence Error Rate (SER)

Given plain text containing words, and an audio (or video) file, produce a phonetic transcription and compute phone error rate of the audio as it relates to the text file as though the text were a "gold standard".

Speaker is non-aphasic control, Production is Climate script

Command: speech2per.sh test2.mp3 test2.txt

```
%PER   14.69   [ 21 / 143, 7 ins, 5 del, 9 sub ]
%SER   100.00  [1 / 1 ]
```

We validated correspondence of human judgment of general severity with machine judgment error rates.

Conclusions and Future Directions

These methods will allow us to:

- develop automated methods for evaluation and training of spoken language in aphasia and AoS
- greatly improve processing and analysis of data from common measures in which the target is known, such as confrontation naming tests, oral reading assessments, and repetition tasks
- use the detailed time alignment and error type data produced by these systems to understand fluency processes
- characterize levels of severity of AoS
- evaluate the success of training methods and to understand the problems that PWAs with different lesion types have producing fluent speech

Ideas for further work include:

- training the system to work on dialects and accents
- improving the accuracy of single word recognition

References

Fridriksson, J., Hubbard, H. I., Hudspeth, S. G., Holland, A. L., Bonilha, L., Fromm, D., & Rorden, C. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Brain*, *135*(12), 3815-3829.

Haley, K. L. (2011). *Chapel Hill Multilingual Intelligibility Test*. http://www.med.unc.edu/ahs/sphs/card/chmit

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Metze, F., Riebling, E., Fosler-Lussier, E., Plummer, A., & Bates, R. (2015). The speech recognition virtual kitchen turns one. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Acknowledgments

This work was funded by NIH-NIDCD grant R01-DC008524 (20012-2017).