

Towards Automatic Assessment of Aphasia Speech Using Automatic Speech Recognition Techniques

Ying Qin¹, Tan Lee¹, Anthony Pak Hin Kong², Sam Po Law³

¹ The Chinese University of Hong Kong, Hong Kong SAR

² University of Central Florida, Orlando, FL, USA

³ University of Hong Kong, Hong Kong SAR



香港中文大學

The Chinese University of Hong Kong

ISCSLP 2016

Tianjin, China

18th, October, 2016

Background



- Aphasia refers to acquired language impairments resulting from a focal brain damage.
 - Adversely affect one or more modalities of language: speaking, listening, reading and writing.
- Subjective evaluation needs not only clinical knowledge about the disease but also relevant linguistic and cultural background. **Effective and reliable methods for objective assessment of aphasia speech are strongly desired.**
- In the past, speech segmentation and feature extraction required a lot of manual work. Aphasia speech materials used for analysis are limited to isolated words and short sentences of pre-designed content.

Background



- **Automatic speech recognition (ASR) techniques** make it possible to efficiently process a large amount of natural speech and extract different types of speech features for assessment purpose.
- The use of **automatic forced alignment** for analyzing supra-segmental duration characteristics of aphasia speech was investigated by Lee et al. ^[1]
 - 7 duration parameters could be useful features to differentiate aphasia speech from unimpaired speech.
- In Lee et al. ^[2], a general **ASR system** was applied to aphasia speech.
 - Recognition accuracy was low. Acoustic & language models were trained by mismatched speech materials.

Aims



- Use of **ASR techniques** to assist assessment of aphasia speech.
- Improve aphasia speech recognition performance.
 - The acoustic model and language model are trained with domain- and style-matched speech data from unimpaired control speakers.
- Analyze the recognition outputs in the aspects of supra-segmental duration and linguistic content.
 - Investigate the feasibility of using ASR for aphasia speech assessment.



Introduction to speech materials

- About Cantonese
 - Cantonese is a major Chinese dialect.
 - Tonal language: tonal syllable is the smallest meaningful unit.
 - A Cantonese syllable is composed of an *'Initial'* part and a *'Final'* part.
 - character: 踢; base syllable: tek; Initial & Final: I_t F_ek
 - Over 600 base syllables in Cantonese
- Database: Cantonese AphasiaBank (Kong et al. [3])
 - 104 aphasia subjects and 149 unimpaired subjects. All are native speakers.
 - 8 recordings for each subject including narrative tasks of personal monologue, picture descriptions and story-telling.



Introduction to speech materials

- Database: Cantonese AphasiaBank (Kong et al. [3])
 - Cantonese Aphasia Battery (CAB): A standard assessment involves sub-tests measuring fluency, information content, comprehension, repetition and naming ability.
 - Aphasia Quotient (AQ): Overall score, total is 100, higher means better. (Yiu [4])

Subject	Aphasia type	CAB AQ ^a	Age ^b	Gender	Years of Education
ANF04	Anomic	88.1	61;06.09	F	11
ANM13	Anomic	87.7	43;08.05	M	11
ANM06	Anomic	84	51;11.04	M	13
TSM01	Transcortical sensory	81.2	54;01.00	M	8
TSM02	Transcortical sensory	76.3	54;02.10	M	7
WNF01	Wernicke's	73.2	60;07.16	F	9



Cantonese ASR

- Syllable level speech recognizer
 - 20 *Initials* & 53 *Finals* are basic units in acoustic modeling
 - 3-state HMM model for each unit
- Kaldi speech recognition toolkit [4]
- Two methods of acoustic modeling

GMM-HMM	DNN-HMM
<p>13-dimension MFCC feature per frame Contextual feature vector (7 frames) Projected to 40 dimensions by linear discriminant analysis (LDA), followed by maximum likelihood linear transform (MLLT) Speaker adaptive training (SAT) is applied by feature-space maximum likelihood linear regression (fMLLR) transform Context-dependent modeling by decision-tree state tying</p>	<p>40-dimension fMLLR features computed from context window of 11 frames Deep neural network (DNN) has 6 hidden layers with 1024 neurons per layer The number of output neurons is equal to the number of context-dependent HMM states Restricted Boltzmann machine (RBM) is applied to do initialization Backpropagation(BP) algorithm by stochastic gradient descent.</p>

ASR on aphasia speech



- Previous study shown in Lee et al. [2]
 - Test set: 17 aphasia speakers and 17 unimpaired speakers on a specific recording.
 - Training set: CUSENT [5] corpus; 20378 utterances (about 20h)
 - Language model: syllable unigram
- **Present study**
 - Test set: 33 aphasia speakers and 8 unimpaired speakers on 8 tasks.
 - Training set: 106 unimpaired speakers from AphasiaBank; 4790 utterances (13.3h)

	CUSENT	AphasiaBank
Cantonese	服務區域包括廣東省	gam2 aa6 跑咗上去睇
Syllable	fuk6-mou6-keoi1-wik6-baa1- kut3-gwong2-dung1-saang2	gam2-aa6- paau2- zo2 - soeng6-heoi3-tai2
English	Service areas include Guangdong Province.	(Two oral words) Run there to see.



→ Read style vs. Spontaneous

- Language model: syllable bigram

ASR on aphasia speech recognition



- Compare with syllable error rate (SER) shown in Lee et al. [2]
 - For the same 17 aphasia speakers on the same recording

Training data	CUSENT (previous)	Domain-matched data (new)
GMM-HMM	58.2% 17.4% 	40.8%
DNN-HMM	57.8% 23.7% 	34.1%

- SER decreases significantly (GMM-HMM:17.4% ; DNN-HMM: 23.7%)
- The advantage of DNN-HMM over GMM-HMM is more noticeable



ASR on aphasia speech recognition

- SER shown in this study
 - For 33 aphasia speakers and 8 unimpaired speakers with 8 tasks

Test data	Aphasia (33)	Unimpaired (8)
GMM-HMM	43.7%	27.3%
DNN-HMM	40.9%	22.7%

- SER for unimpaired speech is relatively good: 22.7%
- High SER for aphasia speech
 - Acoustic mismatch: aphasia speakers may suffer from impairments on voice and articulation, but healthy speakers do not.
 - Language model is simple: Bigram language model cannot capture the filler words, non-speech or unintelligible speech sound, repeat or repaired words.



ASR on aphasia speech recognition

- Analysis of each aphasia speaker
 - Top 5 with lowest SER

Subject ID	SER %	AQ	Average AQ
A065	19.91	92.1	92.1
A082	21.29	97.8	
A026	21.65	99	
A004	21.74	96.1	
A028	23.04	73.2	

- Similar to the performance level for unimpaired speakers.
- Most of them have high AQ value.



ASR on aphasia speech recognition

- Analysis of each aphasia speaker
 - SER higher than 50%

Subject ID	SER %	AQ	Average AQ
A037	50.30	41.3	69.12
A053	51.10	72.8	
A017	54.38	76.3	
A020	57.93	65	
A021	60.76	66.4	
A022	63.21	88.1	
A054	67.72	80.7	
A046	72.39	86.9	
A030	73.78	71.7	
A049	91.50	42	

- Higher accuracy for less severe speakers; Lower for more severe speakers.



Feasibility of automatic assessment with ASR

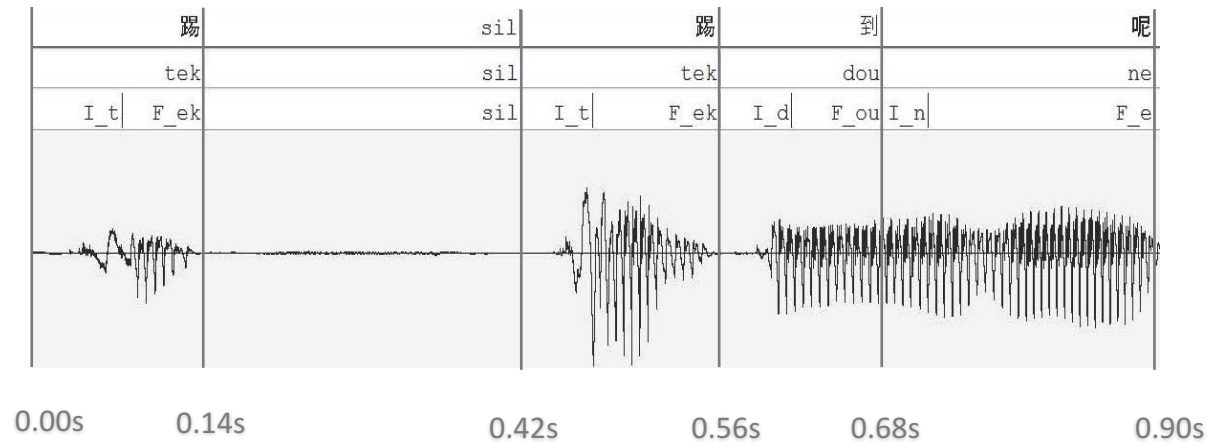
- Characteristics of aphasia speech
 - Duration characteristics
 - Need more short breaks
 - Difficult to speak continuously in a long time span
 - Speak with fewer words
 - Pronounce more slowly
 - Linguistic characteristics
 - Word repetitions occur more frequently
 - More function words, less content words
- Recognition accuracy on aphasia speech could be poor, but characteristics of aphasia speech can still be reflected from ASR results



Duration characteristics

- Duration parameters

- Total number of pauses
- Total number of speech chunks
- Total number of syllables
- Average duration of pauses
- Average duration of speech chunks
- Average duration of syllables
- Average number of syllables in speech chunk



- Duration parameters have been shown to be useful in differentiating aphasia speech from normal speech [1]



Duration characteristics

- Extraction of duration parameters: ASR vs. forced alignment
 - We expect durations given by ASR are comparable to forced alignment.
 - Choose an aphasia speaker with AQ 73.2, SER 23.04%.

Duration parameters	GMM-HMM forced alignment	DNN-HMM ASR
Total number of pauses	173	169
Total number of speech chunks	180	179
Total number of syllables	648	680
Average duration of pauses	1.03	1.07
Average duration of speech chunks	1.37	1.31
Average duration of syllables	0.33	0.29
Average number of syllables in speech chunk	3.6	3.8

- Results are very similar.
- ASR has the potential to analyze duration characteristics.



Linguistic characteristics

- Analysis on contents of ASR result
 - Three sample utterances are spoken by a patient with mild aphasia.
 - AQ: 96.1 SER:21.7%

Transcription

1.e6 搵隻蛋 e6 [x4] 敲碎佢.

2.e6 [x3] 煎蛋啦.

3.跟住加火 [/] 火腿.

Recognition result

1.e wan dei daai e sil e sil e sil e sil gau zo heoi

2.e e sil e sil zin daan laa

3.gan zyu sil gaa fo fo teoi

- Some syllables are wrongly recognized.
- Filler 'e' can be detected.
- Repetitions can be seen.
- ASR on aphasia speech is feasible.



Conclusion

- Automatic speech recognition (ASR) techniques have potential for doing aphasia speech assessment
 - With a reasonable level of aphasia speech recognition accuracy, the ASR can be used for quantifying speech and language impairments.
- More to be done in the future
 - Model unintelligible speech and filler words (acoustic & language modeling).
 - ASR performance needs be further improved for severely impaired speech.

Reference



- [1] Tan Lee, Anthony Pak Hin Kong, Victor Chi Fong Chan, Haipeng Wang, “Analysis of auto-aligned and auto-segmented oral discourse by speakers with aphasia: A preliminary study on the acoustic parameter of duration,” *Procedia, social and behavioral sciences*, vol. 94, p. 71, 2013.
- [2] Tan Lee, Yuanyuan Liu, Pei-Wen Huang, Jen-Tzung Chien, Wang Kong Lam, Yu Ting Yeung, Thomas K. T. Law, Kathy Y. S. Lee, Anthony Pak-Hin Kong and Sam-Po Law, “Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech”, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6475-6479, Shanghai-China, March 2016.
- [3] Anthony Pak Hin Kong, Sam Pow Law, “The Cantonese AphasiaBank”
<http://shsweb.esu.hku.hk:8080/search/>.
- [4] Edwin M-L.Yiu, “Linguistic Assessment of Chinese-speaking Aphasics: Development of a Cantonese Aphasia Battery,” *J.Neurolinguistics*, Volume 7, Number 4, pp. 379-424, 1992.
- [5] Tan Lee, W.K. Lo, P.C. Ching, Helen Meng, “Spoken language resources for Cantonese speech processing,” *Speech Communication*, vol. 36, no. 3, pp. 327-342, 2002.

Thank you!