

Deep learning approach to automatic detection of language impairment in spontaneous speech 深度学习方法在言语障碍自动检测中的应用

Tan Lee 李丹

DSP and Speech Technology Laboratory The Chinese University of Hong Kong 香港中文大学电子工程系

> December 17, 2018 Beijing



香港中文大学 CUHK









语言资源与智能2018









语言资源与智能2018



Speech Research @ CUHK

- Spoken language technology
 - > ASR, TTS, Speaker ID, Language ID, ...
 - Cantonese-focused, multilingual
 - Real-world and real-life speech
 - Deep learning approach
- Wide range of applications
 - Low-resource languages
 - Atypical speech
 - Hearing impairment
 - Music and audio classification



Perspectives

With deep learning methods, high-performance systems depend on

data data data data data data

- All about match and mismatch
- > Typical vs. atypical
- Specificity vs. generalization





Atypical speech

- Unpopular languages: African languages, ethnic minority languages in China, regional dialects, ...
- Code-switching, code-mixing
- L2 speech
- Child voice
- Elderly voice
- Pathological
- Emotional
- Expressive

Atypical speech is meant to have little data. Use of deep learning methods is not straightforward

Inter-disciplinary collaboration





语言资源与智能2018

Collaboration projects



Pathological voices

- CUHK Speech Therapy
- ENT, Sun Yat-Sen Memorial Hospital

Pathological speech (aphasia)

- University of Central Florida Speech and Language Pathology
- University of Hong Kong Speech and Hearing Science

Speech sound disorders of pre-school children

CUHK Speech Therapy

Language acquisition of deaf children

CUHK Linguistics

Speech perception of hearing impaired listeners

- CUHK Audiology
- CUHK Psychology

Behavior therapy: speech in motivational interviewing

CUHK Educational Psychology

Pentatonic Scale Body Constitution (五音体质)

CUHK School of Chinese Medicine



Speech and language impairment

- Difficulties and problems in production of spoken language
- Dysfunctions in linguistic-symbolic planning and/or speech motor control
- Outcome: atypical and abnormal speech
 - > Poor voice quality, hoarseness
 - > Articulatory deficiency, phonological errors
 - > Unnatural, low intelligibility
 - > Stuttering, dis-fluency, lacking intonation
 - Language deficits



Verbal language deficits

- > Abnormality in language expressing in speech
- Not about voice and articulation problems
- > Deficits in
 - Naming
 - Lexical representation
 - Semantic relation
 - > Vocabulary coverage
 - > Discourse organization



Language disorders & brain condition

- Speech and language difficulties reflect the neurological conditions
- Aphasia: acquired language impairment caused by injury or pathology of certain brain areas
- > Neurological injury: stroke or traumatic injury
- Neurodegeneration:
 - > Primary Progressive Aphasia (PPA)
 - Parkinson's disease
 - > Alzheimer's disease (AD)
- Spontaneous speech starts to deteriorate at very early stage of AD progression



Туре	Language	Task	Demo
Aphasia	Cantonese	monologue	
	English	conversation	11 m
	English	dialogue	and a second
MCI	Cantonese	Picture description	







www.aphasia.org





Cantonese AphasiaBank

- A large-scale multi-modal database created to support research on Cantonese-speaking aphasia population
- > Audio recordings of spontaneous speech of native Cantonese speakers
- > 104 aphasic and 149 unimpaired subjects
- Prescribed speech tasks: picture descriptions, procedure description, story telling, monologue



Task	Recording	Description	
Single picture description	CatRe	Black and white drawing of a cat on a tree being rescued	
Single picture description	Flood	A colour photo showing a fireman rescuing a girl	
Sequential picture description	BroWn	Black and white drawing of a boy accidentally break a window	
Sequential picture description	RefUm	Black and white drawing of a boy refusing an umbrella from his mother	
Procedure description	EggHm	Procedures of preparing a sandwich with egg, ham and bread	
Story talling	CryWf	Telling a story from a picture book "The boy who cried wolf"	
Story tening	TorHa	Telling a story from a picture book "The tortoise and the hare"	
Personal monologue	ImpEv	Description of an important event in life	



Picture description



"Now I'm going to show you these pictures."

"Take a little time to look at these pictures. They tell a story. Take a look at all of them, and then I'll ask you to tell me the story with a beginning, a middle, and an end. You can look at the pictures as you tell the story."



Procedure description



"Let's move on to something a little different."

"Tell me how you would make an egg and ham sandwich."





Subjective assessment

- All subjects were assessed using Cantonese Aphasia Battery (CAB)
- Comprehensive assessment: sub-tests on comprehension, repetition and naming abilities, information content, fluency
- Each sub-test results in a numerical score
- Sum of sub-test scores = Aphasia Quotient (AQ)
- > AQ ranges from 0 to 100,
- > 11 to 99 in Cantonese AphasiaBank



ID	Туре	AQ	Information	Fluency	Comprehension	Demo
A003	Anomia	95.8	10	10	10	
A020	Broca's	65.0	9	4	7	(»))
A088	Broca's	27.0	1	1	4.9	

Describe how to make an egg and ham sandwich





Automatic assessment

- Based on analysis of speech signals
- > Evidence-based
- > Non-invasive, objective and repeatable
- > Applying signal processing and machine learning techniques
- Spontaneous and free-style speech is a big challenge

Our approach



语言资源与智能2018





> Prediction of AQ

> Automatically generate AQ to measure the severity of impairment

> Binary classification

- ≻ High-AQ: AQ ≥ 90 (mild/normal)
- > Low-AQ: AQ < 90 (severe)</p>







- Provide a range of information to facilitate extraction of speech and language features
- Linguistic features computed from text
- Duration features derived from time alignment
- Phone posteriors: data-model matching degree





ASR performance

- > ASR for spontaneous and impaired speech is not straightforward
- > DNN acoustic models trained by multi-task learning
 - Domain-matched data: unimpaired speech in Cantonese AphasiaBank
 - Domain-mismatched data: CUSENT and K086
- > Test data: 7 stories from 82 impaired speakers

Acoustic model	Training data	Syllable error rate %
Conventional DNN	CanAB	48.08
TDNN-BLSTM	CanAB	43.35
MT-TDNN-BLSTM	CanAB, K086,CUSENT	38.05

语言资源与智能2018



Design of text features

To compute feature from ASR output text to robustly reflect topic-specific content of a story and differentiate unimpaired story from impaired ones

Cont102. (control)	A023_ANM ~(impaired)		
aa6 有 個 男仔 踢波.	唎 個 就 xxx 空地 度 踢波.⊷		
踢爛 咗 個 玻璃 窗.⊷	gam2 直至 有 日 佢 踢波 嘅 時候.ቍ		
gam2 個 玻璃 窗 有 個 男人.⊷	就食咗個e6 •		
同埋 有 個 燈.⊷	喺邊度去咗個咩個。		
跟住 有 個 男人 攞 住 個 波.⊷	gam2zau6 波 呢 入 咗 去 個 窿 嗰		
就 望 出去 個 爛 咗 個 玻璃 窗 出	度.•		
便 嘞.⊷	gam2 xxx 嗰 個 人 就 攞 住 個 波		
m6.₽	响度 鬧.⊷		

Vector representation of word



- Word transcription is discrete representation, not good for similarity measuring
- Word-embedding: to learn low-dimension continuous-value representation of discrete linguistic units
- Semantic relation learned from data

```
e.g., "King"- "Man" + "Woman" = "Queen"
```



Word embedding

Initially each word represented by a 1-hot vector



- > The 1-hot word vectors are used as input to a neural network
- > A NN model is trained to predict central word from context words





central word



context words



Word embedding



After training, each row of matrix W gives a continuous-value word vector representation



Vector representation of a story



- For Cantonese spontaneous speech, word transcriptions are unreliable and sometimes not attainable
- The 1-best ASR output is in the form of a syllable sequence
- Each syllable is represented by a 50-dimensional vector
- The whole story is represented by taking the average of all syllable vectors
- > The story-level vector reflects topic and content
- > Impaired speech have distorted story vectors







Effectiveness of story vectors

With ground-truth (manual) transcriptions, let us visualize the distributions of story vectors.







If ASR output transcriptions are used, story vectors of impaired speech are even confused







Quantifying the impairment

Intra-story feature: deviation from reference story vectors of impaired speech

 \rightarrow Correlation with AQ: 0.61

Inter-story feature: number of story vectors that are confused with other topics

 \rightarrow Correlation with AQ: 0.82



More robust representation

- > 1-best ASR output contains many errors (~38%)
- > Richer and more inclusive representation from ASR *n*-best, lattice, confusion network
- More robust story vectors could be computed by taking weighted average of all syllable vectors
 - 1. in n-best output
 - 2. In confusion networks



Supra-segmental duration features



- > To capture dis-fluency of impaired speech
- Supra-segmental duration: time length of speech units beyond phonemes

```
Description of features (correlation with AQ)
Non-speech-to-speech duration ratio (-0.684),
# silence segments (-0.460),
average duration of silence segments (-0.633),
average duration of speech segments (0.655)
# spoken syllables (0.448),
# syllable per speech chunk (0.693),
ratio of # silence to # syllables (-0.683),
average # syllable per second (0.688),
ratio of aver. duration of silence to speech segments (-0.667)
ratio of # fillers to the length of speech segment (-0.368),
# long silence segment (-0.420),
# short silence segment (0.146),
average syllable level confidence score from ASR (0.6328)
```





Prediction of AQ: by applying linear regression and random forest (RF) regression to 2 text-based features and 4 best duration features

Correlation between predicted AQ and reference AQ

Text features	LR	RF
1-best of MT-TDNN-BLSTM	0.819	0.839
10-best of MT-TDNN-BLSTM	0.825	0.841
CNs of MT-TDNN-BLSTM	0.827	0.842

For 75.6% (62/82) subjects, prediction errors are smaller than 10, e.g., 45 predicted as 55



Classification: High-AQ vs. Low-AQ



	Binary Decision Tree	Random Forest	SVM
Text features only	0.851	0.896	0.841
Acoustic features only	0.792	0.821	0.789
All features	0.891	0.903	0.874

89.4% (42/47) Low-AQ speakers corrected classified 88.6% (31/35) High-AQ speakers correctly classified







Posteriorgram features

For each time frame of input speech, ASR acoustic models generate a time-posterior matrix known as "**posteriograms**"





What information posteriorgram features might represent

- > Paraphasia: unintended words, extraneous substitutions, e.g., car \rightarrow lar, cat \rightarrow dog
- > Voice disorder: change of voice leads mismatch between acoustic models and input speech
- > Dis-fluency: inappropriate pauses, lengthening of phonemes

Feature design



> Weakly constrained recognizer:

generating posteriograms only based on sounds

Strongly constrained recognizer:

generating posteriorgrams with domain-specific language models

Understandably the "strong" posteriorgram is closer to what is intended to be spoken

Deviation in the "weak" posteriorgram could be due to speech and language impairment







(Each segment is 3 second long)





Unimpaired speech





Mild impairment



uncertainty of strong recognizer





Severe disorder

Vagueness of posterior



low speaking rate and long silences





Binary classification F1 score: ~89%, comparable to the text+duration features



An end-to-end approach

- One single model directly from utterance to prediction score
- Speech features for assessment learned implicitly by neural networks without feature design





End-to-End framework

- > Binary classification: High-AQ vs. Low-AQ
- Classifier: 2-layer unidirectional sequence-to-one
 GRU model and CNN model

	GRU-RNN	CNN
Accuracy (F1 score)	0.77	0.79

- The performance is slightly worse than ASRgenerated features but is comparable
- It's much more efficient than ASR-generated features



End-to-end model is able to learn acoustic features:



Low-AQ speaker (AQ: 55.4)





End-to-end model fails to learn linguistic features:



> This speaker has fluent speech but mostly function words with few content words



Data resource of pathological speech

Publicly available databases of pathological speech are scarce

Dimension of difficulty:

- Language variation
- Variety of pathology

Audio recording in regular clinical work has not become a standard practice in most places



Not just a matter of data ...

- To speech technologists, clinical knowledge is often incomplete, inadequate, disconnected pieces, not well documented and continually evolving
- Clinical goals are diversified and may not be clearly described
- Reliable ground-truth are not available
- Large individual differences among patients
- Multiple types of disorders co-exist



Thank you !

tanlee@ee.cuhk.edu.hk

http://dsp.ee.cuhk.edu.hk

